



Artificial Intelligence and Robotics National Institute

Presented by the
American Bar Association
Section of Science & Technology Law,
Criminal Justice Section,
Government and Public Sector Lawyers Division,
Section of Public Contract Law,
Infrastructure and Regulated Industries Section,
Tort Trial and Insurance Practice Section,
Solo, Small Firm and General Practice Division,
Senior Lawyers Division,
Antitrust Law Section,
Judicial Division,
Cybersecurity Legal Task Force
and ABACLE

ABA
AMERICAN BAR ASSOCIATION™

Artificial Intelligence and Robotics National Institute

Presented by the
American Bar Association
Section of Science & Technology Law,
Criminal Justice Section,
Government and Public Sector Lawyers Division,
Section of Public Contract Law,
Infrastructure and Regulated Industries Section,
Tort Trial and Insurance Practice Section,
Solo, Small Firm and General Practice Division,
Senior Lawyers Division,
Antitrust Law Section,
Judicial Division,
Cybersecurity Legal Task Force
and ABACLE



**American Bar Association
ABACLE
321 North Clark Street, Suite 2000
Chicago, IL 60654-7598
www.abacle.org
800.285.2221**

The materials contained herein represent the opinions of the authors and editors and should not be construed to be the action of the American Bar Association Science & Technology Law Section, Criminal Justice Section, Government and Public Sector Lawyers Division, Public Contract Law Section, Infrastructure and Regulated Industries Section, Tort Trial and Insurance Practice Section, Solo, Small Firm and General Practice Division, Senior Lawyers Division, Antitrust Law Section, Judicial Division, Cybersecurity Legal Task Force or ABACLE unless adopted pursuant to the bylaws of the Association.

Nothing contained in this book is to be considered as the rendering of legal advice for specific cases, and readers are responsible for obtaining such advice from their own legal counsel. This book and any forms and agreements herein are intended for educational and informational purposes only.

© 2019 American Bar Association. All rights reserved.

No part of the publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

Printed in the United States of America.

Product Code: CE2001AIRCMD

Artificial Intelligence and Robotics National Institute

Table of Contents

Faculty and Author Biographies

Sponsors

Keynote 1: Chris Gerdes

Designing Automated Vehicles Around Human Values (*Springer Nature Switzerland*, 2019)

Gereon Meyer and Sven Beiker

Use Cases for Autonomous Driving (Chapter 2 of *Autonomous Driving*, 2016)

Walther Wachenfeld, Hermann Winner, J. Chris Gerdes, Barbara Lenz, Markus Maurer, Sven Beiker, Eva Fraedrich and Thomas Winkle

Keynote 2: R. Patrick Huston

Guest Post: BG Pat Huston on “Future War and Future Law” (*Lawfire*, December 3, 2018)

Charlie Dunlap

Section A: Introduction to the Legal Issues of AI & Robotics

Introduction to the Legal Issues of AI & Robotics (Presentation Slides)

Jimmy Kim, Giancarlo Mori, Christopher Savoie and Stephen S. Wu

AI Product Liability Issues and Associated Risk Management (Chapter 16 of *The Law of Artificial Intelligence and Smart Machines: Understanding A.I. and the Legal Impact*, ABA, 2019)

Stephen S. Wu

See also Section B: AI, Automation, and the Future of Transportation—Unfair and Deceptive Trade Practice Claims Against Manufacturers of Automated Vehicles (Summer 2018)

Section B: AI, Automation, and the Future of Transportation

AI, Automation, and the Future of Transportation (Presentation Slides)

Nandi Chhabra, Selina Pan and Phillip Zackler

The SciTech Lawyer Articles

Unfair and Deceptive Trade Practice Claims Against Manufacturers of Automated Vehicles (Summer 2018)

Stephen S. Wu

When Law and Ethics Collide With Autonomous Vehicles (Fall 2017)

Stephen S. Wu

Law and Liability (Part V of *Autonomous Driving: Technical, Legal and Social Aspects*, Springer, 2015)

Tom Michael Gasser

[https://unglueit-](https://unglueit-files.s3.amazonaws.com/ebf/883c597a81c34af6ad130c08ead0c4d6.pdf#page=517)

[files.s3.amazonaws.com/ebf/883c597a81c34af6ad130c08ead0c4d6.pdf#page=517](https://unglueit-files.s3.amazonaws.com/ebf/883c597a81c34af6ad130c08ead0c4d6.pdf#page=517)

Section C: AI and Robots in the Healthcare Setting

AI & Robotics in the Healthcare Setting (Presentation Slides)

Marissa Urban, Steve Mutkoski and Christopher Hess

AI-Enabled Technologies and Healthcare

Steve Mutkoski

Artificial Intelligence: How to Get it Right (NHS, October 2019)

Indra Joshi and Jessica Morley

Separating Fact from Fiction: Recommendations for Academic Health Centers on Artificial and Augmented Intelligence (Report 23, Winter 2018-2019)

Blue Ridge Academic Health Group

How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature (JAMA, November 12, 2019)

Yun Liu, Po-Hsuan Cameron Chen, Jonathan Krause and Lily Peng

<https://jamanetwork.com/journals/jama/article-abstract/2754798>

Potential Liability for Physicians Using Artificial Intelligence (JAMA, October 4, 2019)

W. Nicholson Price II, Sara Gerke and I. Glenn Cohen

<https://jamanetwork.com/journals/jama/article-abstract/2752750>

Section D: AI in Financial Services

AI in Financial Services (Presentation Slides)

Ted Claypoole

Artificial Intelligence, FTC Authority and International Best Practices in Emerging Technologies and Financial Services (Presentation Slides)

Deon Woods Bell

Section E: Ethics of the Use of AI in the Practice of Law

Ethics of the Use of AI in the Practice of Law (Presentation Slides)

Drew Simshaw, Adam Nguyen and Rafael Baca

Digital Ethics, Morality and the Law

Rafael Baca

Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law (*Hastings Law Journal*, December 2018)

Drew Simshaw

Section F: AI and Robots in the Workplace

AI and Robots in the Workplace (Presentation Slides)

Natalie A. Pierce, R. Jason Straight and Austin Tarango

Section G: Data Privacy, Data Security, and Information Governance

AI Data Privacy, Data Security, and Information Governance (Presentation Slides)

Ian C. Ballon, Kristin J. Madigan, Lucy L. Thomson and Ruth Hill Bro

Excerpts from *E-Commerce and Internet Law: Legal Treatise With Forms*, 2nd Edition, by Ian C. Ballon

Chapter 5: Database Protection, Screen Scraping and the Use of Bots and Artificial Intelligence to Gather Content and Information

Chapter 26: Litigation Risks and Compliance Obligations Under the California Consumer Privacy Act

Chapter 27: California's IoT Law on the Security of Connected Devices

Excerpts from the *ABA Cybersecurity Handbook*, 2nd Edition

Chapter 4: Lawyers' Legal Obligations to Provide Data Security

Thomas J. Smedinghoff and Ruth Hill Bro

Chapter 13: Get SMART on Data Protection Training and How to Create a Culture of Awareness

Ruth Hill Bro and Jill D. Rhodes

Crowell & Moring Articles

Secretary of Defense Esper Calls on Private Sector to Join Forces with

DoD in the Development of AI (November 21, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/Secretary-of-Defense-Esper-Calls-on-Private-Sector-to-Join-Forces-with-DoD-in-the-Development-of-AI>

DOE Invests Again in AI (October 16, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/DOE-Invests-Again-in-AI>

NIST Announces Plan for Federal Engagement in Artificial Intelligence (September 20, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/NIST-Announces-Plan-for-Federal-Engagement-in-Artificial-Intelligence>

Ninth Circuit Rejects Facebook's Article III Argument; Biometric Lawsuit Will Proceed (August 16, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/Privacy-Law-Alert/Ninth-Circuit-Rejects-Facebooks-Article-III-Argument-Biometric-Lawsuit-Will-Proceed>

The U.S. Announces Endorsement of OECD's Principles for Responsible AI (June 4, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/The-US-Announces-Endorsement-of-OECDs-Principles-for-Responsible-AI>

Algorithmic Accountability Act Reflects Growing Interest in Regulation of AI (April 22, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/Algorithmic-Accountability-Act-Reflects-Growing-Interest-in-Regulation-of-AI>

The U.S. Takes Steps to Ensure Its Dominance in the AI Arms Race (February 15, 2019)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/The-US-Takes-Steps-to-Ensure-Its-Dominance-in-the-AI-Arms-Race>

A New Privacy and Data Control Framework in California (June 29, 2018)

<https://www.crowell.com/NewsEvents/AlertsNewsletters/Privacy-Law-Alert/A-New-Privacy-and-Data-Control-Framework-in-California>

GDPR Compliance: The Beginning—Not the End (May 25, 2018)
<https://www.crowell.com/NewsEvents/AlertsNewsletters/all/GDPR-Compliance-The-Beginning-Not-the-End>

With the GDPR, the Dawn of the New Normal Approaches (March 23, 2018)
<https://www.crowell.com/files/20180323-With-The-GDPR-The-Dawn-Of-The-New-Normal-Approaches-3.pdf>

New GDPR Guidance from EU Commission (January 25, 2018)
<https://www.crowelldatalaw.com/2018/01/new-gdpr-guidance-from-eu-commission/>

Additional Online Resources

Competition and Consumer Protection Implications of Algorithms, Artificial Intelligence, and Predictive Analytics (Remarks at Competition and Consumer Protection in the 21st Century, November 14, 2018)
Bruce Hoffman
<https://www.ftc.gov/public-statements/2018/11/competition-consumer-protection-implications-algorithms-artificial>

Human Rights and Technology (Discussion Paper, Australian Human Rights Commission, December 2019)
https://tech.humanrights.gov.au/sites/default/files/2019-12/TechRights2019_DiscussionPaper.pdf

Draft: Chapter 20: California Consumer Privacy Act (October 10, 2019)
California Department of Justice
<https://oag.ca.gov/sites/all/files/agweb/pdfs/privacy/ccpa-proposed-regs.pdf>

California Consumer Privacy Act (CCPA) Fact Sheet
California Department of Justice
https://oag.ca.gov/system/files/attachments/press_releases/CCPA%20Fact%20Sheet%20%2800000002%29.pdf

Connected Cars Workshop, Staff Perspective (January 2018)
Federal Trade Commission
https://www.ftc.gov/system/files/documents/reports/connected-cars-workshop-federal-trade-commission-staff-perspective/staff_perspective_connected_cars_0.pdf

Cybersecurity for Small Business Campaign
Federal Trade Commission
<https://www.ftc.gov/tips-advice/business-center/small-businesses/cybersecurity>

Cybersecurity for Small Business: Understanding the NIST Cybersecurity Framework (November 2, 2018)
Federal Trade Commission
<https://www.ftc.gov/news-events/blogs/business-blog/2018/11/cybersecurity-small-business-understanding-nist>

Section H: Civil and Human Rights Implications of AI

Presentation Outline

Steve Crown, Jessica Fjeld and Vivek Krishnamurthy

Principled Artificial Intelligence: A Map of Ethical and Rights-Based Approaches to Principles for AI

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy and Madhulika Srikumar

Artificial Intelligence & Human Rights: Opportunities & Risks (Berkman Klein Center for Internet & Society at Harvard University, September 25, 2018)

Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz and Levin Kim

The Future Computed: Artificial Intelligence and Its Role in Society

Microsoft

Facial Recognition: It's Time for Action (Microsoft on the Issues, December 6, 2018)

Brad Smith

<https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/>

Section I: Investigations in an Era of AI

Investigations in the Era of AI (Presentation Slides)

Serge Jorgensen, David K. A. Mordecai and Paul Starrett

Online Resources

Section J: Promoting the Progress in AI through IP

Promoting the Progress in AI through IP (Presentation Slides)

Tyler T. Ochoa, Christian Mammen, Brian Adams, Hogene Choi and Colleen Chien

Section K: AI and Robotics Standards, Certifications, and Auditing

AI and Robotics Standards, Certifications, and Auditing (Presentation Slides)

Eric Hibbard and Sumit Kalra

Section L: AI, Robotics, Ethics, and the Public Good

AI, Robotics, Ethics, and the Public Good (Presentation Slides)

Ryan Budish, Theresa Harris, Shannon Vallor and Cynthia Cwik

***Ethically Aligned Design*, First Edition**

Institute of Electrical and Electronics Engineers

Recommendation of the Council on Artificial Intelligence (2019)

OECD

New Artificial Intelligence Tools for Deep Conflict Resolution and Humanitarian Response (*Procedia Engineering*, 2015)

Daniel J. Olsner

Artificial Intelligence & Human Rights: Opportunities & Risks (Berkman Klein Center for Internet & Society at Harvard University Research Publication No. 2018-6, September 25, 2018)

Filippo A. Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz and Levin Kim

Location-Based Data in Crisis Situations: Principles and Guidelines (American Association for the Advancement of Science, March 2019)

Jessica Wyndham, Ellen Platts and Jonathan Drake

Accelerating Social Good With Artificial Intelligence: Insight From the Google AI Impact Challenge

Google

AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (*Minds and Machines*, November 26, 2018)

Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke and Effy Vayena

Understanding Media and Information Quality in an Age of Artificial Intelligence, Automation, Algorithms and Machine Learning (Berkman Klein Center for Internet & Society at Harvard University Research, July 12, 2018)

Yochai Benkler, Robert Faris, Hal Roberts and Nikki Bourassa

Section M: Looking Into the Crystal Ball

Looking Into the Crystal Ball (Presentation Slides)

Keith Abney, Ryan Clark, Daniel Dries and Stephen Wu

The Rise of AGI/Robots and Artificial Personhood: Near- to Far-Term Ethical & Legal Implications (Presentation Slides)

Keith Abney

Quantum Leap: Accelerating America's Growth in Quantum Computing

Ryan C. Clark

Neural Devices Will Change Humankind: What Legal Issues Will Follow? (*The SciTech Lawyer*, Winter 2012)

Stephen S. Wu and Marc Goodman

AI Conceptual Risk Analysis Matrix (CRAMSM)

Martin Ciupa and Keith Abney

Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed (Robot Ethics: The Ethical and Social Implications of Robotics)

Keith Abney

Robot Ethics: Mapping the Issues for a Mechanized World (*Artificial Intelligence*, 2011)

Patrick Lin, Keith Abney and George Bekey

Robots and Space Ethics (Chapter 23 of *Robot Ethics 2.0*)

Keith Abney

Space War and AI

Keith Abney

See also *Section A: Introduction to the Legal Issues of AI & Robotics—AI Product Liability Issues and Associated Risk Management* (Chapter 16 of *The Law of Artificial Intelligence and Smart Machines: Understanding A.I. and the Legal Impact*, ABA, 2019)

See also *Section G: Data Privacy, Data Security, and Information Governance—Get SMART on Data Protection Training and How to Create a Culture of Awareness* (Chapter 13 of the *ABA Cybersecurity Handbook*, ABA, 2018)

Faculty and Author Biographies



Keith Abney is a senior lecturer in the Philosophy Department and a research fellow of the Ethics + Emerging Sciences Group at California Polytechnic State University in San Luis Obispo. His areas of expertise include many aspects of emerging technology ethics and bioethics, especially issues in space ethics and bioethics, robotics, AI and cyberethics, autonomous vehicles, human enhancements and military technologies. He is a co-editor of *Robot Ethics* (MIT Press) and *Robot Ethics 2.0* (OUP), as well as an author and contributor to numerous other books, journal articles and funded reports.



Brian Adams is an associate attorney at Qualcomm Technologies, Inc., in San Diego, California.



Rafael Baca is an attorney at R Baca Law Firm, PLLC, in Palo Alto. Mr. Baca has decades of experience as a federal regulatory and tech lawyer and a registered U.S. patent and IP lawyer. He has a depth of experience in complex commercial litigation, legal transactions, business intelligence, trade secrets, patent prosecution and global asset portfolio strategy and regulatory work in tech and engineering. He is an active speaker in software and technology legal matters and GDPR data privacy regulations and is a national expert in artificial intelligence and big data. Over the years, Mr. Baca worked in patent litigation management and patent prosecuting in software and decentralized and cryptographic computing, as well as telecom and wireless systems, in-house patent transactions, international licensing and IP valuation and large-firm litigation.



Ian C. Ballon is a litigation shareholder and co-chair of the Global Intellectual Property & Technology Practice Group at Greenberg Traurig LLP in East Palo Alto. He represents clients in copyright, DMCA, trademark, trade secret, right of publicity, privacy, security, software, database and Internet- and mobile-related disputes and in the defense of data privacy, security breach, behavioral advertising, TCPA and other Internet-related class action suits. Mr. Ballon, who splits his time between the firm's Silicon Valley and LA offices, is the author of the five-volume legal treatise, *E-Commerce and Internet Law: Treatise With Forms 2d Edition* (Thomson Reuters West, 2008 & 2018 Cum. Supp.) and the earlier first edition. He is also the author of *The Complete CAN-SPAM Act Handbook* (West, 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West, 2009).



Ruth Hill Bro is a privacy and cybersecurity attorney in Chicago. She has focused her legal career on advising businesses on privacy and information management strategy, cybersecurity, global compliance, the electronic workplace and e-business. She has been featured as a speaker on these issues over 160 times and has over 90 published works on these topics. She is a longstanding leader in the American Bar Association, where she co-chairs the ABA Cybersecurity Legal Task Force, serves on the ABA E-Mail Stakeholder Committee and is a leader in the ABA Section of Science & Technology Law (SciTech). In SciTech, she is a senior advisor for the Privacy, Security, and Emerging Technology Division, a member of the planning committee (2015-2019) for the ABA's first four Internet of Things (IoT) National Institutes and the section's liaison to the ABA Commission on Women in the Profession.



Ryan Budish is an assistant research director at the Berkman Klein Center for Internet & Society in Cambridge, Massachusetts, where his current research areas include the ethics and governance of artificial intelligence, with a focus on global governance, corporate and multistakeholder governance and algorithmic transparency and accountability. Since joining Berkman Klein in 2011, he has worked with national and international organizations on a range of important topics, such as the OECD's development of AI governance principles, the UN's creation of a System Wide Action Plan for AI and the NTIA's review of ICANN's accountability mechanisms. Before arriving at Berkman Klein, Mr. Budish was an associate at Covington & Burling LLP, where he focused on privacy, media, technology and health care.



Nandi Chhabra is general counsel at Peloton Technology in Mountain View, California. Mr. Chhabra provides strategic legal advice and counseling to Peloton and leads its legal department. He advises on all legal matters, including corporate, commercial, product liability, intellectual property and regulatory matters. At Peloton, Mr. Chhabra draws on his prior experience in the technology startup sector, and his work in the White House and U.S. Senate. Mr. Chhabra holds a J.D./M.B.A. from Stanford University and an A.B. from Princeton University, majoring in the Woodrow Wilson School of Public Policy and International Affairs. Mr. Chhabra was a Fulbright Scholar to China.



Colleen V. Chien is a professor at Santa Clara University School of Law, where she teaches, mentors students and conducts empirical research on innovation, intellectual property and the criminal justice system. From 2013 to 2015, she served in the Obama White House as a senior advisor on intellectual property and innovation, working on a broad range of patent, copyright, technology transfer, open innovation and other issues. Professor Chien is nationally known for her research and publications on domestic and international patent law and policy issues. She has testified on multiple occasions before Congress, the DOJ, the FTC and the U.S. Patent and Trademark Office on patent issues; frequently lectures at national law conferences; and has published several in-depth empirical studies. Prior to entering academia, Professor Chien did stints as an investigative journalist, strategy consultant and practicing lawyer (as an associate, then special counsel at Fenwick & West LLP in San Francisco).



Hogene L. Choi is a partner at Baker Botts in Palo Alto. She works on a range of intellectual property matters, focusing primarily on patent prosecution, transactions and counseling, and she co-chairs the Patent Counseling & Strategic Portfolio Development Practice Group. Ms. Choi has patent prosecution and transactions experience covering technologies related to machine learning and artificial intelligence, blockchain, computer vision, cloud infrastructure and services, internet applications and server-side architecture, desktop applications and operating systems, graphics and audio/video, as well as semiconductors, medical devices, electronics, nanotechnology and the mechanical arts.



Ryan C. Clark is a patent attorney and associate in the Austin office of Baker Botts and a member of the Intellectual Property Department. Mr. Clark helps clients protect their ideas, designs, brands and expressions through a focused practice centered on patent and trademark litigation, patent and trademark prosecution and counseling. Mr. Clark represents clients in federal court patent and trademark litigation, including cases involving multiple patents and multiple defendants. From a counseling and prosecution perspective, he has experience managing the patent process from ideation through enforcement, including managing global patent and trademark portfolios and building clients' internal patent processes. He also counsels companies in intellectual property due diligence for M&A transactions.



Theodore F. Claypoole leads the Womble Bond Dickinson IP Transaction Team and its FinTech Team in Atlanta, and until recently, its Privacy and Cyber Security Team. Mr. Claypoole is immediate past chair of the ABA Cyberspace Law Committee in the Business Law Section. He serves on the Leadership Council of the ABA Business Law Section and as the liaison between the ABA Business Law Section and the Standing Committee on Law and National Security. He formerly worked as in-house technology counsel for Bank of America and for CompuServe. He is coauthor with Theresa Payton of two books from Rowman & Littlefield Publishers, titled *Privacy in the Age of Big Data* and *Protecting Your Internet Identity: Are You Naked Online?* Mr. Claypoole has assembled and edited a book called *The Law of Artificial Intelligence and Smart Machines* for the American Bar Association, published in August 2019.



Steve Crown is vice president and deputy general counsel for human rights at Microsoft in Redmond, Washington. Mr. Crown and his team focus on development of internal policies and business practices, as well as advocacy and external engagements, that serve to advance realization of human rights. Focus areas include freedom of expression and privacy on the global internet and development of rights-respecting norms to govern development and deployment of artificial intelligence. Mr. Crown holds leadership positions on the executive committees of the Seattle Chamber, the Global Network Initiative, the American Society of International Law and the International Bar Association (Media Law Committee). He received his J.D. from Yale Law School.



Cynthia Cwik is a continuing fellow at the Stanford University Distinguished Careers Institute in Palo Alto. She is a nationally recognized leader in issues at the intersection of law, science and technology. At Stanford, Ms. Cwik is focusing on cutting-edge technology issues, including the ethical, economic, legal, policy and societal implications of emerging technologies, such as artificial intelligence. She also participated in programs and courses on corporate governance and global perspectives for business entities. Ms. Cwik has practiced law at two global law firms, including serving five years as a partner with Jones Day and 15 years as a partner with Latham & Watkins. She represented Fortune 500 corporations in high-stakes, high-profile matters involving science and technology issues, including counseling clients concerning federal and state regulatory and compliance issues.



Heather Deixler is a corporate counsel in the San Francisco office of Latham & Watkins. Ms. Deixler counsels public and private companies operating in the healthcare and life sciences industries on transactional and regulatory matters. She advises hospitals and physician organizations, as well as digital health, pharmaceutical and medical device companies on privacy and security, physician self-referral (i.e., the federal Stark law and its state counterparts) and fraud and abuse. Ms. Deixler advises clients on innovative healthcare delivery systems, including Medicare Accountable Care Organizations, clinically integrated networks, IPAs and other value-based payment programs.



Daniel F. Dries is lead IP counsel at Psi Quantum in San Francisco.



Eric Y. Drogin is a fellow of the American Academy of Forensic Psychology, a diplomate and former president of the American Board of Forensic Psychology and a diplomate of the American Board of Professional Psychology. Dr. Drogin currently holds faculty appointments with the Harvard Medical School (as a member of the Program in Psychiatry and the Law, and on the staff of the Forensic Psychiatry Service, in the Department of Psychiatry at Beth Israel Deaconess Medical Center) and the Harvard Longwood Psychiatry Residency Training Program. Additional positions have included chair of the American Psychological Association (APA) Committee on Professional Practice and Standards, chair of the APA Committee on Legal Issues, chair of the APA Joint Task Force with the American Bar Association, and president of the New Hampshire Psychological Association.



Jessica Fjeld is a lecturer on law and the assistant director of the Cyberlaw Clinic at the Berkman Klein Center for Internet & Society in Cambridge, Massachusetts. She focuses her legal practice on issues impacting digital media and art, including intellectual property; freedom of expression, privacy and related human rights issues; contract; and corporate law. Recently, she has emphasized work with AI-generated art, the overlap of existing rights and ethics frameworks on emerging technologies and legal issues confronted by digital archives. She is a member of the board of the Global Network Initiative, a multistakeholder organization that protects and advances user freedom of expression and privacy around the world. Before joining the Clinic, Ms. Fjeld worked in Business & Legal Affairs for WGBH Educational Foundation, where she advised the American Archive of Public Broadcasting, along with numerous WGBH productions.



Julie A. Fleming is the founder of Fleming Strategic in Cheyenne, Wyoming.



Chris Gerdes is the director of the Stanford Center for Automotive Research Dynamic Design Lab. His laboratory studies how cars move, how humans drive cars and how to design future cars that work cooperatively with the driver or drive themselves. From February 2016 to January 2017, Mr. Gerdes served as the first chief innovation officer at the U.S. Department of Transportation. In this role, he worked with Secretary Anthony Foxx to foster the culture of innovation across the department and find ways to support transportation innovation taking place both inside and outside of government. He was part of the team that developed the Federal Automated Vehicles Policy and represented the Department on the National Science and Technology Committee Subcommittee on Machine Learning and Artificial Intelligence. He continues to serve U.S. DOT as vice chair of the Federal Advisory Committee on Automation in Transportation.



Peter J. Gillespie is a partner at Laner Muchin, Ltd., in Chicago. He represents and counsels management on a wide array of employment law–related issues, including workplace safety and health, wage and hour laws, prevailing wage issues, covenants not to compete, discrimination and harassment, wrongful discharge, whistleblower claims, class actions, hiring, discipline, promotion and dismissal decision-making, workplace privacy, and statutory compliance. He provides employers with strategic advice to help meet their objectives while reducing potential litigation risks. Mr. Gillespie handles litigation in both federal and state courts, as well as claims pending with state and federal administrative agencies including the Occupational Safety and Health Administration, the Equal Employment Opportunity Commission, the U.S. Department of Labor, the Illinois Department of Human Rights and the Illinois Department of Labor.



Theresa L. Harris is a project director in the Scientific Responsibility, Human Rights and Law Program at the American Association for the Advancement of Science (AAAS) in Washington, D.C. Her interests are the implementation of internationally recognized human rights principles in domestic practice and the intersection of information technology, human rights and law. Prior to joining AAAS, she led Human Rights USA as its executive director, where she represented survivors of human rights violations before U.S. courts, the Inter-American human rights system and United Nations human rights mechanisms. Ms. Harris has served on the Board of Directors of Amnesty International USA and the governing body of the World Organization Against Torture. She holds a B.A. in Anthropology, an M.S. in Urban and Regional Planning and a J.D. from American University Washington College of Law.



Matt Henshon is a partner at Henshon Klein LLP in Boston. He delivers sophisticated expertise to its clients, usually relating to transactions and/or functioning as outside general counsel to emerging and mid-sized firms. HKLLP also represents individuals in a range of related business contexts, including risk mitigation, governance issues and estate planning. Clients include high-net-worth individuals, board members, and executives. Immediately prior to forming HKLLP (and its predecessors), Mr. Henshon served as "traveling" chief of staff to Senator Bill Bradley (D-NJ) during Senator Bradley's campaign for the presidency. He writes frequently on business, political and legal topics, and his work has appeared recently in the *Boston Business Journal* and on the *New York Times* op-ed page.



Michele Herman is a partner at Justech Law in Seattle.



Christopher P. Hess is professor and chairman of the Department of Radiology and Biomedical Imaging at the University of California, San Francisco (UCSF). He completed clinical residency and fellowship training at UCSF after obtaining his doctorate in electrical engineering at the University of Illinois, studying signal processing and magnetic resonance imaging. His research interests lie clinically in computational neuroimaging of brain development and degeneration, epilepsy, vascular disease and the development of techniques for ultra-high-field and diffusion MRI. A fellow of the American Institute for Medical and Biological engineering, he has published over 150 peer-reviewed papers in clinical and scientific journals. His current research is funded by the National Institutes of Health and a number of industry partners.



Eric A. Hibbard is Hitachi Data Systems' CTO for security and privacy in Santa Clara, where he leads the Hitachi product-oriented security and privacy strategy activities with an emphasis on data and storage security. He is a senior security professional with expertise in information assurance, privacy, storage, cloud computing, Internet of Things, e-discovery and enterprise ICT. He leverages this expertise and extensive experience in the public and private sectors in leadership roles within the ABA, CSA, INCITS, IEEE and SNIA. Mr. Hibbard currently serves as the ISO Editor of *ISO/IEC 27050 (Electronic Discovery)*. He speaks internationally and is published. Mr. Hibbard holds a BSCS along with the CISSP-ISSAP, ISSEP, ISSMP, CCSP and CISA certifications.



Brigadier General R. Patrick Huston is stationed at The Pentagon in Washington, D.C., as the assistant judge advocate general. He oversees the Army's international legal engagements, criminal prosecutions and government appeals. He also supervises the legal teams that provide advice on national security law, contracts, administrative law and criminal law. He is focused on the legal and ethical development and use of artificial intelligence, autonomous weapons, cybersecurity and other emerging technologies. He also supports diversity and inclusion initiatives as part of talent management for the JAG Corps, one of the world's largest legal organizations. General Huston started his military career as an Army ranger and helicopter pilot stationed in Europe. He then attended law school and became a military prosecutor in Korea. General Huston has completed five combat tours in Iraq and Afghanistan.



Serge Jorgensen is the president and founding partner in the Sylint Group in Sarasota, Florida. He provides technical development and guidance in the areas of cybersecurity, counter-cyberwarfare, e-discovery, system design and incident response. Mr. Jorgensen is a patented inventor in engineering and math-related fields. He directed development of domain name server (DNS) tracking applications, provided response and remediation guidance to multibillion dollar international espionage and cybersecurity attacks and directed, tasked and managed multimillion-dollar litigation, forensic and e-discovery efforts. Prior to the Sylint Group, Mr. Jorgensen ran the Research and Development Department for Locast Corporation, developing a HIPAA-compliant patient location and status tracking device.



Sumit Kalra is a board member at Miracle League of San Francisco Peninsula.



Jimmy Kim is the business operations lead at Built Robotics, a company building automated guidance systems for heavy equipment. Previously, Mr. Kim was a consultant at BCG, driving strategic initiatives for technology and retail clients. He graduated from Washington University in St. Louis, where he studied finance and economics. A native of New Jersey, Mr. Kim has been interested in technology and manufacturing ever since his first internship at Boeing.



Vivek Krishnamurthy is the Samuelson-Glushko professor of law and director of the Samuelson-Glushko Canadian Internet Policy and Public Interest Clinic (CIPPIC). Mr. Krishnamurthy's teaching, scholarship and clinical legal practice focus on the complex regulatory and human rights-related challenges that arise in cyberspace. He advises governments, activists and companies on the human rights impacts of new technologies and is a frequent public commentator on emerging technology and public policy issues. Along with his former colleagues at Harvard's Berkman Klein Center, Mr. Krishnamurthy is the author of a landmark study for Global Affairs Canada that evaluates the risks and opportunities for human rights that artificially intelligent systems present.



Laura O. I. Lemire is an attorney at Microsoft in Redmond, Washington. Leveraging her data protection expertise, she supports the Worldwide Commercial Business, including the Cybersecurity Solutions Group. Additionally, Ms. Lemire focuses on Microsoft's field readiness efforts. She counsels the Microsoft Learning Team and develops training on data protection and compliance matters related to cloud services and emerging technology, such as artificial intelligence.



Kristin J. Madigan is a counsel in Crowell & Moring's San Francisco office and a member of the firm's Litigation and Privacy & Cybersecurity Groups. Ms. Madigan focuses her practice on representing clients in high-stakes complex litigation with a focus on technology, as well as privacy and consumer protection matters including product counseling, compliance, investigations, enforcement and litigation that typically involves existing and emerging technologies such as the Internet of Things, artificial intelligence, autonomous vehicles, enterprise and cloud-based software, blockchain and distributed ledgers. Prior to joining Crowell & Moring, Ms. Madigan served as an attorney at the FTC in Washington, D.C., in the Bureau of Consumer Protection, Division of Privacy and Identity Protection. In that role, Ms. Madigan led nonpublic investigations and prosecuted unfair and deceptive practices in privacy and data security matters within the FTC's consumer protection authority.



Christian E. Mammen is a partner at Womble Bond Dickinson in Silicon Valley. He has more than 20 years of experience guiding Silicon Valley and global tech and life sciences clients in high-stakes patent and intellectual property litigation. He has substantial lead counsel experience and has led both large and small trial teams. He has also served as lead counsel on appeals before the Ninth and Federal Circuits. Dr. Mammen is an accomplished scholar, with significant teaching and academic experience. His clients include companies in the software, telecom, microelectronics and pharmaceutical/biotech/life sciences sectors. He holds a doctorate in law from Oxford University and has held visiting faculty positions at UC Hastings School of Law, UC Berkeley Law School, Stanford Law School and Oxford University. He clerked for Judge Robert Beezer on the U.S. Court of Appeals for the Ninth Circuit.



Richard M. Martinez is a partner at Jones Day LLP in Minneapolis. His practice focuses on technology and its impact on society. He has extensive domestic and international experience in technology and intellectual property matters and in cybersecurity, data privacy and information law. Mr. Martinez has been involved in patent licensing campaigns that have generated eight- and nine-figure settlements and has represented Fortune 100 corporations, leading educational institutions, technology start-ups and individuals in privacy and cybersecurity matters, including wire transfer fraud, Wiretap Act claims, privacy torts and disputes relating to the ownership of data.



Peter McLaughlin is a partner at Womble Bond Dickinson LLP in Boston. While maintaining a broad privacy practice, Mr. McLaughlin focuses on innovative uses of data, especially with the life sciences and digital health sectors. He also guides clients in their domestic and international handling of personal information; new product development; and the assessment of legally defensible cybersecurity programs. He spent several years in-house at a global Silicon Valley technology company and as assistant general counsel and global privacy officer for a multinational health firm. He has represented clients across industry sectors with respect to governing personal information; responding to regulators from the Federal Trade Commission, the U.S. Department of Health and Human Services and state attorneys general; and supporting post-enforcement compliance obligations.



Lois Deborah Mermelstein is the founder of the Law Office of Lois D. Mermelstein in Austin. An attorney and software engineer, she focuses her practice on intellectual property and technology matters, mostly involving patents.



David K. A. Mordecai is president and co-founder of Risk Economics, Inc., a New York City–based advisory firm. As lead for the RiskEcon Litigation, Regulation and Arbitration Expert Advisory Practice, Mr. Mordecai serves as an expert on loss causation and economic damages related to market structure, financial institutions governance, and complex issues related to finance, economics and market standards and practices within securities, derivatives, reinsurance and commodities markets, as well as market structure within a broad range of nonfinancial industry sectors. His expertise includes financial engineering, the valuation of fixed-income securities and structured products, including over-the-counter derivatives (in particular fixed income and credit derivatives), complex insurance and reinsurance liabilities, as well as asset liability and risk management models and practices.



Giancarlo Mori is the CEO of Movyl Technologies in San Francisco. Mr. Mori has more 20 years of software development experience, 17 of which were spent at the senior and executive levels in the video games/mobile industry. Mr. Mori has recently served in C-level positions in several major entertainment companies, such as Glu Mobile, Atari and Animal Logic. Prior to that, Mr. Mori held senior and executive positions at Electronic Arts, Microsoft and Activision. For the last four years, he has worked exclusively in the mobile and social games space. Mr. Mori has experience with small as well as large and complex organizations, establishing and scaling up a business and all operational and business aspects of the mobile industry. Mr. Mori holds an MSc in Geosciences for the University of Florence.



Stephen Mutkoski is the worldwide health policy director at Microsoft in Mountain View, California.



Adam V. Nguyen is the cofounder and senior vice president of eBrevia in San Francisco. Mr. Nguyen formulates and executes strategies with respect to eBrevia's revenue, product and corporate development. Mr. Nguyen is responsible for many of eBrevia's key customers and partners and is a regular speaker at industry events. In addition, Mr. Nguyen is involved in the company's financing activities and daily operation. Previously, Mr. Nguyen practiced law at the global law firms of Paul, Weiss and Shearman & Sterling, where he focused on mergers and acquisitions and investment fund formation, and at AQR Capital Management, where he helped expand the firm's alternative investment products and managed accounts. He clerked for the Hon. Judge Faith Hochberg of the U.S. District Court of New Jersey.



Tyler Trent Ochoa is a professor Santa Clara University School of Law. Professor Ochoa is a recognized expert in copyright law and rights of publicity. He joined the Santa Clara University School of Law faculty in 2003, and he served as academic director of the High Technology Law Institute for the 2005-2006 academic year. Prior to joining Santa Clara Law, Professor Ochoa served as a professor and co-director of the Center for Intellectual Property Law at Whittier Law School. He has also served as a clerk for the Hon. Cecil F. Poole of the U.S. Court of Appeals for the Ninth Circuit and as an associate with the law firm of Brown & Bain in Palo Alto, where he specialized in copyright and trade secret litigation involving computer software. He is also a two-time *Jeopardy!* champion and a champion on *Win Ben Stein's Money*.



Selina Pan is a research scientist at Toyota Research Institute in Los Altos, California, where she works on control, integration and human-vehicle interaction on the Guardian shared autonomy system. Prior to joining Toyota, she was a controls research scientist at Ford Research & Innovation Center. She was a postdoctoral scholar in the Dynamic Design Laboratory at Stanford University, where her research focused on ethics, driver adaptation and integrated path planning and path tracking in autonomous vehicles. She received her M.S. and Ph.D. degrees from the University of California, Berkeley, in mechanical engineering, researching automotive engine control and unmanned aircraft. While at Berkeley, she also taught as an adjunct lecturer at California Maritime Academy. She currently serves as the Industry Liaison of the ASME Dynamic Systems and Control Division's Automotive Transportation Systems Technical Committee.



Natalie A. Pierce is a shareholder and co-chair of the Robotics, AI and Automation Practice Group at Littler Mendelson PC in San Francisco, California. She advises employers of all sizes, ranging from start-ups to global corporations, on all aspects of the employer-employee relationship. She focuses on identifying and avoiding problems before they arise through the use of compliance audits and development of workplace policies and procedures. She handles workplace investigations, competition and confidentiality matters, employee classification issues, employee terminations, and reductions-in-force. Additionally she is a litigator who represents corporations in all types of employment litigation matters before state and federal courts and agencies defending clients against claims involving discrimination, harassment, wrongful discharge and retaliation.



Christopher J. Savoie is the CEO and founder of Zapata Computing, Inc, in Boston, Massachusetts. He is an expert in distributed artificial intelligence and inventor of the NLU engine behind Apple's Siri. He is also a published scholar in medicine, biochemistry and computer science.



Brian Scarpelli is senior policy counsel at ACT The App Association in Washington, D.C.



Deborah Shelton is a partner at Arent Fox in Washington, D.C. Ms. Shelton helps clients navigate complex regulations in the life sciences industry, with a specific emphasis on biotechnology, medical devices and pharmaceuticals. She has more than 20 years of experience as an FDA regulatory attorney, with a focus on biotech products and medical devices, including digital therapeutics. In addition to representing clients in matters before FDA, Ms. Shelton advises clients in matters regulated by the DEA, FTC, and USDA. Her clients include companies of all sizes, from startup enterprises to multinational corporations, for whom she provides strategic counsel in navigating complex regulations that impact every stage of product development, approval and marketing.



Drew Simshaw is an assistant professor at the Gonzaga University School of Law in Spokane, Washington. He researches and writes about the interplay between artificial intelligence and legal ethics, access to justice and legal education. His article, "Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law," was recently published in the *Hasting Law Journal*. Before joining the Gonzaga Law faculty, Professor Simshaw taught at the Georgetown University Law Center as a visiting associate professor of law, legal practice. As a supervising attorney with the Institute for Public Representation in Washington, D.C., he specialized in communications and technology law and represented public interest organizations in rulemakings and adjudications before federal agencies and in litigation before federal appellate courts.



Paul Starrett is the founder of Starrett Law in Saratoga, California. Mr. Starrett specializes in high-profile investigations, cyber risk and compliance consulting. He has investigated thousands of cases over a 12-year career as a corporate-security executive for Fortune 500 companies to include internal and third-party fraud, conflicts-of-interest, violations of compliance laws, other criminal and civil matters and significant violations of company policy and procedure. He has managed hundreds of cases over a 10-year period as an information-governance and litigation-management professional, including some of the largest and most notable matters that came about as a result of the financial collapse during the start of the great recession.



R. Jason Straight is a senior managing director and chief privacy officer at Ankura, based in the New York Office. Mr. Straight is a leader in the cybersecurity and privacy consulting practice and oversees Ankura's internal data privacy program. He has extensive experience managing complex cybersecurity investigations and data breach events in a wide variety of industries involving a range of threat actors, including malicious insiders, organized criminal operations and state-sponsored groups. In addition, Mr. Straight has overseen and led large data risk and privacy compliance consulting matters for global companies facing regulatory challenges arising from the General Data Protection Regulation, the California Consumer Privacy Act, HIPAA, federal and state financial services regulations and other frameworks.



Austin Tarango is product counsel supporting Google's research and artificial intelligence group. He counsels a wide range of product areas, including robotics, quantum computing, natural language understanding and AI-enabled education applications. Prior to working at Google, Mr. Tarango was an attorney in private practice, focusing on commercial technology transactions and intellectual property litigation.



Lucy L. Thomson is a principal at Livingston PLLC in Washington, D.C., where she focuses her practice on legal and technology issues related to cybersecurity, global data privacy, and compliance and risk management. She helps organizations develop and implement practical solutions to address the multitude of complex security challenges arising from new technologies, and to prevent and respond to data breaches and secure critical infrastructure. Previously a senior engineer at CSC, a global technology company, she gained extensive hands-on experience conducting privacy and risk assessments and developing FISMA security plans on two of the government's largest technology modernization projects—Customs and Border Protection (CBP-ACE) and the Internal Revenue Service (IRS). While at CSC she was appointed a Department of Homeland Security (DHS) Information System Security Officer (ISSO).



Marissa R. Urban is associate product counsel at Google in Mountain View, California.



Shannon Vallor is currently the McKenna professor in philosophy at Santa Clara University in Silicon Valley and recently was appointed to take the Baillie Gifford chair in the ethics of data and artificial intelligence at the University of Edinburgh in February 2020. Professor Vallor's research addresses the ethical implications for human character of emerging science and technology, especially AI, robotics and new media. In addition to her many articles and published educational modules on the ethics of data, robotics and artificial intelligence, she is the author of the book *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press, 2016) and the forthcoming *Lessons from the AI Mirror: Rebuilding Our Humanity in an Age of Machine Thinking*.



Deon Woods Bell is a senior international attorney in the U.S. Federal Trade Commission's Office of International Affairs and serves as counsel for International Consumer Protection and Data Privacy. She manages the FTC's consumer protection and data privacy technical assistance portfolio and is the agency's lead staff at the UN Consumer Protection Intergovernmental Group of Experts. She is currently focusing on FinTech and artificial intelligence policy and enforcement matters, among other issues. Since joining the FTC in 2007, Ms. Woods Bell has been responsible for a variety of consumer policy and legal issues. She has worked on bilateral and regional cooperation, technical assistance, privacy and legal enforcement matters, in addition to matters involving the nexus between consumer protection and competition policy.



Stephen S. Wu is a shareholder with Silicon Valley Law Group in San Jose, California. Mr. Wu advises clients concerning transactions, compliance, liability risk management, privacy, security and governance of emerging information technologies, such as artificial intelligence, autonomous and connected vehicles, robotics, big data and the Internet of Things. He negotiates technology agreements, resolves disputes for clients and serves as an outside general counsel for emerging companies. Mr. Wu also advises clients on governing and assessing corporate programs to promote AI compliance and ethics. An author of seven IT legal books and numerous other publications on artificial intelligence, automated driving and robotics, Mr. Wu served as the 2010-2011 chair of the ABA Science & Technology Law Section. In 2007, he helped found the Section's Artificial Intelligence and Robotics Committee. He serves as chair of the ABA Artificial Intelligence and Robotics National Institute.



Phillip Zackler is the director of legal and compliance at the Toyota Research Institute in Los Altos, California.

Artificial Intelligence & Robotics 2020 Sponsors

Gold Sponsor



SoftBank Group Corp. is a strategic holding company that invests in AI and other transformative technologies for the betterment of humanity.

Bronze Sponsor



Baker Botts is a globally respected law firm with 725 lawyers and 14 international offices. We are driven by the highest ethical and professional standards. This professionalism, combined with industry knowledge and insights and our understanding of the law, helps us to deliver effective, innovative solutions for our clients.

With a presence in the U.S., Europe, Asia and the Middle East, Baker Botts provides our clients with responsive service both domestically and internationally. This broad reach enables us to help companies around the world go wherever their industry goes—to successfully respond to challenges, minimize risks and maximize opportunities,

For more than 175 years, Baker Botts has delivered results-oriented services, establishing us as a leading law firm. Our reputation is complemented by our leadership in government, the judiciary and our communities. Regardless of size, sector or jurisdiction of a client, our commitment is to help achieve their business objectives. BakerBotts.com

Break Sponsor

Womble Bond Dickinson (US) LLP

Keynote 1: Chris Gerdes

Lecture Notes in Mobility

Gereon Meyer
Sven Beiker *Editors*

Road Vehicle Automation 6



Designing Automated Vehicles Around Human Values

J. Christian Gerdes¹, Sarah M. Thornton¹, and Jason Millar²(✉)

¹ Stanford University, 416 Escondido Mall, Building 550, Room 136,
Stanford, CA 94305, USA

gerdes@stanford.edu, smthorn@alumni.stanford.edu

² University of Ottawa, 5026C-800 King Edward Ave, Ottawa,
ON K1N 6N5, Canada

jmillar@uottawa.ca

Abstract. The impact of automated vehicles will reverberate across society in many dimensions, changing our expectations of mobility, safety, employment and other aspects of life we value. These major societal changes will, in turn, be the result of a number of small engineering decisions that, when aggregated, determine the system behavior. For automated vehicles to have the benefits their advocates envision, we must bridge the gap between these individual decisions and the societal impacts they create. This paper discusses some of the challenges faced by engineers in bridging this gap and proposes a value-centered approach to the design of automated vehicles. Such an approach engages stakeholders early in the process, identifying values and tensions with enough specificity to drive subsequent engineering choices.

Keywords: Autonomous vehicles · Human values · Ethical programming

1 Introduction

Have you ever driven down a street you have never driven down before because you were guided there by a navigation app? With in-car navigation systems and cell phones, such experiences have become a regular aspect of driving for most Americans. In our quest for reduced travel time and less stressful driving, we have delegated the choice of what streets we travel to routing algorithms that can account for factors such as real-time traffic about which we can only guess.

Yet, while drivers appreciate the time and stress saved, neighborhoods now have to contend with increased traffic. Streets that were once only known to local residents are now regular commute routes, changing the character of the neighborhoods through which they wind. Stories abound of quiet streets turned into “speedways” by the presence of these new road users [1, 2].

This battle over neighborhood streets is a classic example of a value tension. There is an obvious conflict between travelers wanting the shortest driving time and neighborhoods wanting to preserve their relative tranquility. It is easy to argue that increased traffic through residential neighborhoods with children, pedestrians and bicycles poses

an increased threat to safety. But, similarly, roads are a public good. Shouldn't they be used as efficiently as possible? When faced with such tensions, who should decide?

In the past, these competing demands were balanced by traffic engineers using standards for road design that were themselves the result of decades of research and public comment. Increasingly, however, these decisions are made by the engineer who designs the routing algorithm, often without much awareness of the values implicated by that choice or the tensions that might arise as a result. What is most striking is how quickly we have moved from road usage determined by standards to road usage determined by weights in a routing algorithm. As the reach of systems like navigation apps expands and automated vehicles take to the streets, small engineering decisions will increasingly, and rapidly, produce large societal impacts. Instead of societal aspects such as road safety or traffic being the aggregate of many human decisions or carefully crafted standards, huge swaths of our society will hinge on small engineering decisions.

In such a future, it becomes critical to center engineering decisions around the human values and societal characteristics impacted by those decisions. But what exactly are human values? Borning and Muller [3] define “value” as “what a person or group of people consider important in life.” When considering transportation, mobility is an important human value but far from the only one. The human act of driving requires a constant balancing of competing values such as mobility, safety and legality, the results of that balancing act differing according to how each driver handles the many driving scenarios encountered in a single trip. This balancing act can be complicated. Depending on the scenario, one value may clearly take precedence over another, such as when drivers decide to respect a stop sign at a busy intersection—here, safety trumps mobility.

In designing automated vehicles, engineers cannot avoid making decisions that implicate safety, mobility and legality. Several designers of motion planning algorithms have accordingly sought to address and deliberately resolve value tensions with their algorithm designs. Bouton, Nakhaei, Fujimura, and Kochenderfer [4] explicitly balanced mobility and safety in the design of a motion planning algorithm for navigating occluded scenarios by penalizing collisions and rewarding the vehicle for completing maneuvers. Wongpiromsarn, Karaman, and Frazzoli [5] encoded traffic laws in order to synthesize a motion planning algorithm that navigates a two-lane roadway with an obstacle and a double yellow line. The autonomous vehicle is able to navigate around the obstacle because the traffic rules are encoded such that crossing a double yellow is allowable after the vehicle first comes to a stop behind the obstacle, thus balancing legality, safety and mobility.

These examples show how engineers can incorporate or resolve value tensions when those tensions are clear from the start. The fact that humans care about values, and that engineers can incorporate identified values in their designs, motivates us to consider designing autonomous vehicles with human values in mind. The next crucial piece is to consider *how* to go about designing with human values in mind. In the following sections, we present a value-centered approach to designing automated vehicles that focuses the design process on human values from the beginning. In what we propose, human values serve as common ground in all discussions of automated vehicle design across stakeholder groups. While resolving tensions between different

values can be challenging, starting the discussion from the standpoint of values and having meaningful conversations about those tensions need not be. Human values are vitally important; where we strive for our designs to function well in society, it makes sense to situate values at the core of engineering decision-making.

2 Some Value Tensions with Automated Vehicles

To further illustrate the need for considering human values and the challenges engineers may face when attempting to do so, we detail a few examples of design problems that can arise during the development of an automated vehicle. These examples not only cover different aspects of automation but also three different challenges to a value-centered design approach: identifying the underlying values, resolving tensions among those values and acknowledging scenario-dependent immediacy effects.

2.1 Routing Algorithms

A foundational feature of an automated vehicle is to deliver occupants (or goods) from point A to point B. In order to accomplish this feature, a route must be planned for the automated vehicle to follow. A popular routing algorithm is a shortest path search algorithm known as A^* (pronounced A-star) [6], which considers both cost and heuristics for traversing a road segment. There are many factors that can be included in the creation of the cost and heuristics. For example, speed limit and length of the road segment could be included in the cost calculus. The heuristic may be a rough guide, like the Manhattan distance, in order to encourage exploration of the search space in the direction of the target destination.

From the standpoint of an AV programmer, it makes perfect sense to consider routing in terms of the travel time for the occupants. But, as discussed in the Introduction, this is only one of many values associated with vehicle routing. Vehicle routing affects the tranquility of neighborhoods, the environmental impact of the transportation system and, in a mobility-on-demand system, waiting times and service quality. Safety may vary across potential routes, with routes through areas with high concentrations of vulnerable road users decreasing safety for those outside the vehicle and routes through areas of high crime potentially decreasing safety for vehicle occupants. Should these dimensions be added to routing algorithms for AVs or does doing so raise the prospects of neighborhoods being redlined and unable to share in the benefits AVs will provide? Routing algorithms demonstrate how difficult it can be for engineers to properly scope the range of human values relevant to a design choice in isolation. Stakeholder engagement to form an inclusive and diverse understanding of those values becomes critical.

2.2 Vehicle Platooning

One approach to automating vehicles involves creating platoons of vehicles that can follow each other on the highway at close spacing, as depicted in Fig. 1. In the case of heavy trucks, the focus of such systems has been to reduce fuel consumption [7] and,

through coordinated braking, reduce the likelihood of collision in certain accident scenarios. There are many values to consider beyond these improvements in fuel consumption and safety, however. Other road users need to be able to maneuver around or between members of a platoon, particularly at highway exits. When both trucks are operated by human drivers, visibility becomes a key aspect of both safety and human comfort. Furthermore, there is a central tension between the two main objectives of the system. Traveling at a closer spacing can improve fuel economy but raise safety concerns about the vehicles colliding with each other due to, for instance, different braking capability.



Fig. 1. Two heavy truck vehicles platooning on the highway (photo: Peloton Technologies)

The challenge for engineers designing platooning systems is not to surface these values or tensions; all of these are clear from a single platooning experience on the highway. Rather, the challenge is that all of these values hinge on a single decision – the choice of following distance. While this number is trivial to set or modify in a following algorithm, all aspects of system performance and public acceptance stem from that single choice. Platooning is therefore a great example of the need to develop ways of prioritizing or balancing values when so many values are determined by a single design parameter.

2.3 Pedestrian Interactions

Automated vehicles navigating urban environments must determine the appropriate speed at which to approach pedestrian crosswalks. In the California Vehicle Code, §21950 states that drivers must yield the right of way to any pedestrian within a marked crosswalk. When the pedestrian is actually in the crosswalk, the law’s requirement and the necessary vehicle action are rather straightforward. But in order to respect this law

and ensure pedestrians' safety, automated vehicles must also anticipate whether or not the pedestrian will be in the crosswalk at the time the vehicle arrives. As human drivers know, such predictions can be far from trivial. Figure 2 depicts a pedestrian standing in an influential area next to the crosswalk. The pedestrian is completely stopped while looking down at a device he is holding, presenting an ambiguous situation for the automated vehicle. While sometimes this ambiguity resolves as the vehicle gets closer, in this case it does not and the AV must choose a speed through the intersection accordingly. The value tension manifests as conflicts over mobility of different road users and among safety, mobility and legality in setting a speed.



Fig. 2. An automated vehicle approaches a pedestrian crosswalk. At approximately 10 m away (left), it is difficult to discern the intent of the pedestrian. Closer to the crosswalk (right), it is still difficult to discern the intent of the pedestrian.

Viewing this as a tension between the vehicle and the pedestrian alone, however, obscures other value tensions that arise. While legal concerns and the immediacy of the pedestrian may prompt the engineer to frame this design problem narrowly, and take a very conservative approach to speed setting in these cases, such choices impact a broader set of indirect road users as well. Vehicles behind the AV may see their mobility unnecessarily decreased if the algorithm is too conservative, while the broader effects could be an unacceptable decrease in overall systemwide mobility. Furthermore, some human drivers may attempt to overtake the AV as a result, increasing other risky behaviors overall. Crosswalks present an example where narrowly considering value tensions of the immediate stakeholders may fail to adequately reflect the values of others impacted by these decisions.

These three examples are just a sampling of the value tensions that arise when designing automated vehicles, a technology that will alter society in myriad dimensions across an array of stakeholder groups. While these examples highlight different challenges in value-centered design, they share a common theme: all of these tensions involve what seem like rather mundane decisions. In fact, the societal impact of automated vehicles will ultimately be determined by a vast number of seemingly small engineering decisions. The ethical design of automated vehicles, therefore, requires that these decisions be grounded in human values.

3 Ethical Programming, Not Programming Ethics

This focus on the broader impact of small engineering decisions runs counter to the way that ethics is often discussed with respect to automated vehicles. In the past few years, the topic of ethics for automated vehicles has often been equated to finding solutions for the infamous “trolley problem” arising from philosophy. The trolley problem, first invented as a philosophical thought experiment, asks one to consider a scenario in which an uncontrollable trolley is barreling unstoppably down a set of railway tracks because of its broken brakes. If the trolley continues on its way, it is guaranteed to kill five unsuspecting people on the tracks ahead. You (a bystander) are standing by a switch, giving you the power to intervene and divert the trolley to another set of tracks on which there is a single unsuspecting person that would surely perish [8]. Moral psychologist Greene [9] has demonstrated that variations of the trolley problem can provide deep insights into the way the human brain processes these ethical dilemmas.

It’s easy to replace the trolley with an autonomous vehicle and imagine that an algorithm would need to determine whether to continue on the collision course with five pedestrians or swerve and kill one other. Given that this would be a tragic scenario for an autonomous vehicle (or anyone) to encounter on the roadway, the trolley problem, imagined as a real-world hypothetical, has jumped into the discussion of automated vehicles. Researchers around the world have conducted surveys and human subject experiments centered around trolley car scenarios [10, 11]. These studies attempt to crowdsource public opinion on what an autonomous vehicle should do in a no-win crash scenario when it is confined to the choice of hitting one entity over another.

Such a narrow framing of the problem contrasts with the way engineers actually program automated vehicles. The motion of a vehicle is not chosen at a single point in time by weighting known outcomes but rather evolves from a series of choices of the desired speed and path to take through an uncertain and ever-changing environment. Despite the lack of connection to the challenges faced by engineers developing motion planners, the plethora of recent papers and stories deploying no-win scenarios in this field has come to dominate the discussion of ethics for automated vehicles. It’s common to find people who have the impression that engineers are (or should be) designing machines that solve these no-win moral dilemmas.

Early on in our work, we also interpreted the challenge of designing automated vehicles with ethics as a question of how to program ethics explicitly into computer algorithms [12]. To that end, in June 2015 we hosted an interdisciplinary workshop at the Center for Automotive Research at Stanford, including philosophers, engineers, lawyers and psychologists, titled “Towards Programming Ethics in Automated Vehicles.” This workshop was a great success. Not because we came away with a clear answer on how to program ethics, but because one philosopher, Dr. Shannon Vallor of Santa Clara University, challenged us to think about the problem differently. In her opening talk, she stated that the problem we faced wasn’t one of trying to program ethics into vehicles, but rather to find ways to program ethically. According to Vallor, the name of our workshop misplaced the focus, in that we should be focusing on

developing processes and techniques that encourage engineers to program ethically. By using the language of values, we hope to put the focus back where it belongs and address the real ethical issues that arise with automated vehicles.

4 A Value-Centered Approach

So how do we put human values at the center of automated vehicle design? In one sense, the answer is strikingly obvious: we simply need to begin the conversation from the standpoint of human values. Values can form a common language across stakeholder groups, helping to identify potential value tensions up front and informing the search for resolutions. Admittedly, though, this is not a language that most people use in their day-to-day discussions of engineering problems or automated vehicles. Some tools and methods are therefore needed to facilitate this discussion.

We are far from the only ones to recognize the importance of human values in engineering design; some engineers routinely approach design from this perspective. Nor are we the only ones to highlight the need for methodologies that better connect human values and engineering. A variety of design processes with overlapping terminology such as “value”, “values” and “worth” [13] have been developed to make some of these connections. Much of our inspiration comes from Value Sensitive Design (VSD), a general design methodology that formalizes the connection between human values and technology [14, 15]. As a methodology, VSD asks the designer to focus on a broad set of stakeholders that may be implicated by the designed technology (e.g. users, policy makers, the environment, the public), the values of those stakeholders, and the value tensions that may exist between different competing values. VSD is not restrictive to a formal process on achieving the integration of human values into the designed technology and, hence, is very open-ended in that respect.

We have been developing a structured approach to workshops that facilitate convening stakeholders, familiarizing them with the language of values and identifying values and value tensions from their input. Our approach endorses VSD’s core considerations and may look very similar to VSD in specific applications [16]. In many ways our approach acts as a “wrapper” around VSD, constraining it for use in an engineering environment by providing a process and some specific tools for facilitating value-based conversations.

In this value-centered approach, we start by bringing together a wide-range of stakeholders as part of a stakeholder engagement session. When designing automated vehicles, such stakeholders could be executives and engineers from an automated vehicle company, policy makers, transportation officials, uniformed services or even local business owners and citizens. The broader the group of stakeholders, the wider the range of perspectives the designers will be able to consider. Bringing together this group of stakeholders helps to surface underlying values at stake in the design task. It also helps to identify value tensions and potential types of resolutions. The various perspectives can also help to frame the underlying problem in the design task. We have found that the key to an effective workshop is to base discussions around a scenario that is specific enough for people to understand, reveals the key value tensions and is accessible by stakeholders with different backgrounds [17].

After uncovering the value tensions in the design task, we can strategically approach the process of resolving these tensions. While there are many possible approaches to resolving value tensions, three that we have found to be particularly effective are:

- (1) **Prioritize** – resolve the tension by choosing one value as the most important. Remaining values can then be addressed within the design space where this primary value is satisfied. For instance, in the case of truck platooning, safety can be chosen as the primary value. Once an appropriate level of safety has been set, the other values can be fulfilled to the greatest extent possible while maintaining this level of safety.
- (2) **Compromise** – resolve the tension by balancing among competing objectives. In this approach, the goal is to partially fulfill the wishes for multiple values. Such an approach is a natural resolution to the conflicts in a routing algorithm, where neighborhood concerns could be balanced with travel time and traffic impacts in a cost function. When values can be appropriately quantified, this approach lends itself well to analysis of pareto optimality.
- (3) **Reframe** – resolve the tension by changing the problem and dissolving the tension. One of our example scenarios involves tensions between a pedestrian and vehicle at a crosswalk. It is possible to envision road designs or future traffic control approaches where rights of way are handled entirely differently, making the original problem disappear [18]. Solutions of this type are often the most effective but can require significant change.

While workshops can be highly effective at identifying values, tensions and strategic approaches to resolving value tensions, the task of ensuring this information is incorporated into the design falls to the engineering team. Within the framework of VSD, this occurs during the technical implementation phase. To harness the information derived from the workshop, the technical implementation should provide a means for realizing the identified human values. Additionally, it is important to map the design parameters (such as the following distance from the platooning semi-truck example) to the associated values. This serves as clarification and a way to be explicit about what values will be captured in the system. Lastly, the technical implementation should enable treatment of the value conflicts. The strategies identified during the workshop provide an initial framing for how tensions or conflicts can be resolved. An important part of the technical implementation process is to determine whether or not this framing still seems appropriate with the greater level of specificity needed to make hard ethical choices. Even in the event that these approaches need to be revisited and revised, the fact that they were determined in the first place makes it easier for engineers to reengage stakeholders in a structured way.

Once a technology is implemented, we can analyze the system to determine if we captured the values appropriately. This entails revisiting the values and tensions identified during the stakeholder engagement. It may also include reconvening with some of the original stakeholders in the engagement session. This analysis component is a great opportunity for interdisciplinary discussion. Engineers can evaluate the design using simulation or even conduct user studies to gain feedback on the experience. Societal impacts can be estimated if appropriate models - such as traffic

simulations or economic projections - are available. Ethicists can provide a rigorous critical analysis of the underlying values and justificatory approaches for resolving value tensions. Lawyers can provide a legal analysis given a particular implementation by conducting jury research or looking over case law.

A design process that is centered around human values creates common terminology for an interdisciplinary analysis that spans stakeholder groups. We believe a design process focused on human values and tensions will enable designers of automated vehicles to design technology that is societally accepted because potential value tensions have been identified and addressed ahead of time, thus aligning resulting automated vehicles with a robust set of value-centered justifications.

5 Conclusion

Engineers of automated vehicle technology make many small engineering decisions that can accumulate into systemic behavior that impacts society in a myriad of ways. It may not be obvious to all engineers which human values may be implicated in a design task, so we propose that a value-centered approach will help bridge the gap between engineering decisions and positive social impact. First, stakeholder engagement allows for a diverse and inclusive set of perspectives to enter the design process thus surfacing key values and value tensions. Secondly, designing the technology around those identified human values and tensions will help to bring about engineering implementations that help to resolve them earlier on in the design process. Finally, analysis based around human values allows for interdisciplinary discussion by using the values as common terminology. By advocating a human values-centered approach to designing automated vehicles, we believe society will be able to partake in the full benefits automated vehicles promise.

Acknowledgements. The authors would like to thank Daimler AG and the Ford Motor Company for their support of this research and for participating in our workshops.

References

1. Lopez, S.: LA Times. <https://www.latimes.com/local/california/la-me-lopez-encino-waze-20180530-story.html> (2018)
2. Foderaro, L.W.: New York Times. <https://www.nytimes.com/2017/12/24/nyregion/traffic-apps-gps-neighborhoods.html> (2017)
3. Borning, A., Muller, M.: Next steps for value sensitive design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1125–1134 (2012)
4. Bouton, M., Nakhaei, A., Fujimura, K., Kochenderfer, M.J.: Scalable decision making with sensor occlusions for autonomous driving. In: Proceedings of the IEEE International Conference on Robotics and Automation (2018)
5. Wongpiromsarn, T., Karaman, S., Frazzoli, E.: Synthesis of provably correct controllers for autonomous vehicles in urban environments. In: Proceedings of the IEEE Conference on Intelligent Transportation Systems, pp. 1168–1173 (2011)

6. Hart, P.E., Nilsson, N.J., Raphael, B.: A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* **4**(2), 100–107 (1968)
7. Lammert, M.P., Duran, A., Diez, J., Burton, K., Nicholson, A.: Effect of platooning on fuel consumption of class 8 vehicles over a range of speeds, following distances, and mass. *SAE Int. J. Commercial Veh.* **7**(2), 626–639 (2014)
8. Thomson, J.J.: Killing, letting die, and the trolley problem. *Monist* **59**(2), 204–217 (1976)
9. Greene, J.D.: *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. Penguin, London (2014)
10. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. *Science* **352**(6293), 1573–1576 (2016)
11. Wächter, M.A., Faulhaber, A., Blind, F., Timm, S., Dittmer, A., Sütfeld, L.R., Stephan, A., Pipa, G., König, P.: Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for ADVs. *ArXiv e-prints*, 2017. arXiv: <https://arxiv.org/abs/1706.07332> [cs.CY]
12. Gerdes, J.C., Thornton, S.M.: Implementable ethics for autonomous vehicles. In: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (eds.) *Autonomous Driving: Technical, Legal and Social Aspects*, pp. 87–102. Springer, Berlin (2016)
13. Kujala, S., Väänänen-Vainio-Mattila, K.: Value of information systems and products: understanding the users’ perspective and values. *J. Inf. Technol. Theory Appl.* **9**(4), 4 (2009)
14. Friedman, B., Kahn Jr., P.H.: Human values, ethics, and design. In: Jacko, J.A., Sears, A. (eds.) *The Human-Computer Interaction Handbook*, pp. 1177–1201. Lawrence Erlbaum Associates, Mahwah (2003)
15. Friedman, B., Kahn Jr., P.H., Borning, A.: Value sensitive design and information systems. In: Zhang, P., Galletta, D. (eds.) *Human-Computer Interaction and Management Information Systems*, vol. 5, pp. 348–372. M.E. Sharpe, Armonk (2006)
16. Thornton, S.M., Lewis, F.E., Zhang, V., Kochenderfer, M.J., Gerdes, J.C.: Value sensitive design for autonomous vehicle motion planning. In: *Proceedings of the IEEE Intelligent Vehicles Symposium* (2018)
17. Millar, J., Paz, D., Thornton, S.M., Parisi, C., Gerdes, J.C.: Design for human values – a framework for addressing ethical considerations in the engineering of automated vehicles (and other technologies), in preparation (2019)
18. National Association of City Transportation Officials: “Blueprint for Autonomous Urbanism,” Module 1, Designing Cities Edition, Fall 2017

Walther Wachenfeld, Hermann Winner, J. Chris Gerdes,
Barbara Lenz, Markus Maurer, Sven Beiker, Eva Fraedrich
and Thomas Winkle

W. Wachenfeld (✉) · H. Winner
Institute of Automotive Engineering – FZD, Technische Universität Darmstadt,
64287 Darmstadt, Germany
e-mail: wachenfeld@fzd.tu-darmstadt.de

H. Winner
e-mail: winner@fzd.tu-darmstadt.de

J.C. Gerdes
Department of Mechanical Engineering Center for Automotive Research at Stanford,
Stanford University, Stanford, CA 94305, USA
e-mail: gerdes@stanford.edu; gerdes@cdr.stanford.edu

B. Lenz
Geography Department, Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: barbara.lenz@dlr.de

M. Maurer
Institute of Control Engineering, Technische Universität Braunschweig,
38106 Braunschweig, Germany
e-mail: maurer@ifr.ing.tu-bs.de

S. Beiker
Formerly Center for Automotive Research at Stanford, Stanford University,
Palo Alto, CA 94304, USA
e-mail: sven@svenbeiker.com

E. Fraedrich
Geography Department, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
e-mail: eva.fraedrich@geo.hu-berlin.de

T. Winkle
Department of Mechanical Engineering, Institute of Ergonomics, Technische Universität
München – TUM, 85747 Garching, Germany
e-mail: winkle@carforensic.com

2.1 Motivation for the Consideration of Use Cases

Although autonomous driving is characterized (see Chap. 1) by the definition for “fully automated” according to BASt [1] as well as the quote by Feil [2] “self-determination within the scope of an higher (moral) law”, it is possible to come up with a large variety of usage scenarios and specifications for autonomous driving. In order to grasp this variety, proxies are sought, which on the one hand make use of distinguishing characteristics, and on the other hand describe typical usage scenarios for autonomous driving. In the following, these will be called use cases for autonomous driving. Besides the nomenclature, the use cases are defined by their distinguishing characteristics, so that a common understanding can be reached for all writing and reading these book chapters. In addition, the use cases are supposed to serve as reference scenarios for further discussion. It is not intended to exclude other examples. However it is recommended to use the defined use cases to avoid misunderstanding or oversight. The following definitions and assumptions can additionally be expanded for the different book chapters with detailed descriptions. As for the different book chapters, definitions and assumptions are relevant in different ways. For instance the owner relations are less important for a technical point of view than for taking a look at the market impact. Thus, definitions and assumptions are to be examined critically. Desired results from working with these use cases are a founded change of definitions and assumptions as well as possible controversy, which arise in between the different topics (different parameter sensitivity).

The following description of the use cases is structured in 4 sections. Section 2.2, general assumptions, describes the limitations and assumptions that are used and are supposed to apply for all use cases. Section 2.3 introduces the four selected use cases and defines the specific characteristics. Section 2.4 explains the selection and the level of detail for the characteristics describing the use cases. Section 2.5, general definitions, proposes definitions, which facilitate a unique description of the use cases.

2.2 General Assumptions

Besides the characteristics which distinguish the use cases, and which are listed in the following section, there are additional attributes, which apply to the chosen use cases as well. The following general assumptions describe these attributes.

Mixed operation: One basic assumption is that the use cases are deployed at the considered time in a mixed operation of transportation systems with different levels of automation. Road traffic consists of vehicles with all levels of automation ranging from “driver-only” to “assisted” to “fully automated”. During the stepwise introduction of automation, both human vehicle operation and driving robot operation are equally likely.

Failures: Hardware or software failures can also happen with autonomously driven vehicles. However, it is assumed that a vehicle designed according to the state of the art (e.g. ISO 26262) is, with regard to the failures mentioned, at least as reliable and safe as today's vehicles.

Level of detail: The description of the use cases is not a detailed specification. Instead of a detailed description of weather conditions, light conditions, road surface conditions etc. the following simplification is assumed. The quality as well as the success rate with which the driving robot performs the driving task is similar to the human quality and success rate. For example, heavy rain leads only to transition to the safe state and discontinuation of the transportation task when a driver would discontinue the journey as well. This document does not tackle the question of whether this assumption from the user's point of view, the society's point of view etc. is sufficient. Furthermore, in this document the question of how this quality and success rate is quantified and proved remains unanswered.

Conformity with regulations: For all use cases it is assumed that the autonomous journey is performed compliant with the set of rules of the respective jurisdiction (federal/national level, state level in the United States), in which the driving actually takes place. The question about the action in dilemma situations directly arises from this assumption. Is the driving robot permitted, or is it even possible, to disregard rules in order to prevent major damage? For these use cases it is assumed that a legally valid set of rules, respectively meta-rules, exists, which the driving robot follows. In order to do so, the respective authority has granted permission to perform autonomous driving, while it is not further contemplated how such permission can be obtained and what the respective rules might be.

2.3 Description of the Use Cases

The motivations and general assumptions underlying the use cases are laid out above, and the characteristics considered for their description are explained in Sect. 2.4. The combination of these characteristics and/or their values leads to a very large number of use cases, which cannot be described in detail. The four use cases described in the following serve, as mentioned above, as proxies for this multitude of possible use cases. Other use cases are not disregarded but our focus is set on the following four:

- Interstate Pilot Using Driver for Extended Availability.
- Autonomous Valet Parking.
- Full Automation Using Driver for Extended Availability.
- Vehicle on Demand.

The partition of the driving task between human and driving robot, in which the four versions differ, has particularly contributed to the selection of the use cases. The first two use cases are seen as introductory versions, while the two latter use cases present widely developed versions of autonomous driving.

2.3.1 Interstate Pilot Using Driver for Extended Availability

An exemplary use case of the interstate pilot is depicted in Fig. 2.1.

2.3.1.1 Benefit

The driving robot takes over the driving task of the driver exclusively on interstates or interstate-like expressways. The driver becomes just a passenger during the autonomous journey, can take his/her hands off of the steering wheel and pedals, and can pursue other activities.

2.3.1.2 Description

As soon as the driver has entered the interstate, he/she can, if desired, activate the driving robot. This takes place most logically in conjunction with indicating the desired destination. The driving robot takes over navigation, guidance, and control until the exit from or end of the interstate is reached. The driving robot safely coordinates the handover to the driver. If the driver does not meet the requirements for safe handover, e.g. because he/she is asleep or appears to have no situation awareness, the driving robot transfers the vehicle to the risk-minimal state on the emergency lane or shortly after exiting the interstate. During the autonomous journey, no situation awareness is required from the occupant; the definition for fully automated driving according to BASt [1] applies. Because of simple scenery and limited dynamic objects, this use case is considered as an introductory

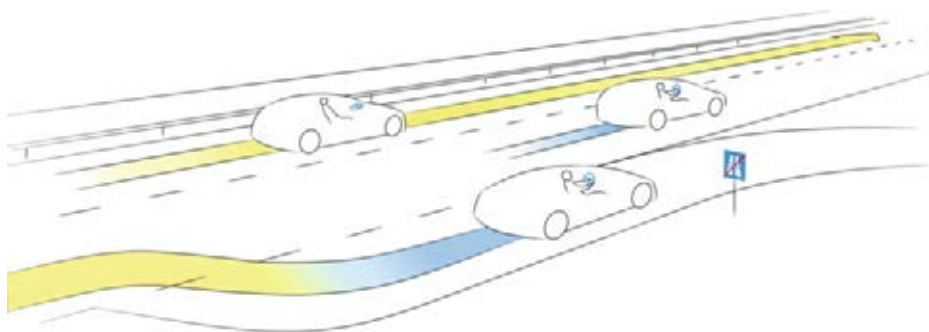


Fig. 2.1 Interstate pilot using driver for extended availability

scenario, even if the comparatively high vehicle velocity exacerbates accomplishing the risk-minimal state considerably.

2.3.1.3 Values of Characteristics

Table 2.1 summarizes the characteristics for the interstate pilot use case. Figure 2.2 shows the intervention possibilities for instances on the levels of the driving task for the use case Interstate Pilot. “The entities which can intervene into the driving task are depicted on the

Table 2.1 Values of characteristics for interstate pilot using driver for extended availability

Characteristic		Value	
A	Type of occupant	3	Person/s with agreed destinations
B	Maximum permitted gross weight	1–3	500 kg to 8 t
C	Maximum deployment velocity	4	Up to 120 km/h
D	Scenery	8 a	Interstate Without permission allowed
E	Dynamic elements	2	Only motor vehicles
F	Information flow between driving robot and other entities	1–4	Navigation optimization, guidance optimization, control optimization, provision of environmental information
G	Availability concept	2	Availability through driver
H	Extension concept	2	Driver
I	Options for intervention		Figure 2.2: interstate pilot options for intervention

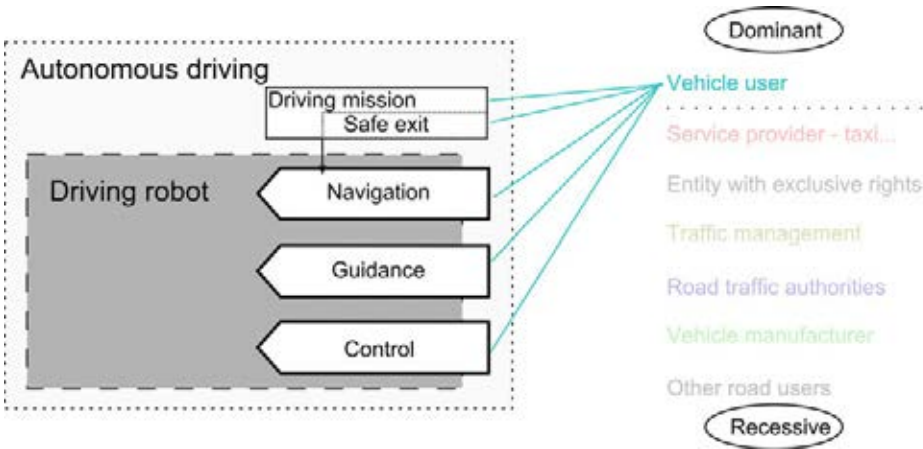


Fig. 2.2 Interstate pilot options for intervention

right side of the hierarchy and are sorted from dominant at the top to recessive at the bottom.” The vehicle user is the only entity which may intervene. It should be emphasized again that the handover is managed in a safe manner through the driving robot. Potential service providers, police and ambulance with specific authority, a traffic coordinator etc. do not have any possibility to intervene with the vehicle control.

2.3.2 Autonomous Valet Parking

An exemplary use case of the autonomous valet parking is depicted in Fig. 2.3.

2.3.2.1 Benefit

The driving robot parks the vehicle at a remote location after the passengers have exited and cargo has been unloaded. The driving robot drives the vehicle from the parking location to a desired destination. The driving robot re-parks the vehicle.

The driver saves the time of finding a parking spot as well as of walking to/from a remote parking spot. In addition, access to the vehicle is eased (spatially and temporally). Additional parking space is used more efficiently and search for parking is arranged more efficiently.

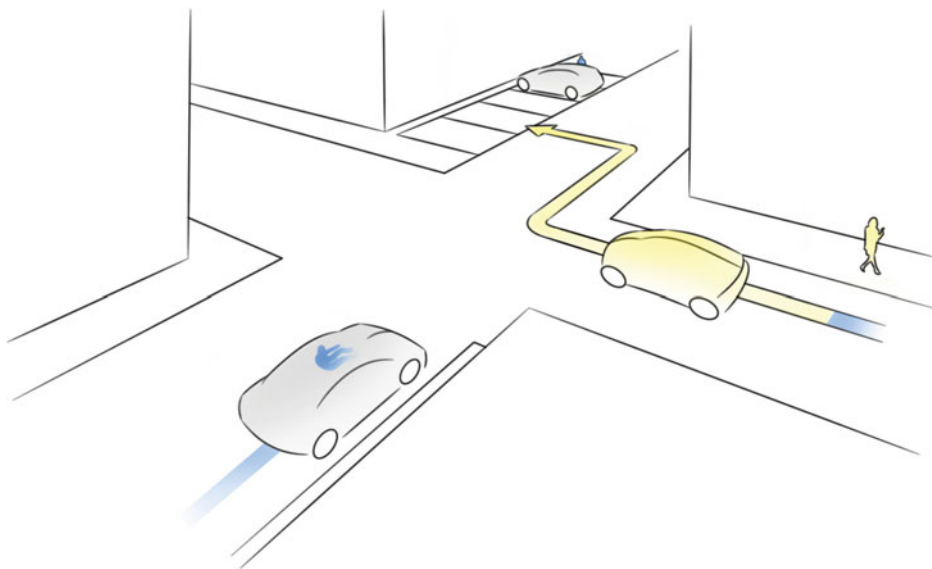


Fig. 2.3 Autonomous valet parking

2.3.2.2 Description

If a driver has reached his/her destination (for example place of work, gym, or home), he/she stops the vehicle, exits, and orders the driving robot to park the vehicle. The vehicle can be privately owned, but might also be owned by a carsharing provider or similar business model. Therefore, the driving robot may now drive the vehicle to a private, public, or service-provider-owned parking lot. It is important to assign a parking lot to the driving robot. The search for the respective parking lot by the driving robot is not taken into consideration for this use case. Therefore a defined destination for the driving robot is always given. Because of the low velocity and the light traffic situation, the deployment of Autonomous Valet Parking is limited to the immediate vicinity of the location where the driver left the vehicle. On the one hand, this limitation reduces the requirements regarding the (driving-) capabilities of the driving robot significantly, because lower kinetic energy as well as shorter stopping distances results from lower velocity. On the other hand, this use case could potentially irritate or frustrate other road users. However, this use case seems to be suitable as an introductory scenario.

An authorized user in the vicinity of the vehicle can indicate a pick-up location to the driving robot. The driving robot drives the vehicle to the target destination and stops, so that the driver can enter and take over the driving task.

If desired by the parking lot administration, the driving robot can re-park the vehicle.

2.3.2.3 Values of Characteristics

Table 2.2 summarizes the characteristics for the autonomous valet parking use case. The entities which can intervene into the driving task are depicted on the right side of the hierarchy and are sorted from dominant at the top to recessive at the bottom (see Fig. 2.4). The vehicle user can change the driving mission from outside of the vehicle and instruct the driving robot to perform a safe exit. The service provider overrules the vehicle user and can

Table 2.2 Values of characteristics for autonomous valet parking

Characteristic		Value		
A	Type of occupant	1	No cargo and no person	
B	Maximum permitted gross weight	1–5	500 kg to 8 t	
C	Maximum deployment velocity	2	Up to 30 km/h	
D	Scenery	3 a–5 a	Parking lot or parking structure, access roads, built-up main traffic roads	Without permission allowed
E	Dynamic elements	1	Without exclusion	
F	Information flow between driving robot and other entities	1 and 3 and 6	Navigation optimization, control optimization monitoring the driving robot	
G	Availability concept	1	No availability addition	
H	Extension concept	2	Driver	
I	Options for intervention		Figure 2.4: autonomous valet parking options for intervention	

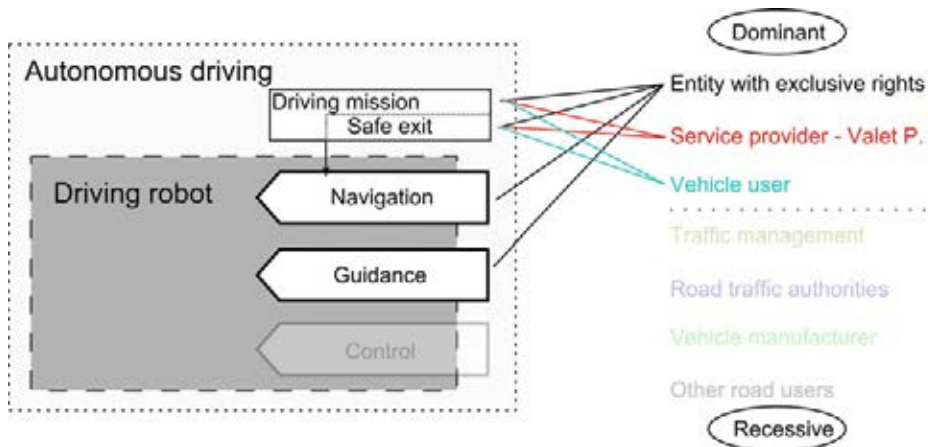


Fig. 2.4 Autonomous valet parking options for intervention

also influence the driving mission and the safe exit. Both entities are overruled by the entities with exclusive rights. For example, the police or ambulance can decelerate the vehicle on the guidance level, change navigation and driving mission, and order a safe exit.

2.3.3 Full Automation Using Driver for Extended Availability

An exemplary use case of the full automation using driver for extended availability is depicted in Fig. 2.5.

2.3.3.1 Benefit

If the driver desires to do so, he/she hands over the driving task to the driving robot in permitted areas. The driver becomes just a passenger during the autonomous journey, can take his/her hands off of the steering wheel and pedals, and can pursue other activities.

2.3.3.2 Description

If the driver desires, he/she can always hand over the driving task to the driving robot, whenever the current scenery is cleared to do so. Almost the entire traffic area in the permitted country is approved for the vehicle; however, such approval is subject to restrictions. If, for instance, the traffic flow is rerouted, a new parking structure opens, or similar changes are undertaken to the infrastructure, then the respective areas cannot be navigated autonomously until further approval. It also appears to be reasonable in this scenario that road sections are excluded from approval permanently or temporarily, e.g. roads with a high frequency of pedestrians crossing. Here again, the handover between driver and driving robot has to be managed in a safe manner.

This use case might come as close as it gets to today's visions for autonomous driving, as it corresponds strongly with today's passenger vehicle usage, and the driving task is

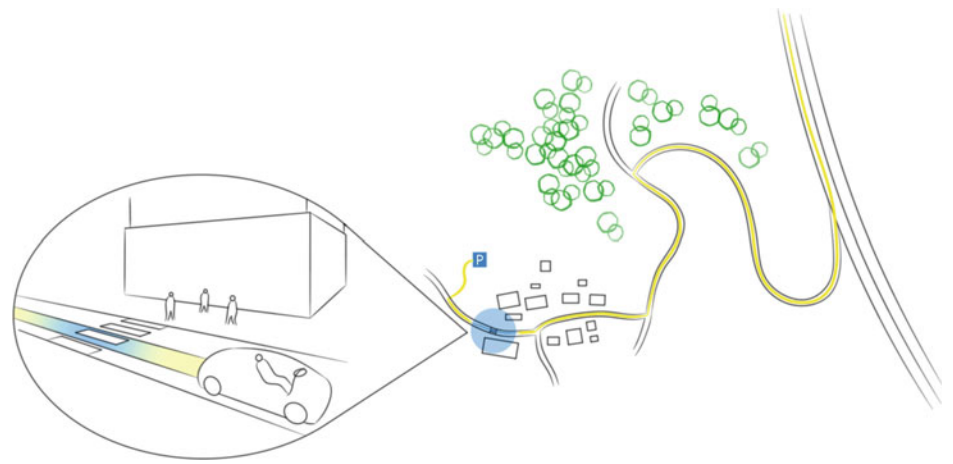


Fig. 2.5 Full automation using driver for extended availability

almost completely delegated to the driving robot while the traditional main user and driver still participate in the journey.

2.3.3.3 Values of Characteristics

Table 2.3 summarizes the characteristics for the full automation using driver for extended availability use case. Figure 2.6 shows, which entity (right) intervenes with a certain driving task (left) on a certain level. If desired, the vehicle user can drive the vehicle the

Table 2.3 Values of characteristics for full automation using driver for extended availability

Characteristic			Value	
A	Type of occupant	1.	Person/s with agreed destinations	
B	Maximum permitted gross weight	1–2	500 kg to 2 t	
C	Maximum deployment velocity	5	Up to 240 km/h	
D	Scenery	2 b– 8 b	Non-standardized road, parking lot or parking structure, access roads, built up main traffic roads, urban arterial road, country road, interstate	Only with permission allowed
E	Dynamic elements	1	Without exclusion	
F	Information flow between driving robot and other entities	1–6	Navigation optimization, guidance optimization, control optimization, provision of environmental information, updating the driving robot’s capability, monitoring the driving robot	
G	Availability concept	2	Availability through driver	
H	Extension concept	2	Driver	
I	Options for intervention		Figure 2.6: full automation using driver for extended availability options for intervention	

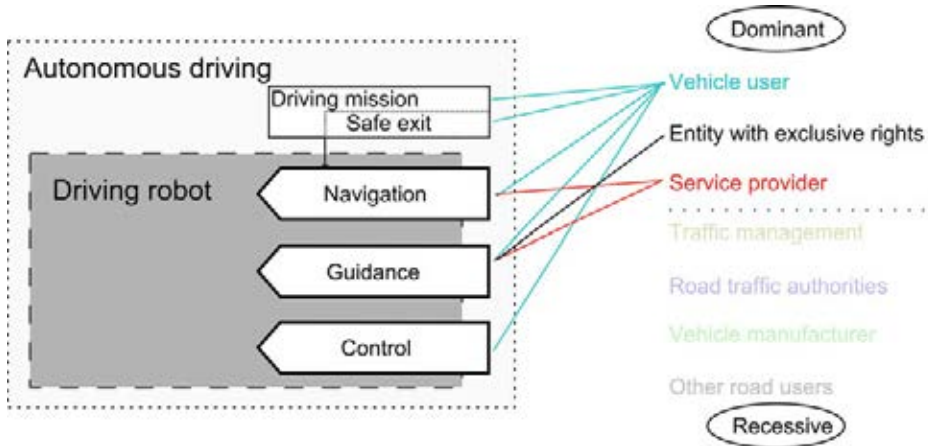


Fig. 2.6 Full automation using driver for extended availability options for intervention

same way as driving a classic driver only automobile, provided that the driving task has been handed over safely from the driving robot. Furthermore, the vehicle user can intervene on the level of the navigation, guidance and control tasks. The vehicle user dominates the entities with exclusive rights. The vehicle user can therefore overrule police or ambulance, which can exclusively intervene on the guidance level. The same is true for the service provider. The service provider can intervene on the navigation and guidance level, as long as not overruled by the vehicle user. It is left open in this document for which services the service provider needs access. Some concepts propose services where the service provider takes over the navigation for commercial use and partly pays for fuel and travel expenses.

2.3.4 Vehicle on Demand

An exemplary use case of the vehicle on demand is depicted in Fig. 2.7.

2.3.4.1 Benefit

The driving robot drives the vehicle autonomously in all scenarios with occupants, with cargo, but also completely without any payload. The driving robot makes the vehicle available at any requested location. Passengers use the travel time completely independently for other activities than performing the driving task. The cabin is designed completely independently from any restrictions of a driver workplace whatsoever. Cargo can be transported with the aid of the driving robot continuously for 24 h a day, as long as it is not restricted by the energy supply for driving.



Fig. 2.7 Vehicle on demand (scenery marked *red* is not part of the operating area)

2.3.4.2 Description

The driving robot receives the requested destination from occupants or external entities (users, service provider, etc.), to which the vehicle proceeds autonomously. Humans do not have any option to take over the driving task. The human can only indicate the destination or activate the safe exit, so that he/she can exit the vehicle as quickly as possible. With this driving robot, a wealth of different business models is conceivable. A mix of taxi service and car sharing, autonomous cargo vehicles or even usage models that goes beyond the pure transportation task. One example could be a vehicle for social networks that uses information from the network directly in order to plan routes, match people or enables further services which have not yet been thought of.

2.3.4.3 Values of Characteristics

Table 2.4 summarizes the characteristics for the vehicle on demand use case. The possibilities for intervention regarding the use case vehicle on demand are especially broad (see Fig. 2.8), due to the enormous (driving) abilities of the driving robot. The driving robot always carries out the control level. An entity with exclusive rights (e.g. police or ambulance) and the entity for traffic management can both intervene on navigation and guidance levels. Vehicle users and service providers can influence the safe exit and therefore instruct the driving robot to a fast and safe stop in order for a passenger to leave the vehicle. It is especially noticeable that service providers and the authority with exclusive rights can overrule the vehicle user. If one authority overrules the user, he/she cannot perform the safe exit anymore and has to stay in the vehicle. This constellation is

Table 2.4 Values of characteristics for Vehicle on Demand

Characteristic			Value	
A	Type of occupant	1–4	No cargo and no person, for transportation approved cargo, person/s with agreed destinations, persons with non-agreed destinations	
B	Maximum permitted gross weight	1–3	From 500 kg to 8 t	
C	Maximum deployment velocity	4	Up to 120 km/h	
D	Scenery	2 a–8 a	Non-standardized road, parking lot or parking structure, access roads, built up main traffic roads, urban arterial road, country road, interstate	Without permission allowed
E	Dynamic elements in the scenery	1	Without exclusion	
F	Information flow between driving robot and other entities	1–8	Navigation optimization, guidance optimization, control optimization, provision of environmental information, updating the driving robot’s capability, monitoring the driving robot, monitoring occupants, occupant emergency call	
G	Availability concept	3	Tele-operated driving	
H	Extension concept	1	No substitute	
I	Options for intervention		Figure 2.8: vehicle on demand options for intervention	

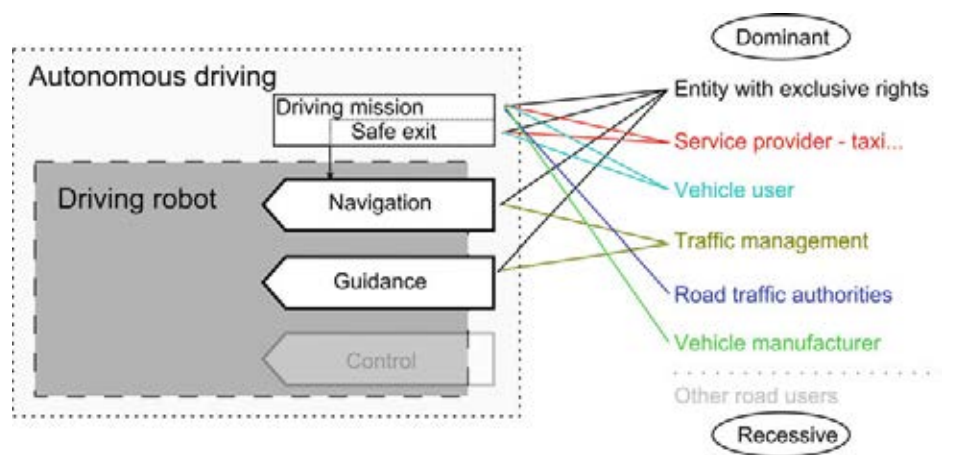


Fig. 2.8 Vehicle on demand options for intervention

similar to that of current taxi concepts. The taxi driver can stop as fast as possible, if the passenger so requests. Generally though, he (the taxi driver) also has the possibility to disregard this request and drive the vehicle as he/she own desires.

2.4 Selected Characteristics to Describe the Use Cases

In this section, the characteristics describing the use cases and their values are explained in more detail. Besides the following few technical characteristics of autonomous driving, it is possible to define further distinguishing attributes, for example regarding the business model or market position. This will be disregarded for now because of the as-yet little knowledge in this area.

The characteristics, in alphabetical order A to I, were derived from the three-level-model for the driving task according to Donges [3] and chosen for the description. In that model, the driving task is divided into the three levels *navigation*, *guidance*, and *control*.

2.4.1 Characteristic A: Type of Occupant

2.4.1.1 Motivation

For today's individual mobility with a vehicle, a human is required to be permanently in the vehicle and to control it under all circumstances [4]. This constraint could change with the automation of the driving task. Thus the vehicle concept and the safety concept depend on the *type of occupant*.

2.4.1.2 Values of the Characteristic

Here, the following values are distinguished:

1. no cargo and no persons, therefore no specific occupant or cargo protection interests
2. cargo approved for transportation
3. person/s with agreed destinations
4. persons with non-agreed destinations.

One use case can be covered by several values of this characteristic. The distinction between value 3 and 4 is made in order to distinguish between individual and public transportation. A vehicle of individual transportation carries persons with agreed destinations. In contrast, a vehicle of public transportation carries multiple persons who have not previously agreed upon a destination. However, persons reach their destinations with public transportation, because a schedule with destinations and intermediate stops is established.

2.4.2 Characteristic B: Maximum Permitted Gross Weight

2.4.2.1 Motivation

The *maximum permitted gross weight* influences safety considerations via kinetic energy. Besides safety considerations, looking at gross weight extends the discussion beyond individual transportation to public transportation, freight transportation as well as road infrastructure. In addition, this characteristic addresses the question of vehicle types, which potentially are not compatible with current vehicle types because of the autonomous driving functions and changing requirements, on a high level. Instead of considering the boundaries of often country-specific vehicle classes, four mass attributes are chosen. They range in values from ultra-light vehicles to heavy trucks and each step spans a factor of 4 between types.

2.4.2.2 Values of the Characteristic

Discrete distinctions have been established in order to describe the imagined use cases and to roughly categorize their mass. An exact determination of the mass is possible for existing use cases and specified deployment. Characteristic B covers the following values:

1. ultra-light vehicles around 500 kg
2. passenger vehicle around 2 t
3. light commercial trucks and vans around 8 t
4. trucks around 32 t.

2.4.3 Characteristic C: Maximum Deployment Velocity

2.4.3.1 Motivation

The characteristic of *maximum deployment velocity* (to be precise the square of the velocity) determines, multiplied with the mass, the kinetic energy of a vehicle, and therefore also needs to be distinguished. In addition, stopping distance is calculated using the square of the velocity. Accordingly, the autonomous system's requirements regarding a risk-minimal state in case of failure or when reaching functional limitations grow with the velocity squared.

Besides safety considerations, travel time and the range achievable in a given time at a given deployment velocity are also values that influence individual mobility. In addition, the deployment velocity directly defines the road type which can be used if a minimum velocity is required for using it.

2.4.3.2 Values of Characteristic

The maximum deployment velocity, characteristic C, has five proxy values, one for walking speed, and four in steps with a factor of two (= factor 4 in terms of kinetic energy and stopping distance). For concrete use cases the values and regulations need to be adapted to the respective deployment. Discrete distinctions have been established in order to describe the imagined use cases and to roughly categorize their velocity. An exact determination of the velocity is possible for existing use cases and defined deployments.

1. up to 5 km/h
2. up to 30 km/h
3. up to 60 km/h
4. up to 120 km/h
5. up to 240 km/h.

2.4.4 Characteristic D: Scenery

2.4.4.1 Motivation

Which spatial areas accessible to the driver through the *driver-only* automobile will also be made accessible with the described use case of autonomous driving? The *scenery* characteristic describes the spatial deployment in which the vehicle drives autonomously. For instance, do standardized structures exist, how many lanes are available, and do other markings exist?

Even static scenery can be diverse and present a challenge for the driving robot. One example of this, as is often mentioned, is traffic lanes covered with snow, or traffic signs hidden by bushes or trees. Such conditions, which are potentially unknown and non-changeable at the beginning of a journey, will not be considered with this characteristic. Determining the extent to which the driving robot can deal with scenery and conditions rests on the assumption that the robot can accomplish the driving task as well as a human driver.

This characteristic therefore describes scenarios that are predictable and that follow existing rules on a high level (location, environment and function of the road).

2.4.4.2 Values of the Characteristic

Dimension: type of scenery

1. off-road terrain
2. agricultural road
3. parking lot or parking structure
4. access road

5. main traffic roads
6. urban arterial road
7. country road
8. interstate
9. special areas.

The scenery characteristic in its first dimension covers 9 values (the scenery types from the German guidelines for integrated network design [5] were expanded) (Table 2.5):

Besides this value that describes the scenery within which a specific use case can be performed, the characteristic has a second dimension, which is the condition whether access to the scenery has to be permitted explicitly or not. The respective values are the following:

- a. *Without permission allowed*: All sceneries of this kind are permitted for driving robot operation.
- b. *Only with permission allowed*: Only selected and permitted sceneries of this kind permit a driving robot to operate autonomously in this area.

For now, it is left open who grants this permission and whether that is a private or public administration. In that sense, the type of permission is not further specified, for example the infrastructure could be in maintenance mode or a map could be provided, enriched with additional information. And also, the permission could include a temporary component and statistical or dynamic cutoff times for specific scenery areas.

Table 2.5 Scenery values description

Value	Description
Terrain (off-road)	<ul style="list-style-type: none"> – Without standardized or known structures such as lanes or other markings – Without apparent traffic coordination – Not paved for driving
Agricultural road	<ul style="list-style-type: none"> – Covers rural roads and similar roads with mainly simple pavement – Is public – Respective traffic rules (e.g. StVO in Germany) apply
Parking lot or parking structure	<ul style="list-style-type: none"> – Explicitly designated and marked for parking vehicles. Markings are not always present for lanes, but standardized marking of the area for the coordinated parking of vehicles exist – Especially in urban areas, parking structures with several levels have at times narrow ramps and little space for maneuvering – Respective traffic rules (e.g. StVO in Germany) apply

(continued)

Table 2.5 (continued)

Value	Description
Access road	<ul style="list-style-type: none"> – Developed roads within developed areas, which primarily serve direct access to the developed properties or serve for general accommodation – Access neighborhoods characterized by residential, commercial and business – Generally single lanes and connected by intersections without traffic lights – Connection with developed main traffic roads, are realized through intersections with or without traffic lights or roundabouts – In special cases they serve public transportation – Mainly open and used by inner-community bike traffic – Respective traffic rules (e.g. StVO in Germany) apply
Urban arterial road	<ul style="list-style-type: none"> – Roads without direct connections or within developed areas – Generally serve a connecting function (connection roads) – Widely spaced buildings often characterize the sides of these roads with facilities for tertiary use, which is why the development remains low – The roads are single or double lane, which are mainly connected by intersections with traffic lights or roundabouts to the remaining road network – Respective traffic rules (e.g. StVO in Germany) apply
Country road	<ul style="list-style-type: none"> – Include single lane roads situated outside developed areas – Includes also short road sections with two lanes, which are single lane roads in the regular case – Connection with roads of the same category is generally realized through intersections or interchanges of different kinds – Respective traffic rules (e.g. StVO in Germany) apply
Interstate	<ul style="list-style-type: none"> – Include non-developed, two-lane roads that are connected with interchanges of different kinds – Run outside, in the perimeter of, or within developed areas and are exclusively used by fast road traffic – Access only possible by special connecting elements like onramps – Respective traffic rules (e.g. StVO in Germany) apply
Special areas	<ul style="list-style-type: none"> – Not open to the public – Their geometry is unknown – General public traffic rules (e.g. StVO in Germany) do not apply – For example an extensive private terrain or industrial facility both indoor and outdoor – The area can have additional infrastructure for autonomous driving, such as a container port with autonomous systems for loading and unloading as well as commissioning

2.4.5 Characteristic E: Dynamic Elements

2.4.5.1 Motivation

Besides the scenery, the complexity of a scene depends largely on dynamic elements. The dynamic elements in the scene with the autonomously driving vehicle extend the requirements on the driving abilities of the driving robot. Therefore this characteristic describes to what extent the use case can be deployed in the current traffic situation and if limitations or exclusions for the dynamic elements are considered.

2.4.5.2 Values of the Characteristic

Four values of the characteristic are distinguished (Table 2.6):

1. without exclusion
2. only motor vehicles
3. only autonomously driving vehicles
4. no other dynamic elements.

The exclusion of other dynamic elements for the values 2–4 is not determined in an absolute way. The scene on a contemporary interstate is described for instance through value (2) *only motor vehicles*. However, while the situation that one person or cyclist steps on the interstate applies in theory, it is disregarded here due to the respective probability of occurrence. According to the assumption in Sect. 2.2 that most likely there will be a mixed operation, only the values 1 and 2 will be used for the use cases.

Table 2.6 Dynamic elements values description

Value	Description
Without exclusion	<ul style="list-style-type: none"> – The most complex scene – Animals, pedestrians, cyclists, vehicles, law enforcement, etc. meet the autonomously driving vehicle in the scene
Only motor vehicles	<ul style="list-style-type: none"> – Interaction of autonomous vehicles and human controlled motor vehicles – Animals, pedestrians, cyclist etc. are excluded
Only autonomously driving vehicles	<ul style="list-style-type: none"> – A scenery exclusive for autonomously moving vehicles
No other dynamic elements	<ul style="list-style-type: none"> – Area exclusive for ONE autonomously driving vehicle

2.4.6 Characteristic F: Information Flow Between the Driving Robot and Other Entities

2.4.6.1 Motivation

As described in Sect. 2.5, the driving robot carries out the tasks of perception, cognition, behavior decision and behavior execution. To do so, information about the state of the vehicle driven by the robot is required, such as position and velocity, but also information about the environment and occupants. This information is derived either from sensors, reading from memory systems, or through communication. How and which information is exchanged between the driving robot and respective entities is defined by the purpose of the information flow. In order to describe the information flow for one use case, the purposes of information exchange are assigned to the use cases.

The availability of the information, its transmission, and the communication partner all have to be suitable for the deployment purpose. As already mentioned, it is additionally assumed that the technology is only introduced onto the market slowly. Therefore not all dynamic elements in the vicinity are able to participate in the information exchange, so that a mixed operation has to be assumed.

The information flow of the driving robot considered herein is a subset of the entire information flow of the vehicle. We shall for the moment disregard purposes that are part of infotainment and convenience systems. Current news, access to social networks, or music streaming may as specific services increase the additional benefit of the autonomous journey; however, the information flow of these services is not primarily relevant for autonomous driving. Therefore only purposes impacting traffic safety, traffic efficiency, as well as purposes that are potentially prerequisites for the autonomous journey, are described as distinguishing attributes.

2.4.6.2 Values of the Characteristic

Eight purposes of the information flow are distinguished (Table 2.7):

1. navigation optimization
2. path-tracking optimization
3. control optimization
4. provision of environmental information
5. updating the driving robot's capability
6. monitoring the driving robot
7. monitoring occupants
8. occupant emergency call.

The first three values might also lead to interactions to negotiate the temporal or spatial usage of the traffic infrastructure. For now this interaction is disregarded.

Table 2.7 Information flow between driving robot and other entities values description

Value	Description
Navigation optimization	<ul style="list-style-type: none"> – Information such as current position, route destination, flow velocity, weather, etc. are exchanged with an inter-regional traffic center. Inter-regional in this context means that the information relevant for navigation lies within the coverage area (several hundred kilometers) of the traffic central unit – Goals of the optimization are, for example, low energy consumption and CO₂-emission, a travel time or travel distance that is as short as possible
Path-tracking optimization	<ul style="list-style-type: none"> – Extensive information about the state (x, v, a, ...) and intention of the vehicle driven by a robot as well as information of the vehicles in the immediate vicinity are exchanged – Information regarding weather, road condition, congestion, road closures, and phase timing of traffic lights are shared with a local traffic center. Local in this context means a coverage area of a few kilometers around the vehicle – The goal is, for example, synchronized drive in lateral as well as longitudinal directions (platooning, intersections without signage, or adaptive lanes...)
Control optimization	<ul style="list-style-type: none"> – The vehicle states selected and the intentions of the driving robot, road users, and further elements in the immediate vicinity of the vehicle are exchanged – The goal is collision avoidance in lateral and longitudinal direction with one or several vehicles in the immediate vicinity, according to already existing V2X concepts
Provision of environmental information	<ul style="list-style-type: none"> – Information about the vehicle environment, which is perceived by the driving robot, is shared with road users as well as with a traffic center in the immediate vicinity – The goal is to serve an optimized map with information as a source for positioning, hazard recognition, navigation, etc.
Updating the driving robot's capability	<ul style="list-style-type: none"> – The manufacturer provides an update, which improves the (driving-) capabilities of the driving robot
Monitoring the driving robot	<ul style="list-style-type: none"> – Information about the status, capabilities, and intentions of the driving robot are shared with authorized entities – The goal is to secure evidence (event data recording) to reconstruct the course of an accident, similar to a black box in aviation – Malfunctions and hazardous situations that are identified through self-diagnosis are transmitted to the manufacturer

(continued)

Table 2.7 (continued)

Value	Description
Monitoring occupants	<ul style="list-style-type: none">– Information (video, audio, heart rate...) about the occupant, which characterize his/her condition, are shared with an emergency call center or a service provider– The goal is to monitor the health and safety of the occupant– Information will be forwarded to authorized receivers without the intention and action of the occupant
Occupant emergency call	<ul style="list-style-type: none">– If the occupant experiences an emergency related either to him-/herself or the autonomous journey, it is possible to contact an emergency call center or the service provider of the autonomous journey– Occupant initiates contact and shares information voluntarily

2.4.7 Characteristic G: Availability Concept

2.4.7.1 Motivation

During normal operation the driving robot controls the vehicle within the permitted area. If the driving robot reaches a generally non-predictable functional limitation, the driving robot hands over to a specified availability concept. This availability concept defines how to continue the driving mission. Such functional limitations can be unknown obstacles on the road, which no longer permit a continuation within the autonomy of decision-making. An example for such an obstacle is a branch extending to the road, so that the vehicle needs to touch the branch in order to continue the journey. The extent to which the availability concept takes over the entire driving task, or just takes over the decision-making, is left open intentionally.

2.4.7.2 Values of the Characteristic

The following *availability concepts* are distinguished (Table 2.8):

1. no additional availability
2. availability through driver
3. tele-operated driving
4. pilot service
5. electric towing.

The handover from the driving robot to the alternative availability concept is to be implemented risk-minimally. The driving robot transfers the vehicle for the handover to that risk-minimal state which is suitable for the transfer to the availability concept.

The respective interfaces for the availability through driver, remote control, a pilot, or towing need to be available.

Table 2.8 Availability concept values description

Value	Description
No availability addition	– The driving robot waits until, through external influence, the scene becomes negotiable again and is covered by the specification of the driving robot
Availability through driver	– One occupant supports the driving robot negotiating the scene – Left open as to whether this is by taking over the driving task or through maneuver commands
Tele-operated driving	– A service provider supports the driving robot negotiating the scene via remote control
Pilot service	– An especially trained person proceeds to the vehicle and supports the driving robot negotiating the scene
Electric towing	– If the hardware necessary for the control task is operational, a tow vehicle with a direct connection can operate it in order to support the driving robot in negotiating the scene

2.4.8 Characteristic H: Extension Concept

2.4.8.1 Motivation

Not necessarily all areas necessary for the transportation task will be covered with the help of autonomous driving, especially not at the beginning of its introduction. Subdomains will remain which cannot be controlled autonomously. Nevertheless, in order to fulfill the mobility needs of customers, portions outside the regime of automated driving can be covered with *extension concepts*. The extension concept describes whether and with what aid it becomes possible to perform the vehicle control outside the area specified for autonomous driving.

2.4.8.2 Values of the Characteristic

Characteristic H has 5 values (Table 2.9):

1. No substitute beyond the operating area, i.e. the autonomous driving area covers the specified transportation tasks completely. The vehicle with this value is an exclusive-autonomous vehicle. If the deployment also covers the entire deployment of current vehicles, it is a fully autonomous vehicle.
2. Driver: A human takes over the driving task.
3. Tele-operated driving: The driving task is performed by an external operator.
4. Pilot service: An especially trained person takes over the driving task in a specific regime.
5. Extra transportation device: At the boundaries of deployment, the driving robot coordinates the handover of the vehicle to an extra transportation device so that this transportation device can continue the transportation task. Possible examples would be the long-distance transport of urban vehicles with the help of a *road train* or a concept similar to an electronic tow-bar.

Table 2.9 Extension concept values description

Value	Description
No substitute	<ul style="list-style-type: none"> – There is no substitute beyond the operating area, i.e. the autonomous driving area covers the specified transportation tasks completely. The vehicle with this value is an exclusively autonomous vehicle – The deployment also covers the entire deployment of current vehicles, it is a fully autonomous vehicle
Driver	– A human takes over the driving task
Tele-operated driving	– The driving task is performed by an external operator
Pilot service	– An especially trained person takes over the driving task in a specific regime
Extra transportation device	<ul style="list-style-type: none"> – At the boundaries of deployment the driving robot coordinates the handover of the vehicle to an extra transportation device so that this transportation device can continue the transportation task – Possible examples would be the long-distance transport of urban vehicles with the help of a road train or a concept similar to an electronic tow-bar

If the driver is considered (the *driver* value), it is inevitably necessary that a vehicle control interface (driver workplace) is available. In addition it is assumed that a capable person holding a driver's license is an occupant for the journey outside the autonomous driving area. For other cases, values that are futuristic from today's perspective (*tele-operated driving* as well as *pilot service*), a necessary service/interface needs to be provided for these alternatives.

2.4.9 Characteristic I: Options for Intervention

2.4.9.1 Motivation

According to Donges [3], the three primary driving tasks of *navigation*, *guidance*, and *stabilization* need to be fulfilled in order to guide a vehicle to the desired destination of the journey. Löper and Flemisch [6] as well as others replaces *stabilization* by the term *control*. The task of control covers the stabilization and in addition vehicle control in situations of vehicle-dynamics instability. Therefore, the primary driving tasks will be *navigation*, *guidance*, and *control*.

According to the definition of fully automated driving, this driving task is transferred completely to the driving robot. When a destination is indicated to the driving robot, it fulfills the *navigation*, *guidance*, and *control* tasks and guides the vehicle to the desired destination. Although the driving robot will execute these tasks, the internal system architecture is not necessarily structured in such a way.

In contrast, with the exception of hazardous situations (electronic stability control, anti-lock braking system, automated emergency braking), the driver is in control of

current production vehicles (overwrite capability). The human fulfills the driving tasks at the driver workplace in the vehicle. Thus he/she currently has the option to correct the actions of assistance systems, i.e. to override them.

Therefore there are two entities, the occupant as well as the driving robot, which basically have the capability to control the vehicle.

In addition, ideas and concepts for remote vehicle operation (tele-operated) exist, in which entities external to the vehicle intervene in the vehicle guidance. If a communication link as well as a respective interface for the world outside of the vehicle exists, these external entities also have the capability to influence the vehicle control. Therefore, in total three groups of entities—*internal*, *vehicle*, and *external*—can be distinguished which can intervene with the vehicle control during the autonomous journey.

To simplify the description of this characteristic, the occupants (adults, minors, people with limiting disability etc.) are summarized as the *internal* group. Influences outside the vehicle (law enforcement (e.g. police), registered vehicle owners (if not part of *internal* group), authorized agents etc.) are summarized as the *external* group.

If the entities are considered independently, the following questions regarding their options for intervention apply:

1. On which level of vehicle control does the entity have the *option* to intervene?
2. For which level of vehicle control does the entity have the *authorization* to intervene?

The first is answered via the vehicle concept of the use case. If the entity is supposed to have the option for intervention, an appropriate interface in the vehicle concept is provided to the entity.

The second question requires a statutory rule that defines which authorization is assigned to entities according to their properties and responsibilities. At this point it will not be further elaborated who sets and checks these rules, whether there is a driving test of some sort for the different levels, and if authorizations such as a driver's license or access codes are needed.

From this, the following combinations of options for interventions that the vehicle provides and the authorization for intervention that the entity possesses result:

- a. The vehicle concept offers the option for intervention on one of the three levels (navigation, guidance and control) and the entity is authorized to intervene on the same level of the driving task. Therefore the entity can intervene.
- b. The vehicle concept offers the option, but the entity is not authorized to intervene on one level. This situation correlates to a child that is in the driver's seat. For the use cases, it is assumed for this situation that the law for this situation regulates the intervention by the entity.
- c. The vehicle concept does not offer the option, but the entity is authorized to intervene on one level. This correlates to a driver in the back seat, who cannot intervene.
- d. The use case offers the option on one level, however the entity is authorized to intervene on a different level of the driving task. Also with this combination, the intervention is not permitted to the entity.

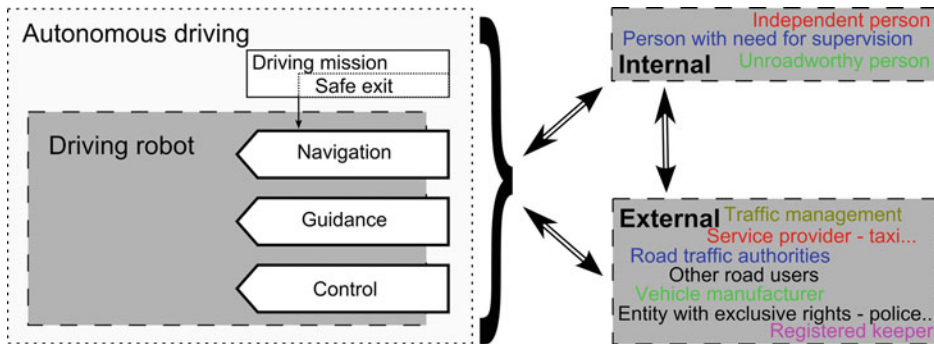


Fig. 2.9 Driving task conflict of interventions between entities

Only with combination (a) can the driving robot be influenced and/or overruled by the entity on one level of the driving task.

For the description of the use cases it follows that those entities are listed for which at least one authorization matches one available option of the vehicle concept.

In addition, it is assumed that statutory regulation will punish and therefore preclude misuse. This assumption also applies to current vehicle concepts. For example, it is not technology that prevents children from driving a vehicle, but rather the respective statutory regulation in combination with required supervision.

If the entities are now considered simultaneously and the entities are therefore able to act simultaneously on the three levels, the third question applies.

3. Which entity is dominant and how is the hierarchy of the entities defined in case of a conflict because of simultaneous interventions (Fig. 2.9)?

In order to answer this question for the description of the use cases, the intervention of the entities has to be attributed with a certain hierarchy. Which entity dominates others and thereby decides the vehicle behavior on the different levels of the driving task? A hierarchy of the entities needs to be implemented in the vehicle design.

In this it needs to be acknowledged that, in addition to the hierarchy of the entities, there also needs to be a hierarchy of the levels for the driving task. Control always overrules guidance and guidance always overrules navigation. Therefore it is additionally defined that internal or external entities can intervene only on one level. The entity with the highest priority suppresses other interventions.

Through autonomous driving it is also possible to exclusively transport persons who are not able to perform the driving task or to change the driving mission. However, in order to provide occupants with the option to exit safely as fast as possible, the safe exit is introduced as a special driving mission. If the occupant gains access to the safe exit with the highest priority, he/she might not necessarily be able to change the destination of the journey, but can exit the vehicle as fast as possible.

2.5 General Definitions

Some basic terms, which will be used in the following sections, are defined as follows: **navigation**—according to Donges [3], navigation includes choosing an appropriate driving route from the available road network as well as an estimation of the expected time requirement. If there is information about current interferences, such as accidents, road works or traffic jams, a change in route planning may be necessary.

guidance—according to Donges [3], the task of guidance is basically to derive the advisable command variables, such as the intended track and the set-point speed, from the road situation ahead as well as from the planned route. Part of the guidance is also to anticipatorily intervene the open-loop control to create favorable conditions for the lowest possible deviations between set and actual values.

stabilization (control)—to fulfill the stabilization task, according to Donges [3], the driver has to ensure with corrective actions that the deviations in the closed-loop control are stabilized and compensated to a level for which the driver is capable of handling.

driver only—ground (0) level of automation according to BAST [1]: “the driver continuously (throughout the complete trip) accomplishes longitudinal (accelerating/braking) and lateral (steering) control.”

assisted—the first (1) level of automation according to BAST [1]: “the driver continuously accomplishes either lateral or longitudinal control. The other/remaining task is accomplished by the automating system to a certain level only.

- The driver must permanently monitor the system.
- The driver must at any time be prepared to take over complete control of the vehicle.”

fully automated—the fourth (4) level of automation according to BAST [1]: “the system takes over lateral and longitudinal control completely within the individual specification of the application.

- The driver does not need to monitor the system.
- Before the specified limits of the application are reached, the system requests the driver to take over with a sufficient time buffer.
- In the absence of a takeover, the system will return to the minimal risk condition by itself.
- All system limits are detected by the system, the system is capable of returning to the minimum risk condition in all situations.”

autonomous driving—for autonomous driving, the driving task [3] is performed in a way that is called *fully automated* (level 4 automation according to BAST [1]). This definition is extended by the assumption that the machine behavior stays within an initially set behavioral framework.

machine driving capabilities—the machine (driving) capabilities are capabilities related to perception, cognition, behavior decisions as well as behavior execution.

driving robot—a driving robot is the implementation of the machine’s (driving) capabilities. The driving robot consists of hardware components (sensors, processors, and actuators) and software elements. It acts as the hardware and software, equivalent to the role of a driver in today’s vehicles as subject¹ (the definition term for this system is still incomplete, so alternative suggestions are welcomed).

fully autonomous vehicle—a fully autonomous vehicle is a vehicle which can drive almost all routes autonomously, on the same level as *driver-only* vehicles. This definition is beyond the BAST [1] definition as it defines the vehicle and not the degree of automation.

exclusively autonomous vehicle (autonomous-only vehicle)—an exclusively autonomous vehicle is a vehicle which can drive all routes for which the vehicle has been specified autonomously from start to destination. This definition is beyond the BAST [1] definition as it defines the vehicle and not the degree of automation.

transportation task—the driving task describes a defined transportation object (vehicle, cargo, passenger etc.) that is transported from one start location to a destination location. Examples of the transportation task include *park vehicle* or *get passenger to the requested destination*.

driving mission—the driving mission describes the journey from start to destination in execution of the transportation task.

safe exit—the safe exit is a special driving mission. It leads the vehicle in the fastest way to a system status which allows the passengers to safely exit the vehicle.

driver—a driver is the human operator of a vehicle, without further specifying the driving capability. This means within a range of humans who have a driver’s license. The driver is the subject of autonomy in case of non-fully automated driving.

scenery—the term scenery according to Geyer et al. [8] refers to the static environment of the vehicle. This takes into consideration the geometry of pre-defined road types, number of lanes, curvature, position of traffic signs and traffic lights, as well as additional stationary objects such as construction areas and natural (e.g. bushes and trees) or man-made objects (e.g. buildings, walls).

dynamic elements—dynamic elements according to Geyer et al. [8] are temporary and spatially variable elements such as other road users, states of traffic lights, light as well as traffic conditions.

scene—the scene according to Geyer et al. [8] is built by the scenery, dynamic elements and optional driving instructions. A scene starts with the end of the earlier scene or—in case of the first scene—with a defined starting scene. Within a scene the elements, their behavior as well as the position of the autonomously driving vehicle are defined. The dynamic elements change their states within a scene.

¹A system which is capable of taking decisions depending on sensor data processed internally has additional degrees of freedom as compared to one with direct sensor data to actuator feedback or one without any capability of control actuation. The former one is termed a ‘subject’, the last one an ‘object’... [7].

situation—a clear definition of the term situation for the use-case description is still to be determined. In particular, an “objective, complete situation (description)” has to be distinguished from a “subjective, projective situation (description)”.

operating area—a spatial and/or temporal area, specified explicitly via the scenery and implicitly via the velocity, in which the vehicle can be moved autonomously through the operation of the driving robot.

operating limit²—the operating limit is specified explicitly through the scenery and implicitly through the velocity, and is therefore a predictable boundary at which the driving task is handed over.

functional limit³—a condition that appears in the permitted operation range but is not predictable in detail, which contradicts with continuing the autonomous journey. Even if the limit is not foreseeable, the driving robot recognizes it at an early stage.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work’s Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work’s Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

References

1. Gasser, T.M., Arzt, C., Ayoubi, M., Bartels, A., Bürkle, L., Eier, J., Flemisch, F., Häcker, D., Hesse, T., Huber, W., Lotz, C., Maurer, M., Ruth-Schumacher, S., Schwarz, J., Vogt, W.: Rechtsfolgen zunehmender Fahrzeugautomatisierung. Gemeinsamer Schlussbericht der Projektgruppe. Berichte der Bundesanstalt für Strassenwesen - Fahrzeugtechnik (F), vol. 83. Wirtschaftsverl. NW Verl. für neue Wissenschaft, Bremerhaven (2012)
2. Based on the understanding of autonomy according to Immanuel Kant interpreted by Feil, E: Autonomie und Heteronomie nach Kant. Zur Klärung einer signifikanten Fehlinterpretation. Freiburger Zeitschrift für Philosophie und Theologie, 29/1-3, (1982), S. 389-441 (Printed in Feil, E. Antithetik neuzeitlicher Vernunft. „Autonomie – Heteronomie“ und „rational – irrational“, Göttingen 1, Teil I, S.25-112.)
3. Donges E.: Fahrerverhaltensmodelle. In: Winner H. et al. Handbuch Fahrerassistenzsysteme, 2. Auflage, Vieweg + Teubner Verlag, Wiesbaden, p 15-23, (2012)
4. Kempen B.: Fahrerassistenz und Wiener Weltabkommen. in 3. Sachverständigentag von TÜV und DEKRA: Mehr Sicherheit durch moderne Technologien, Berlin 25.-26. February (2008)
5. Kategorien der Verkehrswege für den Kfz-Verkehr (3.4.1) out of Richtlinien für integrierte Netzgestaltung Edition 2008 . Distinguished are Autobahn, Landstraße, anbaufreie Hauptverkehrsstraße, angebaute Hauptverkehrsstraße and Erschließungsstraße. The definitions are translated freely.

²The functional limit corresponds to the “Systemgrenze” of category one BASt [1]—Definition.

³The operational limit corresponds to the “Systemgrenze” of category two BASt [1]—Definition.

6. Löper C., Flemisch F. O.: Ein Baustein für hochautomatisiertes Fahren: Kooperative, manöverbasierte Automation in den Projekten H-Mode und HAVEit, 6. Workshop Fahrerassistenzsysteme, Hößlinsülz, (2009)
7. Dickmans E. D.: Subject-object discrimination in 4D dynamic scene interpretation for machine vision, Proc. IEEE-Workshop on Visual Motion (2009)
8. Geyer, S.; Baltzer, M.; Franz, B.; Hakuli, S.; Kauer, M.; Kienle, M.; Meier, S.; Weißgerber, T.; Bengler, K.; Bruder, R.; Flemisch, F. O.; Winner, H.: Concept and Development of a Unified Ontology for Generating Test and Use Case Catalogues for Assisted and Automated Vehicle Guidance, IET Intelligent Transport Systems, Accepted to publish (2013)

Download the free book

Autonomous Driving [:] *Technical, Legal and Social Aspects*, edited by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner.

at: <http://link.springer.com/book/10.1007%2F978-3-662-48847-8>

This book takes a look at fully automated, autonomous vehicles and discusses many open questions: How can autonomous vehicles be integrated into the current transportation system with diverse users and human drivers? Where do automated vehicles fall under current legal frameworks? What risks are associated with automation and how will society respond to these risks? How will the marketplace react to automated vehicles and what changes may be necessary for companies?

Experts from Germany and the United States define key societal, engineering, and mobility issues related to the automation of vehicles. They discuss the decisions programmers of automated vehicles must make to enable vehicles to perceive their environment, interact with other road users, and choose actions that may have ethical consequences. The authors further identify expectations and concerns that will form the basis for individual and societal acceptance of autonomous driving. While the safety benefits of such vehicles are tremendous, the authors demonstrate that these benefits will only be achieved if vehicles have an appropriate safety concept at the heart of their design. Realizing the potential of automated vehicles to reorganize traffic and transform mobility of people and goods requires similar care in the design of vehicles and networks. By covering all of these topics, the book aims to provide a current, comprehensive, and scientifically sound treatment of the emerging field of "autonomous driving".

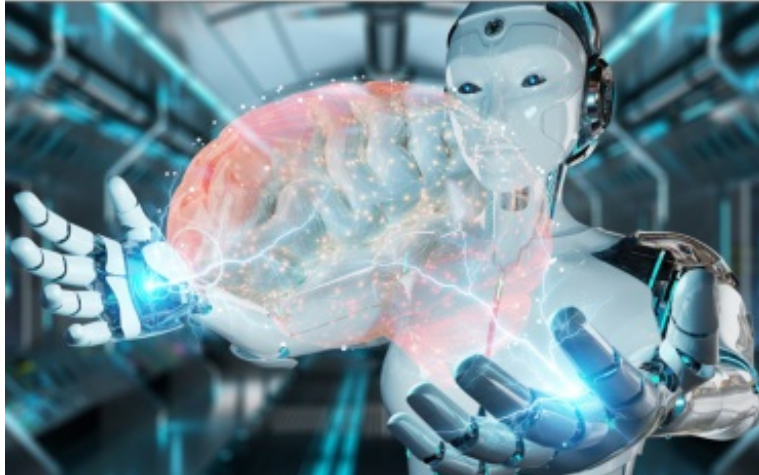
Keynote 2: R. Patrick Huston

Guest Post: BG Pat Huston on “Future War and Future Law”

sites.duke.edu/lawfire/2018/12/03/guest-post-bg-pat-huston-on-future-war-and-future-law/

Charlie Dunlap,
J.D.

December 3,
2018



Interested in how artificial intelligence (AI) and the law will operate together (or not) in future conflicts? If so, then today's guest post is exactly what you are looking for!

*U.S. Army judge advocate **Brig General Pat Huston** will share his perspective – using real-world examples – to provide some context for thinking about AI, and to give us some ideas as to where this technology might be headed.*

I hasten to add that Pat is no ordinary lawyer-warrior. As his bio below notes, he's completed five combat tours in Iraq and Afghanistan. That and the fact that he earned the coveted Ranger tab (and has early military service as a helicopter pilot) gives him a perspective few other lawyers can possibly achieve. His academic credentials are equally as formidable: a degree in engineering from West Point, a law degree from the University of Colorado at Boulder, a Master of Laws degree from the U.S. Army JAG School, and a Master of Strategic Studies degree from the U.S. Army War College. Pat is clearly one of the nation's premier lawyers.

Here is Pat's introductory note: *This essay is based on my remarks during the National Security Law Workshop co-hosted by the University of Texas Law School, the South Texas College of Law, and The Judge Advocate General's Legal Center and School. I sincerely thank Professors Bobby Chesney, Geoff Corn and Todd Huntley for their invitation to speak, and Professor Charlie Dunlap at Duke for his willingness to publish these comments.*

The long title for these remarks is “Future War and Future Law,” but the short title is simply “AI” because Artificial Intelligence is what this conversation is really all about. My goal is to provide you several real-world AI examples to help put this in context. I'll start with a

basic primer about this technology, and then discuss where this is likely headed. I've broken my comments down into 5 topics.

- (1) The rise of artificial intelligence in society.
- (2) The future of AI in the practice of law.
- (3) The future of war, particularly the role of autonomous weapons.
- (4) The future of International Humanitarian Law (IHL) or the Law of Armed Conflict (LOAC) in this context, and finally
- (5) A few legal, ethical, and practical considerations.

I'm not one to keep you in suspense, guessing about my conclusions, so let me start my giving you the top three takeaways from my presentation:

(1) Human Judgment: We must ensure that autonomous weapons allow commanders to exercise appropriate levels of human judgment. In other words, we need humans to make certain key decisions, and we can't unleash an autonomous weapon over which we expect to lose control.

(2) Accountability: All weapons use must comply with IHL, and commanders must remain responsible for all weapons they employ. These are fundamental principles.

(3) The importance of Government cooperation with Industry: The best and brightest AI researchers should insist on legal and ethical conduct for all military AI uses, and they should work with compliant governments (including the US) to this end. If they boycott military projects, the void will be filled by researchers who are less capable, less ethical, or both, and I think that would be a recipe for disaster.

(1) The rise of Artificial Intelligence in society.

So let's jump to the first of the five topics. Artificial Intelligence is not just science fiction like the Terminator movies or HBO's Westworld. AI is real, it is here now, and it is all around us. You know about self-driving cars, "Smart" thermostats, and robot vacuum cleaners, but not everyone uses those. But nearly everyone has done a Google search, which uses AI-enhanced algorithms to optimize the results. So do Amazon and Netflix. They make recommendations for you based on past purchases and past movies, and there are many other everyday interactions with AI, such as Siri, Alexa and Google Translate.

So what's next? Where is this technology headed? 30 years ago, IBM's Deep Blue computer program beat world chess champion Garry Kasparov. That too has evolved. There's a traditional Chinese board game called Go that is far more complex than Chess. Recently, Google's Deep Mind program – called Alpha Go — used a new type of AI to beat the world

champion Go player. This was huge because it shows the potential for a new level of machine learning, where the program starts off knowing nothing about the game, and teaches itself by playing billions of games to see what works and what doesn't. And as it learns, it rewrites its own code. This is the next generation of advanced AI, and is generally referred to as "Deep Learning."

AI-enabled technology is everywhere. But it's not just the US pursuing this technology. China has invested heavily in its AI research and development, and has established fusion centers with universities. China also declared its goal of becoming the global leader in AI and robotics by 2030. In short, AI is all around us and there is a global AI "Arms Race" underway.

(2) The Future of AI in Practice of Law.

Since this audience is mostly lawyers, I'll highlight AI's use in the practice of law simply to illustrate its impact in that context, but you need to understand that AI is also disrupting other industries similarly.

I'll start with some really simple AI uses in the legal profession. Most of you are aware of the speech recognition software used by many court reporters. Those of you involved in litigation – especially civil or commercial litigation – understand how eDiscovery is now being used extensively. AI-enhanced search tools can sort through terabytes of electronic records and e-mails and extract relevant and responsive documents far faster than any human attorney or paralegal. They can also tag documents as potential attorney work product or privileged, so that they're not automatically turned over. Government agencies can use this technology to perform Freedom of Information Act (FOIA) searches. This technology is so effective that some predict due diligence will soon require the use of AI tools in large discovery cases. In other words, failure to use this technology could be legal malpractice in some circumstances.

Some courts in the U.S. are using AI-enhanced tools to help make parole decisions – to evaluate convicted criminals for the likelihood of recidivism. That raises several questions, but think about that from a Due Process standpoint. I recently saw a report that courts in Buenos Aires are using AI to help them assess guilt or innocence, and then to draft their court opinions.

The entire history of our Supreme Court decisions has been fed into an AI system. This system has reportedly identified voting patterns of individual justices resulting in an 80% accuracy rate at predicting future court decisions. This is better than predictions by human SCOTUS experts who routinely practice before our high court.

This is one of my favorites: A study last year pitted British insurance lawyers against an AI system called CaseCrunch to evaluate insurance claims. The AI system was the clear winner. The lawyers had a 62% accuracy rate, and AI scored 87%. That's a D- versus a B+!

Three American universities ran a similar study this year. Stanford, Duke and USC had a lawyer versus AI competition to evaluate contracts for thirty different contract law issues. Again, the AI system won: 85% accuracy for lawyers and 95% for AI. But what was even more impressive than the accuracy was the speed. It took the average attorney 92 minutes to review those contracts. It took the AI system just 26 seconds.

If you're a client, would you rather pay a lawyer for an hour and a half of work that is 85% accurate, or a computer for a half-minute of work that's 95% accurate? The bottom line is that AI could rapidly change the practice of law in some areas and if we ignore these changes, we risk getting left behind.

(3) The Future of War, particularly the role of autonomous weapons.

Let's talk about war ... future war. The biggest looming issue is the use of fully autonomous weapons that unilaterally select and engage targets. And they could do this on a scale and at a speed that could overwhelm humans and render traditional warfare obsolete. This is not as far-fetched as you might think.

Secretary of Defense Jim Mattis is an avid military strategist and a retired Marine General who is not prone to exaggeration or alarm. He has carefully studied weapons developments and previously said "the fundamental nature of war that does not change." But earlier this year, he changed his tune. He's been studying the impact of AI on warfare and essentially concluded that AI is a game-changer. He said the potential development of fully autonomous weapons has caused him to question his entire premise about the future of war.

Now to be clear, most military AI projects are not controversial: smart maintenance scheduling for aircraft, self-driving supply trucks, robots instead of people on "bomb squads," pilotless search & rescue aircraft, tele-medicine to treat wounded soldiers in remote areas, etc. The Navy has several AI-enhanced training programs that have significantly improved the quality of language and other technical skills courses, while also reducing training times. This saves a lot of time and money. The Army has AI-enhanced combat training facilities that are extremely realistic. Secretary Mattis said that he wants every Soldier and Marine to fight twenty-five battles in these synthetic trainers before they ever step foot on a real battlefield. And all of the services use AI to enhance pilot training. This saves money, but more importantly, it saves lives by fostering realistic and effective training in a "safe" environment.

Intelligence is a perfect place for AI tools because it involves massive amounts of data that needs to be sifted through, and it's always time sensitive. Analyzing photos, video feeds, written reports, phone and radio intercepts, emails, texts, and social media accounts. And in many cases, this all needs to be translated first, before any analysis can be done. AI-enhanced software can translate and then analyze everything and predict what the enemy

will do next. This type of “predictive analysis” is the fundamental role of intelligence, and the place where AI can help. But there are concerns. If we are going to target someone based on a computer’s analysis of his threat or their role, how reliable is it? Are there biases?

OK, now let’s talk about weapons. DOD calls these Lethal Autonomous Weapons Systems, or LAWS. Human Right Watch calls them “Killer Robots,” which I admit is much catchier. There are offensive and defensive systems; semi-autonomous and fully-autonomous systems; systems with humans in- or on-the-loop (to monitor or abort) and those with no humans involved.

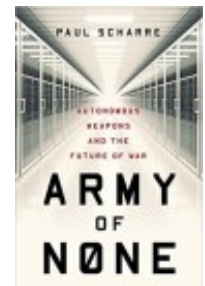
U.S. soldiers who have spent time in Iraq or Afghanistan are familiar with the “Counter Rocket, Artillery and Mortar” system. The C-RAM is a defensive autonomous system that scans the sky for incoming rounds. When it finds one, a loudspeaker – called “the big voice” — blares “Take Cover — Incoming rounds,” and it fires a machine gun to disable the incoming round while it’s in the air. This all happens in a few seconds, which is all the time you have with incoming rounds.



C-RAM

According to Paul Sharre, author of “Army of None,” the Navy has a more complex system called the Aegis to defend its ships against missiles and aircraft, but the basic concept is similar. South Korea has similar “Robot Sentries” on the DMZ. Israel has unmanned patrol boats off its coast. Many other countries have similar defensive autonomous systems to protect borders or military installations.

The US is testing Drone swarms to overwhelm enemy defenses. Initial indications are that those types of swarms can’t be countered without AI-enhanced defenses, so you can see the potential for rapid military escalation.



China and Russia are doing the same things. China has developed Military-Civilian AI Fusion Centers at its top universities. In Russia, President Putin announced that the nation which leads in AI will “be the ruler of the world.” And it’s clear that he intends for that to be Russia. Russia has several autonomous “Warbots” with names like Platform-M, Argo & Wolf-2, a driverless APC called the A-800 MARS and a fully autonomous tank called the URAN-9. I certainly think this trend will continue.

(4) The Future of International Humanitarian Law (IHL) or the Law of Armed Conflict (LOAC) in this context.

So what does this mean for the future of IHL or LOAC? The general consensus – and certainly my opinion — is that current LOAC construct and principles are sufficient to govern this emerging technology, including autonomous weapons. However, even though the legal framework is sufficient, it may be more complex and difficult to make the factual determinations in order to meet the legal standards.

Let's look at the two most important LOAC principles: distinction and proportionality. Starting with distinction: even if you confirm the target is a combatant, fratricide is always a concern, so you have to differentiate between friend and foe. This is not a new concept. During World War II, the Germans developed torpedoes that used sonar to track to enemy ships. These torpedoes were autonomous once launched. But two torpedoes did U-turns and sank the German U-boats that launched them.

But let's say the autonomous system does properly distinguish between friend and foe, and you know it has properly identified the enemy through advanced facial recognition technology that has access to pictures of every member of the enemy's armed forces. It must still go a step further to confirm that the enemy is targetable: that they're not waving a white flag or raising their hands in surrender; that they're not marked with a red cross; not wounded and out of combat. These are important steps that every commander knows to follow. We have to ensure that AI-enhanced systems follow these rules too.

And there are other vulnerabilities. Let's say you're relying on facial recognition technology to identify a specific enemy commander. How confident are you in the program's ability to get it right? Some MIT students spoofed an AI visual recognition program to consistently conclude that a plastic turtle was a rifle. In some ways, this is no different than the wooden decoy tanks we placed all over Britain before D-Day to fool the Germans about troop concentrations and where we would invade France. But because AI is all computer based, there are also cyber vulnerabilities at every turn, so it's just more complex and harder to predict what you know and what you don't know.



We've already said that data is the key to effective AI systems. We've all heard the expression that "garbage in equals garbage out." But war has always been marked by incomplete or inaccurate information. Plus there is often chaos and confusion, called the "fog of war." If all that wasn't enough, both sides are deliberately trying to deceive the other. Deception is an inherent part of any military operation.

Let's say you satisfy the distinction test, you also have to satisfy proportionality. In simple terms, commanders must assess — based on the information available at the time — that the expected collateral damage is not excessive compared to the expected military advantage. There's more room for uncertainty and subjectivity here. And the commander employing an autonomous system has to understand the system well enough to have

confidence in what it will do and what it won't do. This doesn't have to be a perfect prediction – legacy weapon systems and certainly humans can be unpredictable too – but in some cases, an AI system may be so unpredictable that the commander is unable to satisfy this requirement.

This basic principle of accountability or command responsibility can't be abdicated. Commanders are always responsible for the systems they employ, including autonomous weapons.

(5) A Few Legal, Ethical and Practical Considerations.

Let me tell you about the controversy involving Google, and the Campaign to Stop Killer Robots.

First, Google has a contract with the U.S. military to use AI tools to review and analyze video feeds from remotely-piloted aircraft, or drones. AI can do it faster and, in some cases, better than human intelligence analysts. But many Google employees – including some of the Nation's top AI researchers – were concerned that their work was contributing to military operations, and pressured Google to not extend the contract. Google announced a few months ago that they will not continue when the contract expires. They also announced a series of ethical principles that essentially say they will not participate in AI projects related to weapons.

Google won't renew its U.S. Defense Department contract in 2019



This led to a larger movement in July 2018, when more than 2,400 workers at over 160 technology companies signed a letter demanding laws banning lethal autonomous weapons. This parallels an effort called the “Campaign to Stop Killer Robots” led by Human Rights Watch and Harvard’s Human Rights Clinic. This campaign calls for an outright ban on developing or employing lethal autonomous weapons. These initiatives have led to significant public debate.

Let me state up front that I think these groups’ underlying concerns are valid. Nobody wants to see weapons that are unleashed on the world, or robots that turn on their creators like in Terminator. However, I don’t think their proposed solution is the most effective way to address the concerns. In fact, I think their proposed total boycott would significantly worsen the problem.

First, if industry’s best and brightest AI researchers and coders — who are concerned about the legal and ethical compliance of these new technologies — withdraw from the discussion, the void will be filled by others who are less capable, less ethical, and less law-abiding. If our

talented and ethics-minded coders just quit or bury their heads in the sand, it could be a recipe for disaster.

Second, I don't think a boycott will slow Russian or Chinese efforts. They do not appear to be nearly as concerned as the U.S. government about the legal or ethical implications of autonomous weapons.

Third, a boycott seems to ignore the likelihood that AI can in some cases produce better results than humans. This creates ethical obligations, or at least raises significant ethical questions. For example, what if AI can produce more precise weapons that reduce collateral damage and cause less suffering? Would we be legally required to leverage that technology? Morally or ethically required? This goes back to the evolving eDiscovery standards which may require us to use technology that outperforms humans.

I'm also concerned about the risk of escalation. Most AI systems operate at incredible speeds. In the Wall Street context, rapid AI-enhanced stock market trading caused several "Flash Crashes." Do we need some sort of ethical governor, circuit breaker, or kill switch on these systems to prevent a rapid escalation of autonomous weapons to avoid a "Flash War"? I think the answer is "yes," and we should all want our most talented developers solving this problem.

Finally, I'll note that the current U.S. military policy (DoD Directive 3000.09) includes a provision for a human role in the process, which I think is important. It requires that all autonomous systems "allow commanders ... to exercise appropriate levels of human judgment over the use of force" and states that these systems can only be employed when all LOAC principles are satisfied. And I think these standards strike the right balance.

So in conclusion, I want to circle back to, and end with, my three key takeaways:

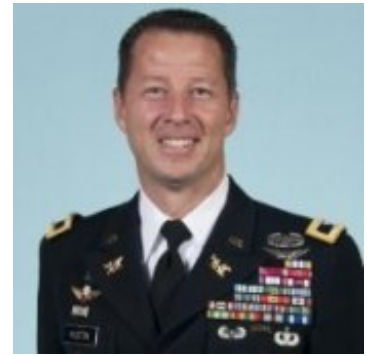
(1) Human Judgment: We must ensure that autonomous weapons allow commanders to exercise appropriate levels of human judgment. In other words, we need humans to make certain key decisions, and we can't unleash an autonomous weapon over which we expect to lose control.

(2) Accountability: All weapons use must comply with IHL, and commanders must remain responsible for all weapons they employ. These are fundamental principles.

(3) The importance of government cooperation with industry: The best and brightest AI researchers should insist on legal and ethical conduct for all military AI uses, and they should work with compliant governments (including the US) to this end. If they boycott military projects, the void will be filled by researchers who are less capable, less ethical, or both, and I think that would be a recipe for disaster.

Thank you all for your time and attention.

Brigadier General Pat Huston is the Commanding General of The Judge Advocate General's Legal Center and School in Charlottesville, Virginia. His current work focuses on the legal and ethical development and use of artificial intelligence, autonomous weapons, and other emerging technologies. General Huston has completed five combat tours in Iraq and Afghanistan, and was the General Counsel (Staff Judge Advocate) of three major defense organizations: the 101st Airborne Division, the Joint Special Operations Command (JSOC), and the U.S. Central Command (CENTCOM).



BG Huston

The views expressed here are those of the author and do not reflect the official position of The Judge Advocate General's Legal Center and School, the United States Army, or the Department of Defense.

*As we like to say at **Lawfire®**, check the facts, assess the arguments, and decide for yourself!*

Introduction to the Legal Issues of AI & Robotics

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

Introduction to the Legal Issues of AI and Robotics

January 9, 2020 / 9:00 a.m.



THE PREMIER SOURCE FOR CLE

Introduction to the Legal Issues of AI & Robotics

January 9, 2020 / 9:00 a.m

Speakers

- Jimmy Kim (Business Operations Lead, Built Robotics)
- Giancarlo Mori (CEO, Myvyl Technologies)
- Christopher Savoie (CEO and Founder, Zapata Computing, Inc.)
- Stephen Wu (Shareholder, Silicon Valley Law Group)
[Moderator]

AI's Impact

“I believe it's going to change the world more than anything in the history of mankind. More than electricity.”

Kai-Fu Lee

On CBS 60 Minutes

Source: 60 Minutes segment at
<http://bit.ly/PLI2019-60Minutes>

AI's Impact

- AI “is the dominant technology of the future.”
- “Gaining access to considerably greater intelligence would be the biggest event in human history.”

Stuart Russell
Human Compatible

What is Artificial Intelligence?

What are Robots? What is Robotics?

Exponential Growth in Computing Power

Legal Issues of AI & Robotics

Industry-Specific Legal Issues

- Key examples of AI- and robotics-impacted fields
 - Transportation
 - Healthcare
 - Financial services
- Future programming to cover
 - Entertainment
 - Oil, gas, and clean energy
 - Real estate
 - Insurance
 - Manufacturing

Cross-Cutting Legal Issues

- Intellectual property protection
- Employment and workplace safety
- Data protection (privacy and security)
- Civil and human rights
- Investigations and evidence law
- Professional responsibility

Role of Lawyers with Other Disciplines

- Lawyers and businesspeople
- Lawyers and technologists
- Lawyers and philosophers (Public Good panel)
- Lawyers and auditors (Standards, Certifications, and Audits panel)

Top Legal Issues for Business Lawyers

- Compliance
- Transactions
- Liability and litigation
- Incident response and investigations
- Governance (e.g., policies and procedures)

Top Legal Themes for the Institute

- Safety and robustness: harm and liability
- What is the standard of care?
- The "human element"/interface with systems
- Human and individual rights
- Data protection compliance and risk management
- Worker replacement and training
- Transparency v. "black box"
- Bias

Top Legal Themes for the Institute

- Machine autonomy and agency
- The “control problem”
- The role of lawyers
- Ethical, legal, and social implications (ELSI)
- Interdisciplinary collaboration
- Lawyers and collaborators acting as part of something “larger than ourselves”

Questions?

- Jimmy Kim | jimmy@builtrobotics.com
- Giancarlo Mori | Giancarlo@movyl.us
- Christopher Savoie | cjs@zapatacomputing.com

- Stephen Wu | ssw@svlg.com



The Law of Artificial Intelligence and Smart Machines

Understanding A.I. and the Legal Impact

Theodore F. Claypoole, Editor

ABA
AMERICAN BAR ASSOCIATION
Business Law Section

AI Product Liability Issues and Associated Risk Management

By Stephen S. Wu

I. INTRODUCTION

On March 18, 2018, an Uber test car driving in autonomous mode, with a safety driver in the vehicle, struck and killed a woman walking a bicycle across a street in Tempe, Arizona. Volvo and Uber worked together on the autonomous driving system controlling the car. After an accident like that, we could reasonably anticipate that the estate of the pedestrian could bring a product suit against Uber, Volvo, or both. As it turns out, within days after the accident, Uber settled with the family of victim.¹ Nonetheless, accidents like these are likely to lead to the question of liability for robots, artificial intelligence systems, and autonomous vehicles (AVs). The news media are full of stories asking who would face liability for accidents involving robots and AVs.

Robots and AI systems hold the promise of bringing widespread and extensive benefits to society. Robots will increasingly have the ability to perform the dull, dirty, or dangerous work that humans perform today, saving humans from boredom, health problems, injuries, or even death. As just one example, AVs may be able to save tens of thousands of lives each year in the United States, and many more worldwide, reducing traffic, saving energy, and providing mobility to those who cannot drive conventional cars. Nonetheless, robots, AI systems, and AVs will inevitably have some accidents. On balance, they are likely to prevent many more accidents and much more harm than they cause, but there will be at least some accidents involving these technologies that would not have occurred with human-controlled systems.

Robots and AVs in particular act in the physical world. Accidents involving these systems are inevitable. Some of these accidents will cause catastrophic injury for those

1. Faiz Siddiqui, *Uber reaches settlement with family of victim killed after being struck by one of its self-driving vehicles*, WASH. POST, Mar. 29, 2018, available at <https://www.washingtonpost.com/news/dr-gridlock/wp/2018/03/29/uber-reaches-settlement-with-family-of-victim-killed-after-being-struck-by-one-of-its-self-driving-vehicles/>.

involved in the accident. Even worse, if a defect or cyber attack could compromise every instance of a particular robot or an entire network, fleet, or industry, the defect or attack could cause widespread simultaneous accidents throughout the country or even the world. Imagine, for instance, a future in which regional transportation centers in metropolitan centers control the dispatch and navigation of AVs in the region. Imagine further that a sudden defect causes all the AVs under control of the system to crash all at once in a major metropolitan area like New York. The impact of such an event in terms of harm, property damage, injury, and deaths could easily exceed an event like the attacks on September 11, 2001.

In 2012, I had the opportunity to speak at the Driverless Car Summit presented by the Association of Unmanned Vehicle Systems International. The conference organizers polled the audience which, although admittedly unscientific, did provide a data point about industry views on product liability. One polling question asked attendees to identify the chief obstacle to the deployment of AVs, and the top answer was “legal issues.” The proceedings of the conference identified this issue as well.² Although the poll did not break down the issues among compliance and liability, I suspect that liability is the larger perceived issue. Indeed, some people have identified product liability suits as an existential threat to autonomous driving.³

In the worst-case scenario for the industry, manufacturers could face numerous suits that force some of them to exit the robotics market and cause others to decide not to enter the market in the first place. They could perceive that the sales are not worth the risk. Such an outcome could be tragic if it results in manufacturers not bringing otherwise life-saving and socially beneficial robots to the market. Manufacturers, however, can implement practices to minimize the likelihood, frequency, and magnitude of accidents, and thereby control the risk of liability. By implementing these practices, manufacturers can maintain the profitability they would need to offer robots in the market.

This chapter discusses the nature of product liability and identifies liability risk for manufacturers of robots, AI systems, and AVs. It covers the sources of product liability and explains how manufacturers can manage these risks. Section 1.2 talks about why product liability suits occur and, in particular, the cause of huge verdicts. Section 1.3 describes how high profile product liability cases impose human and financial impacts. Section 1.4 covers product liability law, focusing on the types of

2. E.g., Autonomous Solutions Inc., 5 Key Takeaways From AUVSI’s Driverless Car Summit 2012 (Jul. 12, 2012) (“Some of the largest obstacles to autonomous consumer vehicles are the legalities.”). Reports from Lloyd’s of London and the University of Texas listed product liability as among the top obstacles for AVs. Lloyd’s, *Autonomous Vehicles Handing Over Control: Opportunities and Risks for Insurance* 8 (2014) [hereinafter, “Lloyd’s Paper”]; University of Texas, *Autonomous Vehicles in Texas* 5 (2014).

3. See, e.g., Tim Worstall, *When Should Your Driverless Car From Google Be Allowed To Kill You?*, FORBES, Jun. 18, 2014 (“the worst outcome would be that said liability isn’t sorted out so that we never do get the mass manufacturing and adoption of driverless cars”), available at <http://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/>.

claims and defenses arising in product liability cases. Section 1.5 examines risk management in design practices and procedures that manufacturers can employ. Section 1.5 includes a discussion of how insurance can help manage risk by transferring risk to insurance carriers, which also promote a number of risk management practices.

II. WHY DO PRODUCT LIABILITY SUITS OCCUR?

Product liability suits frequently follow accidents involving the product in question. For instance, accidents on the roads and highways of the country played a prominent role in the development of American tort law. “Products liability, like America, grew up with the automobile. Prior to the entry of motorcars onto the nation’s highways, ‘there simply were not large numbers of product-related lawsuits.’ Once America embraced the automobile, it inevitably embraced automotive products suits as well.”⁴ In other words, once manufacturers began selling automobiles, and the public began using them on mass scale, accidents occurred, and product liability suits inevitably followed.

One early road accident case in the United States was in reaction to an important British case⁵ about the scope of tort liability: *Winterbottom v. Wright*.⁶ *Winterbottom* involved a road accident in which a mail coach driver was thrown from his horse-drawn mail carriage after it collapsed. The plaintiff’s claim was that the defendant contractor failed to maintain the carriage in a safe condition. The court denied relief to the injured coachman because of a lack of privity between the plaintiff coachman and the defendant contractor. The coachman was not a party to the contract in which the defendant contractor promised to maintain the coach in good working order. Therefore, no duty from that contract could run in favor of the coachman and, in the court’s view, a recovery by parties in the position of the plaintiff would be too broad.⁷

In the 20th century, car accidents led to the development of case law changing the nature of U.S. product liability law. Two key cases in product liability arose for car accidents. First, the court in *MacPherson v. Buick Motor Co.*⁸ rejected privity as a defense and affirmed a verdict for a car owner ejected from his Buick car in an accident caused by a defective wooden wheel on the car that collapsed.⁹ Judge Cardozo’s opinion for the court held that although the plaintiff had bought the car

4. 1 LOUIS R. FRUMER & MELVIN I. FRIEDMAN, PRODUCTS LIABILITY § 1.02 (2018) [hereinafter, “FRUMER”] (footnote omitted).

5. It was common in the 19th century for American courts to cite contemporary British cases as precedents.

6. *Winterbottom v. Wright*, 152 Eng. Rep. 402 (1842).

7. *See id.* at 404-06.

8. 217 N.Y. 382, 111 N.E. 1050 (1916).

9. The car was apparently going 8 miles per hour at the time of the accident. FRUMER, *supra* note 4, § 1.02.

from a retailer, but could still sue the manufacturer despite the lack of a contract with the manufacturer. The court recognized that, to the manufacturer, probable harm from a defective automobile is foreseeable and the manufacturer knew that the car would be used by persons other than the buyer. Given these facts, a manufacturer putting a defective product on the market should be liable for negligence, regardless of the presence or absence of a contract.¹⁰

Likewise, in the car accident case of *Henningsen v. Bloomfield Motors, Inc.*,¹¹ the New Jersey Supreme Court rejected privity, a warranty disclaimer, and limits of liability as defenses to the warranty claim of the wife, driver of the car, and her husband, the purchaser of the car. The wife of the purchaser incurred the accident. She testified at trial that she felt something crack in the car, the steering wheel spun sharply, the car veered off the road, and the car struck a highway sign and brick wall. The court affirmed a jury verdict against the manufacturer Chrysler and a dealer.

Once robots and AI systems are widely deployed and used, we can expect a similar growth and perhaps explosion of product liability suits. With robots and AI systems in common usage, accidents and other mishaps are bound to occur, and litigation will inevitably follow. As with the automobile and other products, product liability suits will occur because of accidents and the resulting damage.

Not only is litigation inevitable, but manufacturers face a risk of enormous verdicts against them. Some of the cases that will arise in the future may result from mass catastrophes or numerous smaller accidents causing many deaths and catastrophic injuries. Counsel for injured plaintiffs can point to these mass accidents and present testimony in unvarnished and sometimes horrific terms. To use an analogy, imagine an opening statement in a case like the famous Ford Pinto case. After the accident, a writer¹² described the accident in stark and gruesome terms:

[A] woman, whom for legal reasons we will call Sandra Gillespie, pulled onto a Minneapolis highway in her new Ford Pinto. Riding with her was a young boy, whom we'll call Robbie Carlton. As she entered a merge lane, Sandra Gillespie's car stalled. Another car rear-ended hers at an impact speed of 28 miles per hour. The Pinto's gas tank ruptured. Vapors from it mixed quickly with the air in the passenger compartment. A spark ignited the mixture and the car exploded in a ball of fire. Sandra died in agony a few hours later in an emergency hospital. Her passenger, 13-year-old Robbie Carlton, is still alive; he has just come home from another futile operation aimed at grafting a new ear and nose from skin on the few unscarred portions of his badly burned body.¹³

10. See MacPherson, 217 N.Y. at 389-91.

11. 32 N.J. 358, 161 A.2d 69 (1960).

12. The writer is using pseudonyms for the names of the crash victims, evidently before the names of the victims became public.

13. Mark Dowie, *Pinto Madness*, MOTHER JONES, Sept. 1, 1977, available at <http://www.motherjones.com/print/15405>.

The 13-year-old boy in the real Ford Pinto case, named Richard Grimshaw, did in fact prevail at trial and received a jury award of over \$2.5 million in compensatory damages. To punish and deter Ford, the jury awarded punitive damages in the amount of \$125 million.¹⁴ One of the factors behind the very large verdict was testimony of Ford's decision to use inexpensive parts in the Pinto instead of more expensive parts that could have prevented the catastrophic explosion following the accident. The jury apparently concluded that Ford placed profits above the safety of its customers and the public. The punitive damages award was later reduced in this case to \$3.5 million.¹⁵ Nonetheless, the Ford Pinto case shows the kind of large verdict that could occur in a product liability case after a catastrophic accident.

Fast forwarding to the future, robot and AI system manufacturers will likely face product liability litigation. If manufacturers decide to try these cases, we can expect that some cases will involve shockingly disfigured plaintiffs injured by out-of-control robots after tragic, frightful accidents. Just as the writer described the Ford Pinto accident in horrific terms, a plaintiff's lawyer could describe the fate of a client injured by a robot. In the courtroom setting, the attorney could point to the victim client, present throughout the trial, badly maimed and perhaps suffering through the trial. The jury would be watching the victim the whole trial.

In assessing its risk in front of the jury, any defendant manufacturer in such a case would need to consider the catastrophic nature of the plaintiff's injuries and the risk that a jury, even one trying to be impartial, cannot help but feel at least some unconscious sympathy for the plaintiff. In this fraught setting, the trial would place the defendant manufacturer's engineering and business practices under exacting scrutiny. It would be up to the jury to determine whether the manufacturer's design decisions could have avoided the accident.

During the design phase of any robot or AI system, the manufacturer's design team will make many engineering and business decisions leading to the final design of a product. The team may have opportunities to implement different safety features but adding additional safety features will likely increase the cost of the product and likely reduce its profitability. When deciding whether to implement these features, design teams can think more clearly and assess risk more effectively by imagining themselves in a courtroom setting, defending their practices in litigation arising from a catastrophic accident.

Why do juries award very large verdicts against manufacturers? The short answer is juror anger. "Angry jurors mean high damages."¹⁶ Juries award very large verdicts when they become angry at a manufacturer based on what the manufacturer did or didn't do in designing the product. When juries become angry at a manufacturer,

14. See *Grimshaw v. Ford Motor Co.*, 119 Cal. App. 3d 757, 771-72 (1981).

15. *Grimshaw*, 119 Cal. App. 3d at 823-24.

16. Robert D. Minick & Dorothy K. Kagehiro, *Understanding Juror Emotions: Anger Management in the Courtroom*, FOR THE DEFENSE, July 2004, at 3 (emphasis added), available at http://www.krollontrack.com/publications/tg_forthedefense_robertminick-dorothyhagehiro070104.pdf.

they see a punitive damages award as a way of sending a message to the manufacturer to say that its conduct is unacceptable.

The jury in the Ford Pinto case heard testimony that Ford was aware of the problems with the fuel system during the design of the car. In the actual case, the rear-end accident split open the Pinto's gas tank, allowing an explosion to occur. Ford had identified the problem and also knew that a part costing \$11 could have prevented the damage to the fuel system. Ford made a cost/benefit comparing the overall cost of adding the \$11 safety part to the vehicle against the value of the lives lost from accidents involving the vulnerability. In its analysis, Ford assigned a value to each human life likely lost. In the end, Ford decided that the cost of using the part surpassed the overall value of the human lives that would have been saved. As a result, it decided not to add the \$11 part to the Pinto's design.

To the jury, Ford's cost/benefit calculation seemed cold-hearted. Ford placed a dollar value on human life. And the price of the part, \$11, seemed small in comparison with the cost of the car. The jury evidently concluded that Ford was more interested in maximizing its profits than in protecting human life. The apparent callous disregard for safety led to the jury anger.¹⁷

The role of jury anger was apparent in another famous product liability case involving the painkiller Vioxx. A jury awarded \$253.5 million verdict against pharmaceutical company Merck to a widow of someone who used Vioxx for eight months and suffered a fatal heart attack.¹⁸ At trial, the jury saw internal Merck documents making it clear that Merck was well aware of Vioxx's heart attack risk but Merck went ahead with sales of the drug anyway. The documents showed the jury that Merck placed profits above the safety of its customers. The jury apparently became angry and used the verdict to send Merck a message that it is wrong to conceal information about a drug's risks.¹⁹ While an appellate court later overturned the jury's verdict,²⁰ the case shows the risks to manufacturers that treat public safety lightly.

III. MORE RECENT HIGH-PROFILE PRODUCT LIABILITY LITIGATION

Although the Ford and Vioxx cases reduced or overturned the large verdicts, the lesson from these cases is that accidents and resulting product liability can lead to large financial consequences for manufacturers. More recent cases underscore the human and financial impact of alleged product defects leading to numerous accidents.

17. Grimshaw, 119 Cal. App. 3d at 813.

18. Alex Berenson, *Vioxx Verdict Raises Profile of Texas Lawyer*, N.Y. TIMES, Aug. 22, 2005, available at <http://www.nytimes.com/2005/08/22/business/22lawyer.html?pagewanted=all>.

19. Lisa Girion and Dana Calvo, *Merck Loses Vioxx Case*, L.A. TIMES, Aug. 20, 2005, available at <http://articles.latimes.com/2005/aug/20/business/fi-vioxx20>.

20. *Merck & Co., Inc. v. Ernst*, 296 S.W.3d 81 (Tex. Ct. App. 2009), *cert. denied*, 132 S. Ct. 1980 (2012).

Section 1.3.1 describes the alleged “sudden acceleration” phenomenon that affected Toyotas. Section 1.3.2 explains the consequences of recent General Motors ignition switch defect problems. In addition to the human toll of deaths and injuries, these two sets of cases show how manufacturers and/or their insurers had to pay huge sums to resolve product liability litigation.

A. Toyota “Sudden Acceleration” Litigation

Starting in the 2000s, media reports about a phenomenon in which Toyota and Lexus drivers said that their cars would suddenly speed up without warning and would not stop. Some of these incidents resulted in accidents and injuries. One typical news report stated, “Nancy Bernstein feels lucky to be alive after her Toyota Prius kept accelerating, no matter how hard she hit the brakes. ‘The car’s going about 70 miles an hour, and I’m beginning to get scared because it’s not slowing down,’ Bernstein described.”²¹ Injured plaintiffs filed suit, and federal cases were transferred to the U.S. District Court for the Central District of California for coordinated or consolidated pretrial proceedings.²²

According to some reports, the Toyota sudden acceleration phenomenon may have claimed the lives of 89 people.²³ Governmental investigations, however, found no evidence that design or implementation defects in the Toyota cars caused the unintended acceleration phenomenon.²⁴ Consequently, the existence of an actual unintended acceleration phenomenon and the role driver error may have played in these accidents are still not certain.

During the course of the ongoing litigation, one key development was the completion of a report by expert witness Michael Barr. Barr concluded, based on his research, that a software malfunction caused the sudden acceleration phenomenon.²⁵ Barr pointed to numerous problems with the software. Moreover, according to Barr,

21. Ric Romero, Sudden Acceleration Issue Spans Beyond Toyota, 6ABC.COM (Jan. 4, 2010), <http://6abc.com/archive/7200572/>.

22. In re Toyota Motor Corp. Unintended Acceleration Marketing, Sales Practices, and Products Liability Litigation, No. 8:10-ML-2151 JVS (FMO) (C.D. Cal. filed Apr. 12, 2010) (cases consolidated by the Judicial Panel on Multidistrict Litigation in the U.S. District Court for the Central District of California).

23. CBS News and Associated Press, Toyota “Unintended Acceleration” Has Killed 89, CBS NEWS (May 25, 2010), <http://www.cbsnews.com/news/toyota-unintended-acceleration-has-killed-89/>.

24. National Highway Traffic Safety Administration, Technical Assessment of Toyota Electronic Throttle Control (ETC) Systems 57-60, 62-64 (Feb. 2011); see NASA Engineering and Safety Center, Technical Support to the National Highway Traffic Safety Administration (NHTSA) on the Reported Toyota Motor Corporation (TMC) Unintended Acceleration Investigation 170-172 (Jan. 18, 2011).

25. Michael Barr, Rule 26(a)(2)(B) Report of Michael Barr April 12, 2013 in Estate of Ida St. John v. Toyota Motor Corporation, et al. 65 (Apr. 12, 2013) [hereinafter, “Barr Report”].

Toyota's own engineers couldn't understand the company's own software. Due to its complexity, these engineers referred to the Toyota code as "spaghetti like."²⁶

Barr presented his findings in testimony during a trial in an Oklahoma state court case. Apparently, based in part of Barr's testimony, the jury awarded damages of \$1.5 million to the driver and \$1.5 million to the family of a passenger who died in the crash.²⁷ After this initial award, the parties in the case, *Bookout v. Toyota Motor Corp.*,²⁸ settled the case before a second phase of the trial for assessing punitive damages against Toyota.²⁹

Even though Toyota could have continued disputing liability in other cases, Toyota began settling the remaining actions against the company. The *Bookout* case may have led Toyota to conclude that settlement was the best course of action.³⁰ According to news reports, Toyota had to pay:

- \$1.6 billion to settle financial loss claims in the multidistrict litigation.³¹
- \$1.2 billion to settle potential criminal charges against Toyota.³²
- \$25.5 million to settle shareholder claims arising the failure to report safety issues.³³
- \$65 million in fines for violations of federal vehicle safety laws.³⁴

These settlement payouts are in addition to other product liability cases. The final amounts to resolve the litigation may have exceeded \$4 billion. The legal fees and other expenses related to investigation and remedial measures are in addition.

B. General Motors Ignition Switch Issues and Recall

In addition to the Toyota sudden acceleration litigation, problems with ignition switches used by General Motors show the financial cost of product liability litigation.

26. *Id.* at 39, 54.

27. Junko Yoshida, *Toyota Case: Single Bit Flip That Killed*, EE TIMES, Oct. 25, 2013.

28. No. CJ-2008-7969 (Okl. Dist. Ct. Okl. Cty. dismissed Nov. 20, 2013).

29. Jerry Hirsch, *After losing verdict, Toyota settles in sudden acceleration case*, L.A. TIMES, Oct. 25, 2013, available at <http://articles.latimes.com/2013/oct/25/autos/la-fi-hy-toyota-settles-sudden-acceleration-20131025>.

30. "Legal analysts said that the verdict most likely spurred Toyota to pursue a broad settlement of its remaining cases." Jaclyn Trop, *Toyota Seeks a Settlement for Sudden Acceleration Cases*, N.Y. TIMES, Dec. 13, 2013, available at http://www.nytimes.com/2013/12/14/business/toyota-seeks-settlement-for-lawsuits.html?_r=0.

31. *Id.*

32. David Undercoffler, *Toyota and Justice Department said to reach \$1.2 billion settlement in criminal case*, L.A. TIMES, Mar. 18, 2014, available at <http://www.latimes.com/business/autos/la-fi-hy-autos-toyota-justice-department-settlement-20140318-story.html>.

33. *Id.*

34. *Id.*

GM started to use new types of switches in small cars starting in the late 1990s. The idea was to make the switches work more smoothly. “But as it turns out, new switches in models such as the Chevrolet Cobalt and Saturn Ion can unexpectedly slip from ‘run’ to ‘accessory,’ causing engines to stall. That shuts off the power steering, making cars harder to control, and disables air bags in crashes.”³⁵ The issue may have caused over 50 accidents. “GM says the problem has caused at least 13 deaths, but some members of Congress put the death toll near 100.”³⁶

As with the Ford Pinto, GM engineers knew of the problems with the switches before accidents occurred. They decided nonetheless to use the new switches. A Congressional probe uncovered an internal GM email saying that a better switch design would add 90 cents to the price of the switch, but would only save 10–15 cents in reduced warranty claims.³⁷ “The part costs less than \$10 wholesale. The fix takes less than an hour. A mechanic removes a few screws and connectors, takes off a plastic shroud, pops in the new switch, and the customer is back on the road.”³⁸ “[T]o many people familiar with the automaker,” GM did not recall the cars sooner due to “a corporate culture reluctant to pass along bad news. When GM was struggling to cut costs and buff its image, a recall of its popular small cars would have been a terrible setback.”³⁹ “It’s pretty clear that somebody somewhere was being penny-wise and pound-foolish,” said Marina Whitman, a professor at the University of Michigan and a former economist at GM.”⁴⁰

As a result of GM’s decision not to recall the cars sooner, GM and/or its insurers incurred huge expenses. It had to respond to investigations by Congress, safety regulators, the U.S. attorney in New York City, the SEC, Transport Canada, and almost all state attorneys general. GM did end up having to recall the cars. Also, GM created a compensation fund, which was expected to cost \$400 to \$600 million to compensate families of crash victims.⁴¹

In addition to the compensation fund, GM said that it will spend \$1.2 billion to repair the cars and trucks recalled during the second quarter, on top of the \$1.3 billion it identified for repair costs in the first three months of the year.

35. Tom Krisher, *GM’s ignition switch: what went wrong*, COLUMBUS DISPATCH, Jul. 8, 2014, available at <http://www.dispatch.com/content/stories/business/2014/07/08/gms-ignition-switch-what-went-wrong.html>.

36. *Id.*

37. Email from John Hendler to Lori Queen et al. (Sept. 28, 2005, 4:07 pm), <http://docs.house.gov/meetings/IF/IF02/20140618/102345/HHRG-113-IF02-20140618-SD036.pdf>.

38. Michael Fletcher & Steven Mufson, *Why did GM take so long to respond to deadly defect? Corporate culture may hold answer*, WASHINGTON POST, Mar. 30, 2014, available at http://www.washingtonpost.com/business/economy/why-did-gm-take-...swer/2014/03/30/5c366f6c-b691-11e3-b84e-897d3d12b816_story.html.

39. *Id.*

40. *Id.*

41. Chris Isidore, *GM to pay victims at least \$400 million*, CNN MONEY, Jul. 24, 2014, available at <http://money.cnn.com/2014/07/24/news/companies/gm-earnings-recall/>.

In addition, the company set aside an additional \$874 million in the quarter for future recalls.⁴²

“All told, GM’s recalls have cost the automaker nearly \$4 billion”⁴³ in 2014. The final total will be beyond that amount to resolve litigation and governmental investigations. These amounts are in addition to the legal fees and other internal expenses it incurred in connection with the investigations and remedial measures.

As is apparent from the Toyota and GM examples, not only do accidents take a significant toll in injury and deaths, the resulting product liability can lead to huge expenses. We can expect the same for robots and AI systems. While not every product is as widely used or causes as much damage as a car, manufacturers are aware the product liability suits can lead to huge awards and expenses. Indeed, product liability suits can end entire businesses or industries.

IV. CLAIMS AND DEFENSES IN PRODUCT LIABILITY CASES

The previous sections covered the human and financial toll of product liability. This section covers what a plaintiff must plead and prove in a case involving an allegedly defective product. It also discusses potential defenses for defendants. In product liability cases arising out of accidents, plaintiffs assert claims for strict product liability, and breach of warranty.⁴⁴ State law typically governs the liability of product manufacturers for allegedly defective products, and laws vary by state.⁴⁵

A. Strict Product Liability Claims

Strict product liability claims are the easiest tort claim for a plaintiff to prove, as they require no showing of fault on the part of the defendant or privity with a manufacturer. Potential defendants include almost every business in the chain of distribution from raw materials or component part manufacturers to manufacturers of the finished product, distributors, and retailers.⁴⁶ Different states recognize strict

42. *Id.*

43. *Id.*

44. Another theory of recovery for plaintiffs is fraud, also known as “deceit” or “misrepresentation,” and is based on false statements made by the seller about a product. Misrepresentations may be intentional, negligent (careless), or innocent. This type of claim, however, is the least used theory of recovery in the product liability context. 1 FRUMER, *supra* note 4, § 2.05[1].

45. For a survey of case law bearing on robot and autonomous vehicle product liability, see Stephen S. Wu, *Unmanned Vehicles and US Product Liability Law*, J. OF LAW, INFORMATION & SCIENCE, 2011/2012, at 234.

46. 1 FRUMER, *supra* note 4, § 5.01.

liability in different ways, although some states do not recognize strict liability as a cause of action. Some states have statutes establishing strict liability and its scope, while others recognize strict liability as a common law development under Section 402A of the Second Restatement of Torts.⁴⁷ Under Section 402A, the essential elements of a strict liability claim are:

- The defendant sold the product in question,
- The defendant is in the business of selling this kind of product,
- The product was defective and unreasonably dangerous,
- At the time it left the defendant's hands,
- And the product is expected to and does reach the user or consumer without substantial change in the condition in which it is sold, and
- The defect was the proximate cause of
- Damages to the plaintiff.⁴⁸

A strict liability case may arise from a defect in the design of a product, in the manufacture of the product in question, or in the failure to warn or instruct a plaintiff about the product. A design defect claim rests on the design of all instances of a product. In the Ford Pinto case, for example, the plaintiffs said that all Ford Pintos of the model in question had a defect in the design of the fuel system. By contrast, under a manufacturing defect theory, a plaintiff alleges that the one product involved in that accident had a problem. There may be nothing wrong with the design, and all the other products of that model may have no defects, but something went wrong with the manufacture of the product involved in the accident.

A manufacturer can address manufacturing defects through quality control processes, and problems would arise only from what is hopefully a small number of products manufactured with issues arising from the manufacturing process. The larger concern for manufacturers stem from design defect cases and failure to warn cases. These defects would affect an entire line of products.

Courts have recognized various tests to show the existence of design defects under state law. Examples include:

- The consumer expectation test: A test based on what an ordinary consumer would expect from a product, typically used where the potential for injury is clear to consumers from the nature of the product.
- The risk-utility balancing test: A test in which the plaintiff contends that the risks from a design outweigh the benefits to the consumer or public from a design.

47. Restatement (Second) of Torts § 402A(1) (1965). "More than three quarters of American jurisdictions incorporate all or part of this section in their own distinct brand of strict liability."
1 FRUMER, *supra* note 4, § 2.04.

48. Restatement (Second) of Torts, *supra* note 47, § 402A(1)(b).

- The prudent manufacturer test: A test asking whether a reasonably prudent manufacturer or seller, aware of the product's dangerous condition, would not have put the product on the market if it had been aware of the product's condition.
- A combination test: A test that shifts the burden of proof to the manufacturer to show a lack of defect in certain situations.
- The ultimate issue approach: A doctrine by which the jury has the discretion to determine whether a design is defective.⁴⁹

Frequently, a plaintiff asserting a design defect will use expert testimony to explain why the defendant's design is defective. Also, using expert testimony, a plaintiff will often attempt to prove that an alternative design could have prevented the accident.

In contrast to a design defect claim, a plaintiff can proceed under a failure to warn theory of strict liability. Failure to warn liability is based on an alleged defect arising because the defendant failed to provide adequate warnings or instructions about the robot or AI system. The plaintiff would need to prove that the warnings did not adequately reduce risks associated with the product or that the instructions were inadequate to inform the user about proper use of the product.

B. Negligence Claims

A plaintiff often pleads a negligence claim in addition to strict liability. Like strict liability, a plaintiff may point to negligence in the design of the product, in the manufacture of a product, or in the defendant's failure to provide adequate warnings or instructions. Unlike strict liability, a plaintiff must show culpability on the part of the defendant, which makes it a harder claim to prove at trial.

Again, defendants are most likely concerned with design negligence or negligent failure to warn. The following are the essential elements of a negligence claim:

- The defendant owed a duty of care to provide a reasonably safe product in terms of design or to warn of dangerous defects—meeting a standard of conduct to protect others against unreasonable risk,
- The defendant breached its duty of care by failing to conform to the standard of conduct required, and
- The defendant's conduct proximately caused
- Damages to the plaintiff.⁵⁰

49. Robert Shields & Carolyn Callaway, *Strict Liability in Tort*, in 1 PRODUCTS LIABILITY PRACTICE GUIDE § 6.04 (John Vargo ed., 2018) [hereinafter, "VARGO"].

50. See Robert Shields & Carolyn Callaway, *Negligence*, in *id.* § 6.03[1].

C. Breach of Warranty Claims

Breach of warranty is another possible claim that a plaintiff can assert to seek damages for alleged defects in a product. Warranties are affirmations or promises concerning a product or its performance, features, or characteristics, such as those concerning the safety of a product. A plaintiff may have a warranty claim if the product does not perform as promised, does not have the promised features or characteristics, or is defective in some other way. As with strict liability and negligence, warranty claim may arise from design defects, manufacturing defects, or failures to warn. Warranty claims turn on whether or not the product adheres to the promises made. The fault of the seller is not relevant to the claim.

Depending on state law, defendants have a number of defenses, including privity, a requirement that the plaintiff provide the seller notice of the alleged breach, and the ability for a seller to disclaim warranties.⁵¹ In most U.S. jurisdictions, privity will not preclude purchasers of a product or their family members to sue companies in the chain of distribution under a warranty theory.⁵²

The essential elements of a breach of warranty claim are:

- The defendant made a warranty,
- The product did not comply with the warranty at the time of the sale,
- The plaintiff's injury was proximately caused by the defective nature of the product, and
- As a result, the plaintiff suffered damages.⁵³

Warranties may be either express or implied. An express warranty is a promise or affirmation based on something the seller actually stated in writing or orally. Express warranties may appear in a sales contract, warranty program documentation, advertisements, or sales collateral. By contrast, the law will sometimes recognize implied warranties regarding the sale of products, which arise by operation of law rather than express statements of the seller.

In the context of the sale of goods, a plaintiff may allege one of two kinds of implied warranties. First, a plaintiff may contend that an implied warranty of merchantability applies to the product in question. An implied warranty of merchantability requires the seller to make sure the product is fit for the ordinary purposes of such product. For instance, the purchaser of a hammer would expect that its head would remain affixed to the handle during its useful lifespan. If the head flies off the first time it is used after purchase, the purchaser could contend that it is not fit for hammering nails, its ordinary purpose. Of the two kinds of implied warranties, this is the more common type asserted in product cases.

51. See Robert Shields & Carolyn Callaway, *Breach of Warranty*, in *id.* § 6.02[4].

52. *Id.*

53. *Id.* § 6.02[1].

Second, a plaintiff could assert that a warranty of fitness for a particular purpose applied to the product in question. Such a warranty arises when the seller knows the particular purpose for which the consumer will use the product, and the buyer is relying on the skill and judgment of the seller to select and furnish suitable products. For example, if a consumer walks into a vehicle dealership and tells the sales representative that he or she wants a pickup truck that is capable of towing a trailer through winding mountainous dirt roads, and the representative recommends a particular truck to satisfy the consumer's needs, the law will recognize that the dealer has warranted that the truck can, in fact, tow trailers in such a setting.

D. Claims under Consumer Protection Laws

As an alternative to strict liability, negligence, or warranty, a plaintiff may seek relief for alleged product defects under various consumer protection laws. While some states do not permit claims under consumer protection laws for bodily injury,⁵⁴ consumer protection laws provide an avenue to seek relief for alleged economic or financial losses. A plaintiff could, for instance, assert that he or she overpaid for a product because in its defective state, the product was not worth what the plaintiff paid for it. Examples of consumer protection laws that a plaintiff could assert include the various state laws that prohibit unfair and deceptive trade practices. For instance, California law includes an Unfair Competition Law (UCL)⁵⁵ prohibiting unfair and deceptive trade practice, a False Advertising Law (FAL)⁵⁶ prohibiting untrue or misleading advertising practices, and the California Consumer Legal Remedies Act (CLRA),⁵⁷ which prohibits various specific types of unfair practices, such as misrepresenting the characteristics or qualities of a product.

A plaintiff seeking remedies under consumer protection statutes would typically need to plead and prove the following essential elements:

- A violation of the statute occurred
- That causes
- Injury to a consumer.

We already have a real-life example of a case involving automation technology—*Sheikh v. Tesla, Inc.*⁵⁸ As of the date of this chapter, *Sheikh* is pending in the U.S. District Court for the Northern District of California. The suit followed a May 2016 accident in which a Canton, Ohio, Tesla driver named Joshua Brown died in a crash in his Tesla, which was under control of the driver assistance system, when

54. JOHN ALEE, et al., *PRODUCT LIABILITY* § 18.02 (2014) [hereinafter, "Alee"].

55. Cal. Bus. & Prof. Code § 17200 et seq.

56. *Id.* § 17500 et seq.

57. Cal. Civ. Code § 1750 et seq.

58. No. 5:17-cv-02193 BLF (N.D. Cal. Complaint filed Apr. 19, 2017).

the Tesla failed to brake for a truck turning left in front of his car. Brown also failed to intervene to stop the car.

The plaintiff purchasers contend that Tesla oversold the capabilities of the Tesla enhanced Autopilot advanced driver assistance system. They contend Tesla's advertising implied that Tesla's system was fully autonomous, as shown in a sales video,⁵⁹ when in fact it is not. According to the purchasers who later filed suit against Tesla, they paid \$5,000 for the enhanced Autopilot system of the kind depicted in the video, but Tesla failed to deliver the Autopilot features in the timeframe Tesla promised. Moreover, they say the Autopilot system that these owners received was unusable and dangerous.⁶⁰

The *Sheikh* plaintiffs allege violations of the California UCL, FAL, and CLRA, as well as fraud by concealment. They also allege similar violations of consumer protection laws under Colorado, Florida, and New Jersey laws. The plaintiffs' theory of recovery is that they paid many thousands of dollars for a product they didn't receive; they say they didn't receive the benefit of their bargain.⁶¹ To bolster their claims, the plaintiffs quote a sentence that allegedly appeared on the Tesla website saying, "All Tesla vehicles produced in our factory, including Model 3, have the hardware needed for full self-driving capability at a safety level substantially greater than that of a human driver . . ."⁶² In addition, they quote the language appearing in the video image: consumers "first would be invited to see a video lasting over two minutes, in which the initial frames shouted: 'THE PERSON IN THE DRIVER'S SEAT IS ONLY THERE FOR LEGAL REASONS. HE IS NOT DOING ANYTHING. THE CAR IS DRIVING ITSELF'"⁶³

Tesla's statements imply that Autopilot is a fully automated system, while in fact they only constitute a driver assistance system, requiring the driver to monitor all driving tasks at all times.⁶⁴ It is possible to draw a line directly from Tesla's statements to the factual allegations of the suit and the causes of action alleged. The plaintiffs used Tesla's own words against it to show that the company allegedly oversold the capabilities of the Autopilot system. As a result, plaintiffs are seeking, among other

59. Tesla, *Tesla Autopilot 2.0*, YouTube (Oct. 20, 2016), <https://www.youtube.com/watch?v=C3DbrYx-SN4> (reposted by Daniel as *Tesla Autopilot 2.0 – Level 5 Autonomy. Full Self-Driving Hardware*).

60. Second Am. Class Action Compl. ¶ 1, at 1, *Sheikh v. Tesla, Inc.*, No. 5:17-cv-02193 BLF (N.D. Cal. Complaint filed Apr. 19, 2017).

61. *Id.* ¶ 4, at 2.

62. *Id.* ¶ 38, at 10-11.

63. *Id.* ¶ 38, at 11.

64. The National Transportation Safety Board called Tesla's Autopilot system an SAE International "Level 2 automated vehicle system." NTSB, *Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston May 7, 2016*, Florida § 1.3.3, at 9 (Sept. 12, 2017). SAE Level 2 means that the human driver is responsible for continuously monitoring the driving environment and is prepared to take over driving at any time.

things, restitution of amounts paid to Tesla for the enhanced Autopilot system and disgorgement of profits to be ascertained at trial.⁶⁵

E. Types of Defects at Issue in Robotics Litigation

Many kinds of defects that could result in product liability litigation. We do not yet have extensive experience with mass-marketed robots and AI systems to establish clear trends. Nonetheless, robots and hardware are likely to suffer from the same kinds of problems seen with conventional equipment. Examples include:

- Mechanical or physical defects or issues, such as metal fatigue, parts separation, and mechanical failures.
- Defects in electrical components or systems other than sensors or control systems for autonomous control, such as the use of wrong kind of components, problems in the performance of the components, or the lack of durability of the components.
- Software or firmware code defects relating to systems other than sensors or control systems for autonomous operation, including information security vulnerabilities.

With these conventional types of defects or issues, a manufacturer's liability would depend on the facts and circumstances that we would see with conventional products.

Robots and hardware making use of artificial intelligence, however, may have defects that would not affect conventional equipment or devices. Again, these defects may be mechanical, electronic, or relating to software or firmware code. They may, for instance, include:

- Mechanical or physical defects in the control systems for autonomous operation or the sensors⁶⁶ used by the autonomous systems.
- Defects in electrical components for sensors or control systems for autonomous operation.
- Defects in the code used for sensors or control systems used for autonomous operation.

65. For a more thorough discussion of this case and claims against automation companies based on unfair and deceptive trade practices, see Stephen S. Wu, *Unfair and Deceptive Trade Practice Claims Against Manufacturers of Automated Vehicles*, THE SCI TECH LAWYER, Summer 2018, at 5.

66. In *Wickersham v. Ford Motor Co.*, No. 9:13-cv-1192-DCN; No. 9:14-cv-0459-DCN, 2016 U.S. Dist. LEXIS 131072 (D.S.C. Sept. 26, 2016), the plaintiff alleged that the plaintiff's decedent's Ford Escape had a defective crash sensor which improperly deployed the car's airbag, causing the decedent's injuries and eventually his suicide. Similar types of claims may arise from defects in robot or autonomous vehicle sensors.

F. Defenses in Product Liability Cases

While plaintiffs may be able to assert a number of claims against defendants for alleged defects in robots or AI products, defendants may have a number of available defenses. Some of the defenses relate to the plaintiff's own conduct. A plaintiff may have caused or contributed in some way to an accident. A plaintiff's failure to exercise reasonable care may constitute contributory or comparative negligence. Also, a plaintiff's misuse or modification of the product may be a cause of an accident. Further, a defendant may be able to allege that the plaintiff assumed the risk of an open, obvious hazard. Finally, a plaintiff's failure in some way to mitigate the alleged damages may reduce any award against the defendant.

Unlike with conventional products, robots and AI systems may operate autonomously. One issue that may arise with any attempt to use the plaintiff's own conduct as a defense is whether the plaintiff was in control of the product at the time of the accident. If the product was acting in autonomous mode, there may have been no way for the plaintiff to have contributed to the accident. For instance, with regard to autonomous vehicles, if the vehicle was driving itself and in control of all driving tasks at the time of the accident, then a plaintiff's carelessness at that moment could not have contributed to the accident.

Nonetheless, events preceding the accident could bear on the plaintiff's conduct. For example, in our autonomous vehicle accident, if the plaintiff modified the vehicle to drive faster than the speed limit and safety limits set by the manufacturer, the plaintiff's conduct could have contributed to the accident even though the vehicle was in autonomous mode right before the accident occurred. We can foresee that some people will modify their robots or try to exceed their limits in some way. Likewise, pedestrian and bystander accident victims might contribute to accidents by abusing robots for fun, for example by darting in front of them to see how they try to avoid a collision.

Given the fast-moving nature of artificial intelligence technology, manufacturers may also want to assert a "state of the art" defense to product liability cases involving robots and AI system. Defendant manufacturers could claim that at the time the product in question left their hands, a safer design was not technologically feasible. Some states recognize a state of the art defense, while others do not.⁶⁷

A defendant may also assert the economic loss defense. This defense precludes tort claims for product liability where the alleged damages are financial and not for bodily injury or damage to property other than damage to the product itself. Another available defense in some cases is preemption. Under this defense, federal law under the U.S. Constitution's Supremacy Clause⁶⁸ preempts inconsistent state law. A manufacturer may be able to shift responsibility to a "sophisticated user" or a "sophisticated" or "learned intermediary." For instance, a pharmaceutical company may have an obligation to provide product warnings about a prescription drug to

67. 1 FRUMER, *supra* note 4, § 8.04.

68. U.S. CONST. art. VI, cl. 2.

a doctor that could prescribe it, rather than a patient. Some products may be appropriate for use by engineers, technicians, and other sophisticated users. In that case, a manufacturer's duty to provide warnings and instructions would only require the level of information needed for the class of likely users and purchasers, rather than the greater level of warnings or instructions needed if a member of the general public were to start using it.⁶⁹ Certain industrial, transportation, and specialty robots could fall within this category of product. Finally, a manufacturer may be able to assert a "government contractor" defense. If the manufacturer made a product in accordance with government specifications, in certain cases, it may have a defense to an action based on issues inherent with the government specifications, rather than any fault of the manufacturer.

V. MANAGING THE RISK OF ROBOT PRODUCT LIABILITY

Given the large human and financial consequences of defective products, manufacturers seek to manage the risk of product liability litigation and costly recalls. What can a robot or AI system manufacturer do to reduce the likelihood of company-ending product liability litigation? Most importantly, if manufacturers can proactively prevent defects and resulting accidents from occurring in the first place, they can prevent the need to defend product liability claims. Planning for improved safety can enable manufacturers to make safer products that are less likely to cause accidents and trigger product suits.

Of course, accidents may occur anyway and with any widely-deployed robot or AI system, a manufacturer can foresee that accidents are inevitable. Nonetheless, a proactive approach to risk management would permit a manufacturer to put itself in the best position possible to prevail in product liability cases based on the inevitable accidents. A proactive approach to design safety means that the manufacturer takes the steps today to implement a commitment to safety, which will minimize its risk from future suits. As mentioned in the discussion above, juror anger fuels outsized verdicts. If a proactive manufacturer takes the concrete and effective steps to implement a commitment to safety, it will be able to tell a future jury why its products were safe and how it truly cared about safety. Such actions will place the manufacturer in the best possible light when, despite all these safety measures, an accident does occur.

Making the commitment to safety upfront is crucial. As one commentator stated, "The most effective way for [counsel for] a corporate defendant to reduce anger toward his or her client is to show all the ways that the client went *beyond what was required*

69. One reported decision focuses on this defense in the context of a robot, *Taylor v. Intuitive Surgical, Inc.*, 187 Wash. 2d 743, 389 P.3d 517 (2017). The court held that a manufacturer's duty to warn regarding a surgical robot purchased by a hospital and operated by a surgeon requires that the manufacturer warn the purchaser hospital in addition to the operating surgeon. See *id.* at 753-56, 389 P.3d at 522-24.

by the law or industry practice.”⁷⁰ Going beyond minimum standards is important because, first, juries may look at minimum standards skeptically, thinking that the industry set the bar too low. Moreover, juries expect that manufacturers know more about their product than any ordinary “reasonable person,” which is the standard for judging a defendant in a negligence action. Juries expect more from manufacturers. “A successful defense can also be supported by walking jurors through the relevant manufacturing or decision-making process, showing all of the testing, checking, and follow-up actions that were included. Jurors who have no familiarity with complex business processes are often impressed with all of the thought that went into the process and all of the precautions that were taken.”⁷¹ The most important thing to a jury is that the manufacturer tried hard to do the right thing.⁷² Accordingly, a manufacturer that goes above and beyond minimum industry standards is in the best position to minimize the likelihood of juror anger and minimize possible product liability risk.

Any proactive approach to product safety should begin with a thorough risk analysis. A risk analysis would look at the types of problems that could arise with a product, how likely these problems could occur, and the likely frequency and impact of these problems. After completing this analysis, a manufacturer can analyze its robot or AI product design in light of the risks. It can change design and engineering practices to address potential issues and prioritize risk mitigation measures based on what it sees as the most significant risks. In implementing this risk management process, a manufacturer may obtain guidance from a number of standards relevant to robots and AI systems. Examples include:

- ISO 31000 “Risk management – Principles and guidelines” (regarding the risk management process).
- Software development guidelines from the Motor Industry Software Reliability Association.
- IEC 61508 Functional safety of electrical/electronic/programmable electronic safety-related systems (safety standard for electronic systems and software).
- ISO 26262 family of “Functional Safety” standards implementing IEC 61508 for the functional safety of electronic systems and software for autos.

Adherence to international standards may not insulate a manufacturer from liability, whether in front of a jury or as a matter of law. Nonetheless, following international standards increases the credibility of a manufacturer’s risk management program. Also, following standards helps a manufacturer create a framework of controls for its risk management process. Therefore, organizing a risk management program based on the methods specified in international standards provides an important basis for defending later product liability litigation.

70. Minick & Kagehiro, *supra* note 16, at 2.

71. *Id.*

72. *See id.*

In addition to adhering to international standards, insurance will play an important role in managing robot and AI product liability risk. Insurance functions to shift product liability risk to insurance carriers. In exchange for paying a premium, a manufacturer's insurance carriers will defend and indemnify manufacturers for losses and pay for settlements or judgments to resolve third party claims. The insurance industry is in the early stages of understanding robot and AI risk and creating coverage that effectively manages risk.⁷³ As businesses and consumers deploy robots and AI systems more broadly, insurers will create insurance programs for third party accident and liability risks. Some of those risks may include privacy and security breaches. One barrier to effective insurance programs is the lack of loss experience data to assist in the underwriting process. To start writing policies for given robots or AI systems, however, insurance carriers are likely to look at analogous conventional products.⁷⁴ In the short run, manufacturers may need to tailor-make insurance coverage with bespoke policies that fit their risk profiles. Over time, carriers will enter the market and create standard policies, reducing premium costs over the longer run.

Beyond the most immediate internal safe design steps and insurance programs, manufacturers of a given type of robot or AI system may be able to act jointly to mitigate risk to the entire industry sector (subject to possible antitrust issues involving joint action). For instance, they may work on safety and information security standards to promote safe practices within the industry sector. Trade groups and purchasing consortia can help manufacturers promote the safety among component manufacturers. Finally, an industry sector may want to create and maintain information sharing groups to develop and promote safety practices among industry participants.

During the design process, effective records and information management (RIM) will help a manufacturer document and evidence its commitment to safety. Documents generated contemporaneously with the design process can memorialize a manufacturer's safety program and the steps it takes to fulfill its commitment to safety. In any product liability suit, a witness could certainly testify about the manufacturer's safety program. Nonetheless, without corroborating contemporaneously recorded documentation, there is a risk that the jury would find any such testimony to be self-serving and thus disbelieve it. In this vein, wholesale destruction of all design documents of a certain age may be as bad as retaining too many documents. Archiving the right documents in preparation of future litigation will help the business defend itself in the future. Effective RIM may win cases, while poor RIM may lose cases.

Finally, some pre-litigation strategies may further reduce product liability risks. For example, manufacturers can work with jury consultants to advise the manufacturer in the defense of a product liability case. They can focus on ways the manufacturer

73. See generally Lloyd's Paper, *supra* note 2.

74. Cf. David Beyer et al., *Risk Product Liability Trends, Triggers, and Insurance in Commercial Aerial Robots* 20 (Apr. 5, 2014) (describing the development of insurance coverage for drones), available at http://robots.law.miami.edu/2014/wp-content/uploads/2013/06/Beyer-Dulo-Townsley-and-Wu_Unmanned-Systems-Liability-and-Insurance-Trends_WE-ROBOT-2014-Conference.pdf.

can place its safety program in the best light to avoid impressions that would anger a jury. Moreover, a manufacturer may want to create a network of defense experts familiar with their robotics or AI technologies. These experts can help educate jurors about various engineering, information technology, and safety considerations. Further, attorneys representing AI and robotics manufacturers may work within existing bar groups or form new ones to share specialized knowledge, sample briefs, case developments, and other information helpful to the defense of product liability cases.

VI. CONCLUSIONS

Product liability lawsuits following accidents create a potentially existential risk to robot and artificial intelligence system manufacturers. Manufacturers that ignore risks to customer or the public may face angry juries that could award huge jury verdicts to plaintiffs. We can see from the benchmarks set by the Toyota sudden acceleration and General Motors ignition switch events that product liability and related expenses can cost manufacturers billions of dollars. Defects may arise from mechanical, electrical, or software issues, permitting plaintiffs to assert various types of claims against manufacturers, although manufacturers may have different kinds of defenses to these claims.

Manufacturers can take steps to manage product liability risk and proactively prepare today, during the design of robots and AI systems, to prevail in future litigation for accidents that have yet to occur. Most importantly, manufacturers that take a proactive approach to analyze risk, adhere to industry standards, and document an effective commitment to product safety will place themselves in the best position possible to defend future product liability litigation. They can further manage their risk through robust insurance coverage, industry collaboration on safe practices, and effective records and information policies to document their safety programs. In short, product liability is a serious risk to robot and AI system manufacturers, but the proactive approach to risk management can maximize safety and minimize liability risk over time.

AI, Automation, and the Future of Transportation

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

AI, Automation, and the Future of Transportation

January 9, 2020

9:15 am – 10:30 am

National Institute on Artificial Intelligence and Robotics

Santa Clara University Law School

Speakers

- Nandi Chhabra
General Counsel, Peloton Technology
- Dr. Selina Pan
Research Scientist, Toyota Research Institute
- Phillip Zackler
Legal Director, Toyota Research Institute

2020 – Where we are supposed to be

- Landed on Pluto
- Vacationing on the moon
- Robots have replaced laundromats
- Food consumption replaced by nanobots
- Conscious Computer with superhuman intelligence

2020 and Beyond

- Micro-mobility
- Car-sharing
- Ride Hailing and Ride Sharing
- Connected Vehicles
- MaaS Industry
- Last mile robotic systems
- EV / AV Development
- Vertical Take-off and Landing (VTOL)

...and much more

Agenda

Topic	Speaker
AV Technology Components	Dr. Selina Pan Research Scientist Toyota Research Institute
From Research to Deployment	Phillip Zackler Legal Director Toyota Research Institute
Truck Platooning	Nandi Chhabra General Counsel Peloton Technology

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



AMERICAN **BAR** ASSOCIATION

Science & Technology
Law Section



TOYOTA
RESEARCH INSTITUTE

Transportation Panel: Technology Components

Dr. Selina Pan, Toyota Research Institute

January 9, 2020

National Institute on Artificial Intelligence and Robotics
Santa Clara University Law School

Agenda

- Overview
- Object Perception
- SLAM
- Motion Planning and Control

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Overview

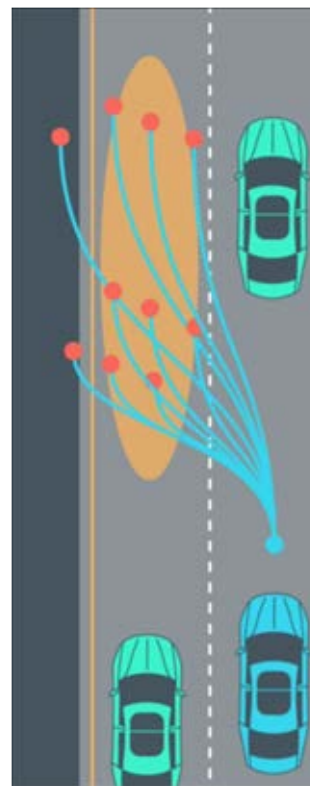
Hello!

- Research Scientist in Planning and Control at Toyota Research Institute
- Previously: Research Scientist in Automated Driving at Ford, 2016-2017
- Postdoc in mechanical engineering at Stanford University, 2014-2016
 - Path planning and control, vehicle dynamics, ethics, human-vehicle interaction
- PhD in mechanical engineering at University of California, Berkeley, 2008-2014
 - engine control, nonlinear control, adaptive control



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Images [Discover Central Massachusetts] [Or Rivlin]

Overview

Analogy for the talk: you need to get from the back of a crowded room to the front.

How do you go about doing this?

Take in the scene with your eyes and detect what all the other entities are: people, things, dynamic, static.

Understand where you are relative to some internal map: is the front of the room to your left, right, front, or back? What is your orientation?

Plan your path: Synthesize the prior information and work out a trajectory. Move body according to trajectory.

Overview

Analogy for the talk: you need to get from the back of a crowded room to the front.

How do you go about doing this?

Take in the scene with your eyes and detect what all the other entities are: people, things, dynamic, static. **[OBJECT PERCEPTION]**

Understand where you are relative to some internal map: is the front of the room to your left, right, front, or back? What is your orientation? **[SLAM]**

Plan your path: Synthesize the prior information and work out a trajectory. Move body according to trajectory. **[MOTION PLANNING AND CONTROL]**

Overview

A self-driving car needs to drive from Point A to Point B.

How does it go about doing this?

Across the industry, the following technologies are generally used:

Object Perception: It uses sensors to perceive all objects surrounding and algorithmic reasoning to identify what the objects are.

SLAM: It uses GPS, maps, and algorithmic reasoning to localize itself.

Motion Planning and Control: It uses all the aforementioned information to plan a trajectory forward, and commands the vehicle actuators to execute that trajectory.

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



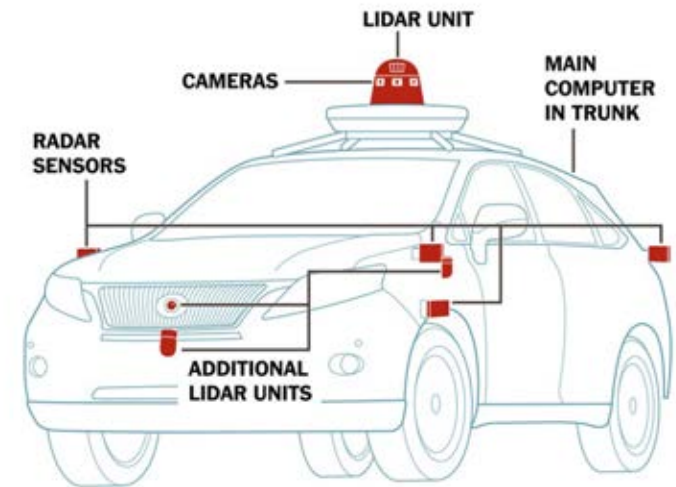
Object Perception

Object Perception

When your eyes look at what's surrounding you and identifies each item

“Eyes” of a self-driving car

- Sensors such as radar, lidar, came
- “Brain” that works on identification
- Algorithms run in an on-board co
 - Object classification, machine learning



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Sample view of self-driving car system with detected objects surrounding

Image [Toyota Research Institute]

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Sample view of self-driving car system with detected objects surrounding

Image [New York Times]

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



TOYOTA
RESEARCH INSTITUTE

SLAM

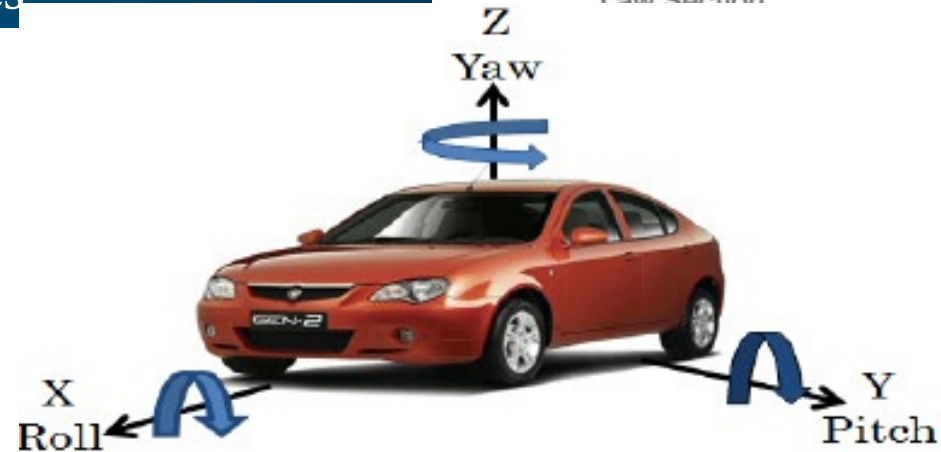
NATIONAL INSTITUTES Artificial Intelligence and Robotics

SLAM

When your brain figures out where you are relative to your environment

“Brain” uses tools to determine this

- GPS, base station, maps
 - These help the car determine not only where in the world it is, but also how it is oriented
 - “Simultaneous Localization And Mapping”



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



TOYOTA
RESEARCH INSTITUTE

Motion Planning and Control

Motion Planning and Control

When your brain decides how you want to move and then makes your body move

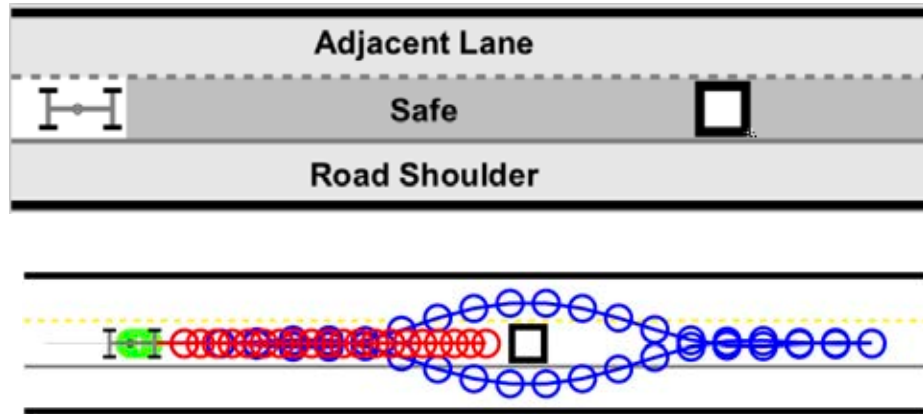
“Brain” of the car

- Algorithms running in the on-board computer
 - Prediction, decision-making, problem generation, trajectory selection, trajectory planning, control

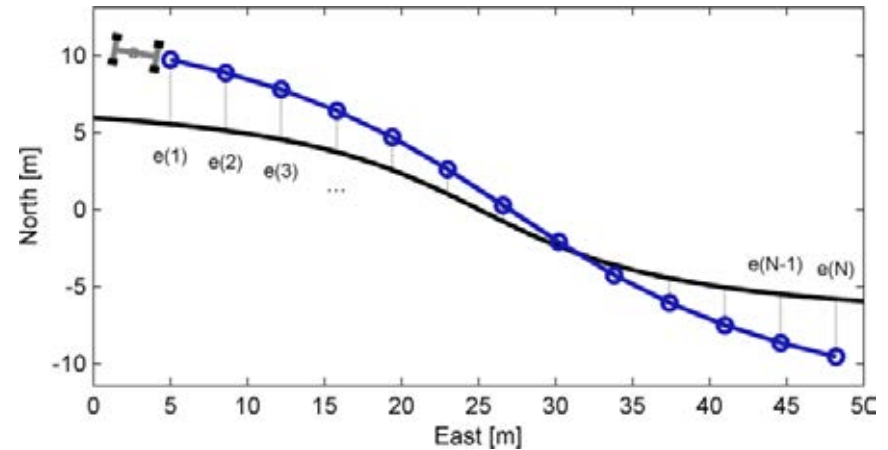
“Body” of the car

- Control: commands the actuators to execute the brain’s plans
 - A variety of control algorithms from traditional to optimization
 - Steering wheel and brake/throttle

NATIONAL INSTITUTES Artificial Intelligence and Robotics



**Decision-Making
Trajectory Selection**



**Trajectory Planning
Control**

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Thank You

AI Systems: Issues from Research to Deployment

Phillip Zackler
Director, Legal & Compliance
Toyota Research Institute

January 9, 2020

Legal Coverage

Corporate	Privacy
Commercial	Real Estate
Dispute Resolution	Regulatory
Human Resources	Safety
Intellectual Property	Standards
Litigation	Tax

Select Topics

1. Confidentiality
2. Software and Data Licensing
3. Export Control
4. Regulatory
5. Standards and Safety
6. Product Liability

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



AMERICAN BAR ASSOCIATION

Science & Technology
Law Section

Confidentiality

Confidentiality

“vs”

Academic Publications
Open Source

Software and Data Licensing

Licensing

- Commercial vs. Non-Commercial
- Open-Source
- Affiliate Companies
- Component integration into software stack
- University Partnerships
- “Derivative” Works

Export Control

Export Control

- Department of Commerce
 - Export
 - Deemed Export
 - Export Control Classification Number
- Department of State
 - ITAR
- Department of Treasury
 - OFAC

Export Control – potentials

- Navigation Equipment /GPS
- LIDAR
- Laser Systems
- Sensors
- Encryption
- Software
- Chemicals

Export Control – Activity Examples

- Licensing software overseas
- Presenting research at an international conference
- Hand carrying material on a flight overseas
- Deemed Export = Sharing information with colleagues that are foreign nationals

Export Control

Penalties

- Civil and criminal penalties for the individual and organization
- Substantial fines
- Imprisonment for violators
- Denial of Export privileges

Export Control

FIRRNA

Foreign Investment Risk Review Modernization Act

Export Control

ECRA

Export Control Reform Act

Critical Technologies - potentials

- Biotechnology
- Artificial Intelligence
- Position, Navigation, and Timing (PNT) technology.
- Microprocessor technology
- Advanced computing technology
- Computer analytics

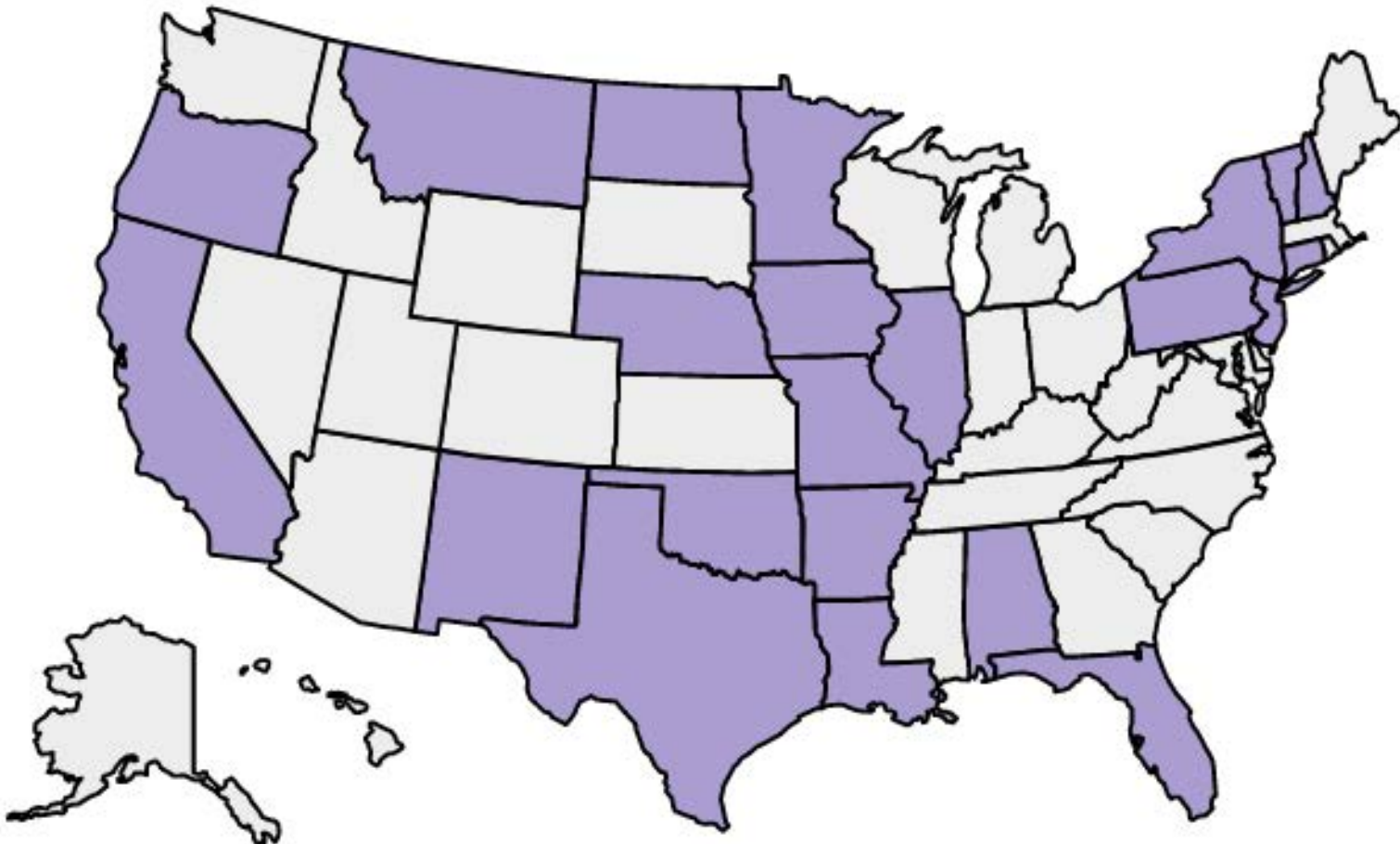
Export Control – Fast moving

Monday January 6, 2020

- Interim Final Rule - New ECCN
 - Software analyzing Geospatial Imagery
- First dual-use export control explicitly for a type of AI software.

Regulatory

State Legal Landscape



U.S. Legislative Action

- 115th Congress
 - House passed SELF DRIVE ACT (Sept 2017)
 - Senate Commerce Comm passed AV START ACT (Oct 2017)
- 116th Congress
 - Drafting in progress

NATIONAL INSTITUTES Artificial Intelligence and Robotics



NHTSA



Automated Vehicles 3.0

PREPARING FOR THE FUTURE OF TRANSPORTATION



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



AMERICAN BAR ASSOCIATION

Science & Technology
Law Section

Standards and Safety

Standards: One Example

SAE Automated Vehicle Safety Consortium	
Participants	Standard Creation
<ul style="list-style-type: none">• Ford• GM• Toyota• Uber ATG	<ul style="list-style-type: none">• Safety principles• Common terminology• Best safety practices

Safety: Design Decisions

- Evaluation of risks/hazards
- Reasonable design criteria
- Test procedures to evaluate performance
- Testing to show adherence to criteria

Safety: Testing and Operations

- Functional Safety
 - Safety Case
- Operations / Processes
 - Technician Training and Guidelines
 - Pre-mission Safety Alignment
 - Incident Response

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



AMERICAN BAR ASSOCIATION

Science & Technology
Law Section

Product Liability

Product Liability - Traditionally

Type of Injury	Serious
Parties	Any party involved in: Design Development Manufacturing Marketing Distribution

Product Liability – Wide Adoption of AVs

Type of Injury	Any
Parties	Traditional Parties Data Service Provider Cloud Service Provider Information Security Vendor Managed Security Vendor

Product Liability – Traditional principles

- Strict Liability for Design Defect
- Negligence
- Breach of Warranty
- Misrepresentation
 - Unfair / Deceptive Trade Practices
 - False Advertising
- Fraud
- Security or Privacy Breach

Product Liability - Defenses

- No Defect
- Comparative Fault
- Failure to Heed Clear Warnings
- Product Abuse
- Assumption of the Risk

Documentation of Design Decisions

- Evaluation of risks/hazards
- Reasonable design criteria
- Test procedures to evaluate performance
- Testing to show adherence to criteria



Peloton

ABA AI & Robotics National Institute

Nandi Chhabra, General Counsel, Peloton Technology

Trucking Industry in America

TRANSPORTATION INDUSTRY

71% of all cargo is transported by trucks



8%

by air



6%

by pipeline



4%

by rail



2%

by water



TRUCKING INDUSTRY

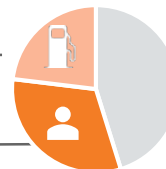
\$796.7B annual trucking industry revenue



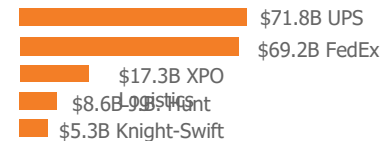
freight revenue is forecast to grow 53.8% by 2030

24% of operating expense spent on fuel

33% of operating expense spent on labor



TOP5 trucking companies by revenues



\$93 Billion spent on commercial vehicle crashes in 2011

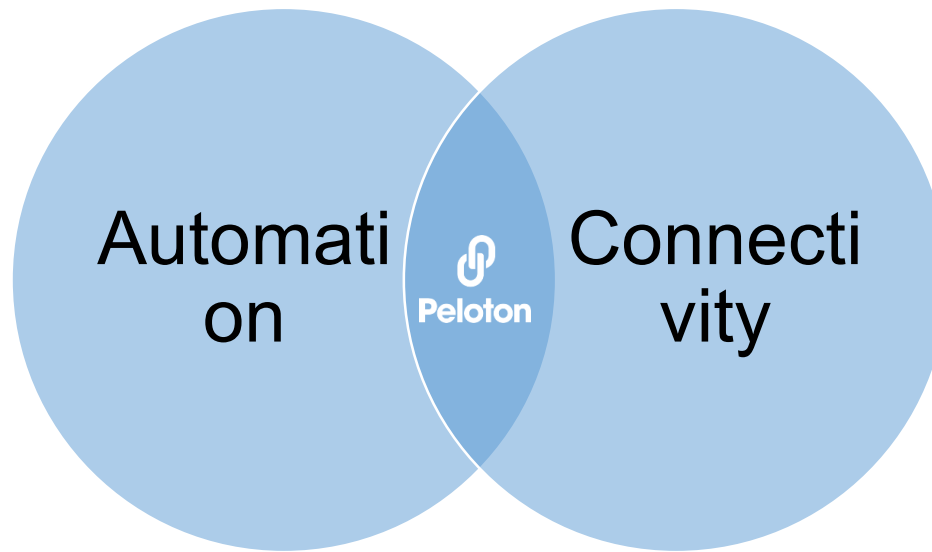
Platooning



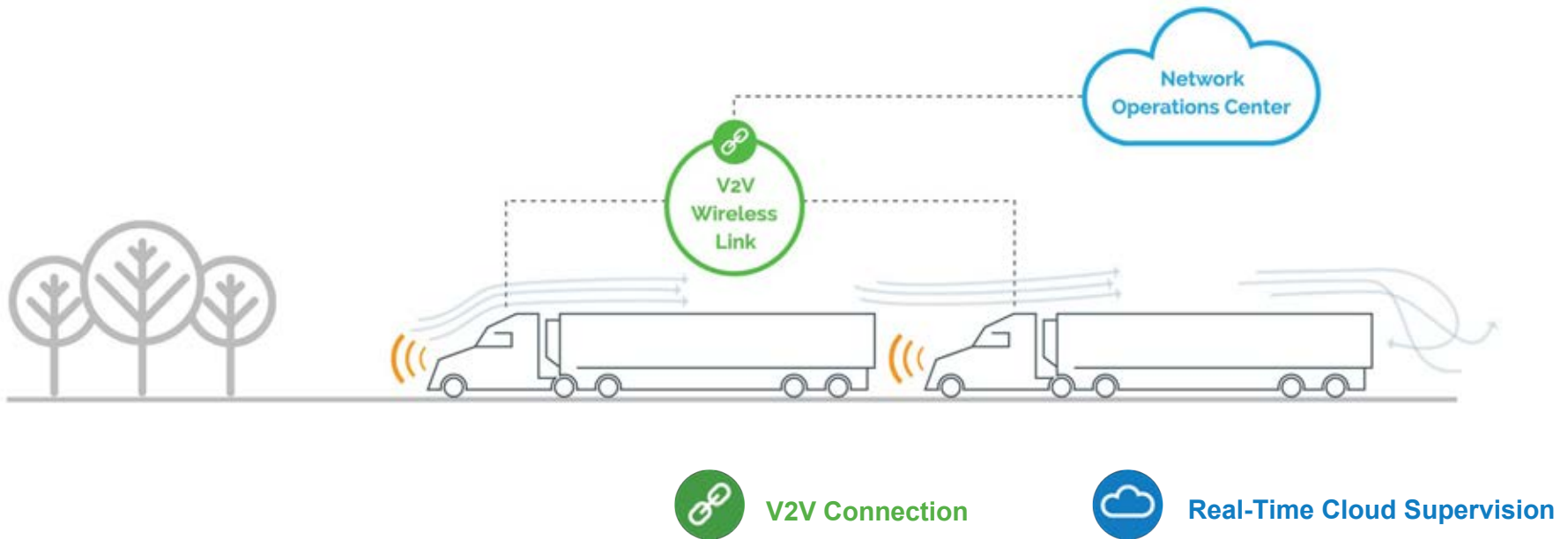
 Peloton
PLATOONPRO

 Peloton
Auto**Follow**

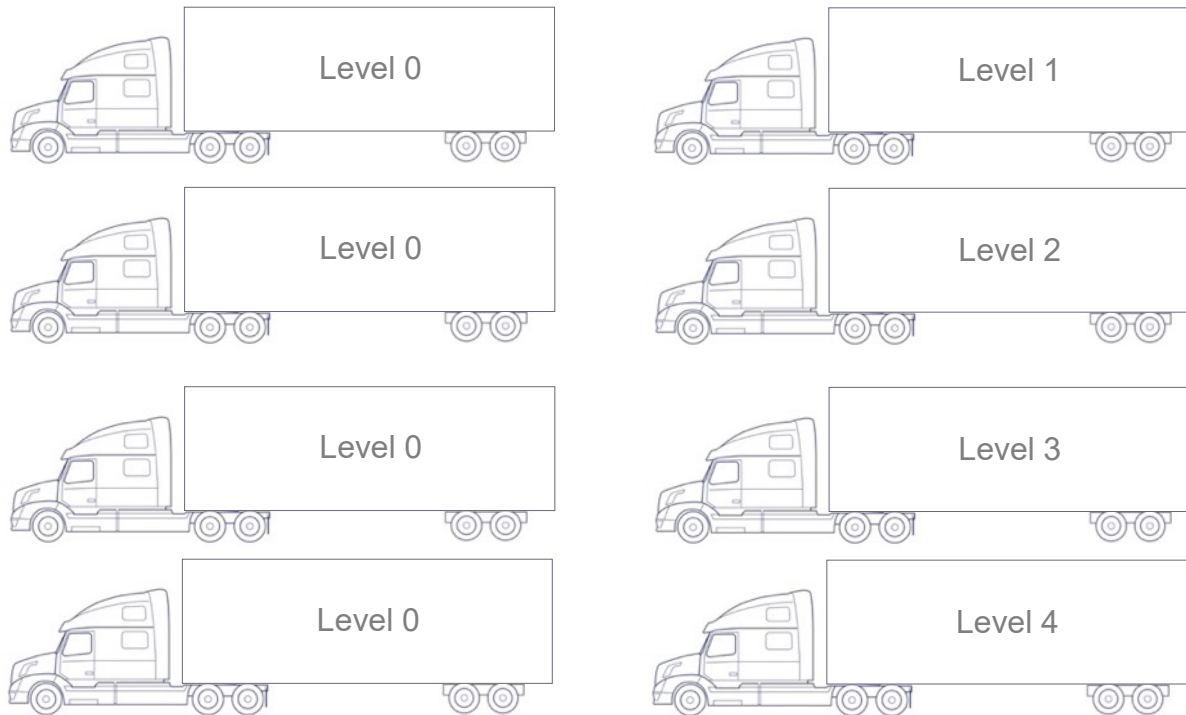
Connectivity and Automation



Connectivity and Automation



Platooning Automation Scenarios



Regulatory Landscape: Federal

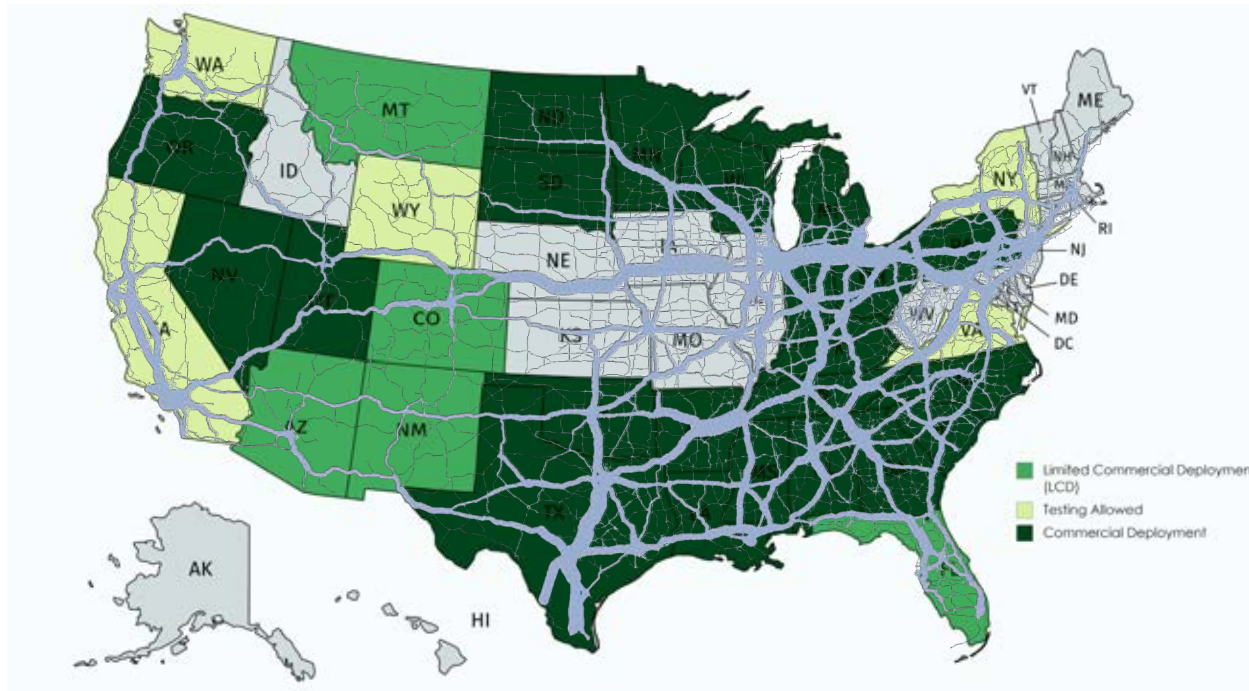


“Best Practices for State Legislatures

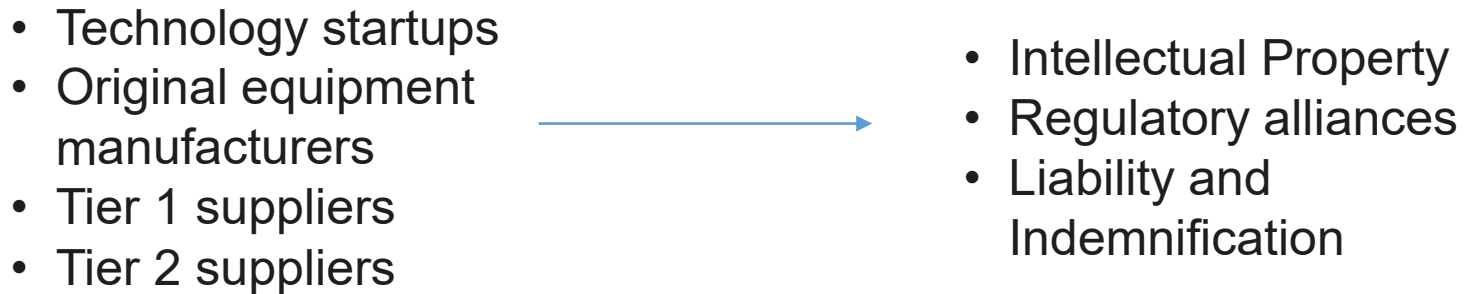
...

A Vision for Safety 2.0 recommended that State legislators follow best practices, such as providing a technology-neutral environment, licensing and registration procedures, and reporting and communications methods for public safety officials. States should consider reviewing and potentially modifying traffic laws and regulations that may be barriers to automated vehicles. For example, several States have following distance laws that prohibit trucks from following too closely to each other, effectively prohibiting automated truck platooning applications.”

Regulatory Landscape: States



Technical Partnerships




Product Liability





IN THIS ISSUE AUTONOMOUS VEHICLES • AV CYBERSECURITY • ROAD SAFETY

THE **SciTech** LAWYER

VOLUME 14 ISSUE 4 | SUMMER 2018 | SECTION OF SCIENCE & TECHNOLOGY LAW | AMERICAN BAR ASSOCIATION 

SELF-DRIVING CARS

MATTHEW HENSHON AND SARAH MCMILLAN, ISSUE EDITORS

Published in The SciTech Lawyer, Volume 14, Number 4, Summer 2018. © 2018 American Bar Association. Reproduced with permission. All rights reserved. This information or any portion thereof may not be copied or disseminated in any form or by any means or stored in an electronic database or retrieval system without the express written consent of the American Bar Association.

UNFAIR AND DECEPTIVE TRADE PRACTICE CLAIMS AGAINST MANUFACTURERS OF AUTOMATED VEHICLES

BY STEPHEN S. WU

On March 18, 2018, an Uber test car driving in autonomous mode, with a safety driver in the vehicle, struck and killed a woman walking a bicycle across a street in Tempe, Arizona. Volvo and Uber worked together on the autonomous driving system controlling the car. After an accident like that, we can reasonably anticipate that the estate of the pedestrian could bring a product suit against Uber, Volvo, or both.

Perhaps the most famous automobile product liability case in history was the Ford Pinto exploding gas tank case titled *Grimshaw v. Ford Motor Co.*,¹ tried in Orange County, California's Superior Court. The appellate decision in the case reports that the plaintiff tried the case to the jury based on theories of strict liability and negligence.² It was common in the 1970s and 1980s to see product liability cases based on accidents that alleged strict liability, negligence, breach of warranty, and common law claims such as fraud.

These claims still appear in accident cases. A recent example is the class action for bodily injury alleging that Toyota cars suddenly accelerated without warning and the drivers could not stop them.³ One of the accident actions, *Spisto v. Toyota Motor North America, Inc.*,⁴ alleged negligence, strict liability based on design defect, a failure to warn strict liability claim, breach of the

implied warranty of merchantability, and fraudulent concealment. To avoid these kinds of suits, manufacturers have focused on safe designs, quality control, and sufficient instructions and warnings provided to consumers. Manufacturers of automated vehicles (AVs) will likely focus on the same kinds of controls in order to avoid these kinds of claims.

More recently, however, plaintiffs are filing complaints on behalf of consumers that never had an accident. How is that possible? At first blush, it might appear that owners that remained safe from accidents have no case against the manufacturer. After all, they experienced no accident, sustained no bodily injury, and were not affected by any damage to their property.

The plaintiffs filing these actions seek the recovery of economic losses only. They contend that they sustained losses because the cars they bought were not worth what they paid for them. The defects and problems they identify, despite not causing an accident, diminish the market value of their cars—what a buyer would be willing to pay for the car. In some cases, they allege they sold their allegedly defective cars to get rid of them at a lower price than if their cars did not have the alleged defects. The economic loss claimed is the alleged defect's diminution in value to the car.

Plaintiffs alleging purely economic loss assert a different set of claims against a manufacturer: claims based on violations of consumer protection laws barring unfair and deceptive

trade practices. Manufacturers of modern cars driven by humans, and in the future manufacturers of AVs, must not only manage the risks of product liability claims, but also unfair and deceptive trade practice claims. In terms of economic effect, each case is different. Nonetheless, manufacturers should be just as worried about unfair and deceptive trade practice claims as they are product liability claims.

Unfair and Deceptive Trade Practice Laws

Manufacturers face liability under both federal and state laws that prohibit unfair and deceptive trade practices. At the federal level, the Federal Trade Commission (FTC) has authority under Section 5 of the Federal Trade Commission Act⁵ to stop unfair and deceptive trade practices. Under Section 5, "Unfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are hereby declared unlawful."⁶ The FTC Act states that the FTC "is hereby empowered and directed to prevent persons . . . from using unfair methods of competition in or affecting commerce and unfair or deceptive acts or practices in or affecting commerce."⁷

With the FTC Act, the risk for a manufacturer is based on a possible governmental investigation or action, rather than a private plaintiff action. Nonetheless, a manufacturer faces legal risk from violating Section 5. Section 5 gives the FTC two kinds of authority: deception

Stephen Wu (ssw@svlg.com) is a shareholder with Silicon Valley Law Group in San Jose, California.

authority and unfairness authority. The authority to strike at deceptive conduct permits the FTC to seek to stop material misstatements about a product or fraudulently concealing information from buyers. In many cases, the FTC can easily prove that a manufacturer of a product promises one thing in public statements but sells products that are inconsistent with those promises.⁸

The FTC's authority to strike at unfair conduct is potentially more difficult to prove. The FTC must prove that an act or practice causes consumer injury that is "(1) substantial, (2) without offsetting benefits, and (3) one that consumers cannot reasonably avoid."⁹ Unfairness claims require the determination of consumer injury based on an investigation and balancing consumer harms against benefits, which is harder than simply finding discrepancies in public statements about a product.¹⁰

In addition, states have enacted counterparts to the FTC Act, sometimes called "Little FTC Acts." One of the most prominent examples is California's Unfair Competition Law (UCL) codified at Business & Professions Code Section 17200 et seq. Under Section 17200, "unfair competition shall mean and include any unlawful, unfair or fraudulent business act or practice and unfair, deceptive, untrue or misleading advertising."¹¹ A plaintiff that has suffered injury and lost money or property as a result of a violation may allege a UCL claim¹² under one or more of the three prongs of the UCL: the "unlawful" prong, the "unfair" prong, or the "fraud" prong.

Under the "unlawful" prong, a plaintiff can point to conduct of the defendant that violates other law. The UCL creates a private right of action for a violation of other law, even if that other law does not have its own private right of action. Similar to the FTC Act, the "unfairness" prong calls for an analysis of harm to consumers balanced against the motives and justification for a business practice. Immoral conduct, unethical conduct, and conduct that violates public policy, or conduct that causes substantial consumer injury are actionable. Finally, to assert a UCL claim under the "fraud" prong, the plaintiff

must allege and prove that the defendant made false statements on which the plaintiff relied.

In California, the False Advertising Law (FAL) supplements the UCL, stating:

It is unlawful for any person . . . with intent . . . to dispose of real or personal property or to perform services . . . to make . . . , including over the Internet, any statement, concerning that real or personal property or those services . . . which is untrue or misleading, and which is known, or which by the exercise of reasonable care should be known, to be untrue or misleading. . . .¹³

In addition to the UCL and FAL, California enacted the California Consumers Legal Remedies Act (CLRA),¹⁴ which identifies certain forms of "unfair methods of competition" and "deceptive acts or practices" in the sale or lease of consumer goods or services.¹⁵ One of the sections of CLRA prohibits "[r]epresenting that goods or services have . . . characteristics . . . , uses, benefits, or quantities that they do not have."¹⁶ Other CLRA sections may apply in a given case.

Unfair and Deceptive Trade Practice Claims in the Automated Vehicle Context

The concern about unfair and deceptive trade practice claims for AV manufacturers is not theoretical. We already have a real-life example of a case of this kind—*Sheikh v. Tesla, Inc.*¹⁷

Sheikh is pending in the U.S. District Court for the Northern District of California. The suit followed a May 2016 accident in which a Canton, Ohio, Tesla driver named Joshua Brown died in a crash in his Tesla, which was under control of the driver assistance system, when the Tesla failed to brake for a truck turning left in front of his car. Brown also failed to intervene to stop the car.

The backstory of the case begins with Tesla's advertising campaign of its "Autopilot" driver assistance system. The name "Autopilot" itself suggests that

the car is driving itself without the need for human monitoring. Moreover, Tesla showed a video on its website of how the Autopilot system works, which is no longer on the Tesla website but has been reposted on YouTube.¹⁸ The video shows a vehicle occupant entering his Tesla car and driving to an office setting. The throbbing beat of the Rolling Stones song "Paint It, Black" plays throughout the video. The camera is behind the driver looking forward towards his hands, which are away from the steering wheel but close by. At the end of the video, the driver leaves the car, and without an occupant, it finds a parking spot and parks itself.

The Tesla video begins with the following introductory text:

THE PERSON IN THE DRIVER'S SEAT IS ONLY THERE FOR LEGAL REASONS.

HE IS NOT DOING ANYTHING.

THE CAR IS DRIVING ITSELF.

According to the purchasers who later filed suit against Tesla, they paid \$5,000 for the enhanced Autopilot system of the kind depicted in the video, but Tesla failed to deliver the Autopilot features in the timeframe Tesla promised. Moreover, they say the Autopilot system that these owners received was unusable and dangerous.¹⁹ As a result of these claims, a number of named plaintiffs filed the *Sheikh v. Tesla, Inc.* suit on behalf of themselves and others similarly situated.

The *Sheikh* plaintiffs allege violations of the California UCL, FAL, and CLRA, as well as fraud by concealment. They also allege similar violations of consumer protection laws under Colorado, Florida, and New Jersey laws. The plaintiffs' theory of recovery is that they paid many thousands of dollars for a product they didn't receive; they say they didn't receive the benefit of their bargain.²⁰ To bolster their claims, the plaintiffs quote a sentence that allegedly appeared on the Tesla website saying, "All Tesla vehicles produced in our factory, including Model 3, have the hardware needed for full self-driving capability at a safety

level substantially greater than that of a human driver”²¹ In addition, they quote the language appearing in the video image appearing above: consumers “first would be invited to see a video lasting over two minutes, in which the initial frames shouted: ‘THE PERSON IN THE DRIVER’S SEAT IS ONLY THERE FOR LEGAL REASONS. HE IS NOT DOING ANYTHING. THE CAR IS DRIVING ITSELF.’”²² In discussing the named plaintiff from New Jersey, a Mr. Tom Milone, the plaintiffs say, “Tom understood the video to explain that a driver was just in the vehicle for legal purposes and that this vehicle could drive anyone from point A to B and even let the occupants out and go park itself.”²³

Tesla’s statements imply that Autopilot is a fully automated system, while in fact it only constitutes a driver assistance system, requiring the driver to monitor all driving tasks at all times.²⁴ It is possible to draw a line directly from Tesla’s statements to the factual allegations of the suit and the causes of action alleged. The plaintiffs used Tesla’s own words against it to show that the company allegedly oversold the capabilities of the Autopilot system.

How AV Manufacturers Can Mitigate Risk

What can AV manufacturers do to mitigate its legal risks of unfair and deceptive trade practice actions? First, in the design phase of an AV, or any product for that matter, manufacturers should analyze what capabilities their products have and what capabilities are beyond the company’s current technology. A well-repeated phrase is probably the best guidance to follow: Only promise what you can deliver, and deliver what you promise. It is the discrepancies between promises and what is delivered that create the most legal risk.

Second, AV manufacturers should implement a process to guide development and advertising about the product. Cross-functional teams with representatives from engineering, marketing, sales, finance, legal, and data protection can meet on a regular basis to track the progress of a product through development. If all major groups within the organization

know what the product can and can’t do from an engineering perspective, marketing and sales groups are less likely to misunderstand, or worse misrepresent, the capabilities of the product.

Finally, AV manufacturers should include product counsel in decisions about design and marketing. It is probably unrealistic to have product counsel approve every advertisement a manufacturer puts out. Nonetheless, if product counsel works with sales and marketing on some advertising, providing guidance on what statements could create legal risk for the company, sales and marketing can internalize product counsel’s guidance on things marketing collateral can say and what would create risk. Regular training sessions and refresher discussions would provide additional guidance.

AV manufacturers can manage their legal risks of unfair and deceptive trade practice claims. To do so, they must take steps and create procedures and infrastructure to support sound legal judgment about product design and marketing campaigns. Implementing these steps would go a long way towards reducing legal risk. ♦

Endnotes

1. 119 Cal. App. 3d 757, 174 Cal. Rptr. 348 (1981).
2. *Id.* at 778, 174 Cal. Rptr. at 363.
3. There is still some debate or uncertainty about the sudden acceleration phenomenon and whether floor mats or human error might have caused these accidents.
4. No. CV11-04479 CBM (RZx) (C.D. Cal. Complaint filed May 24, 2011).
5. 15 U.S.C. § 45.
6. *Id.* § 45(a)(1).
7. *Id.* § 45(a)(2).
8. Some claims or promises may constitute “puffery”—“exaggerated advertising, blustering, and boasting upon which no reasonable buyer would rely.” *Southland Sod Farms v. Sower Seed Co.*, 108 F.3d 1134, 1145 (9th Cir. 1997). I am not referring to avoiding puffery, which may have its place in advertising. A company may refer to its product as “the best in the land.” No consumer would take such a slogan seriously, and therefore no consumer could be misled by it. What I am referring to is statements that are objective, verifiable, or measurable. If an AV manufacturer says its cars have a firewall to protect against cyber attack and the cars, in

fact, don’t have a firewall, then an investigation easily could show that the advertised firewall is missing.

9. J. Howard Beales, Former Dir., Bureau of Consumer Prot., The FTC’s Use of Unfairness Authority: Its Rise, Fall, and Resurrection, Address at Marketing & Public Policy Convention (May 30, 2003), available at <https://www.ftc.gov/public-statements/2003/05/ftcs-use-unfairness-authority-its-rise-fall-and-resurrection>.

10. For a thorough discussion of how the FTC could enforce consumer protection norms using Section 5 of the FTC Act, see Professor Woodrow Hartzog’s article on this topic. Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785 (2015).

11. CAL. BUS. & PROF. CODE § 17200.

12. *Id.* § 17204.

13. *Id.* § 17500.

14. CAL. CIV. CODE §§ 1750–1784.

15. *Id.* § 1770(a).

16. *Id.* § 1770(a)(5).

17. No. 5:17-cv-02193 BLF (N.D. Cal. Complaint filed Apr. 19, 2017).

18. Tesla, *Tesla Autopilot 2.0*, YouTube (Oct. 20, 2016), <https://www.youtube.com/watch?v=C3DbrYx-SN4> (reposted by Daniel as *Tesla Autopilot 2.0 – Level 5 Autonomy. Full Self-Driving Hardware*).

19. Second Am. Class Action Compl. ¶ 1, at 1, *Sheikh v. Tesla, Inc.*, No. 5:17-cv-02193 BLF (N.D. Cal. Complaint filed Apr. 19, 2017).

20. *Id.* ¶ 4, at 2.

21. *Id.* ¶ 38, at 10–11.

22. *Id.* ¶ 38, at 11.

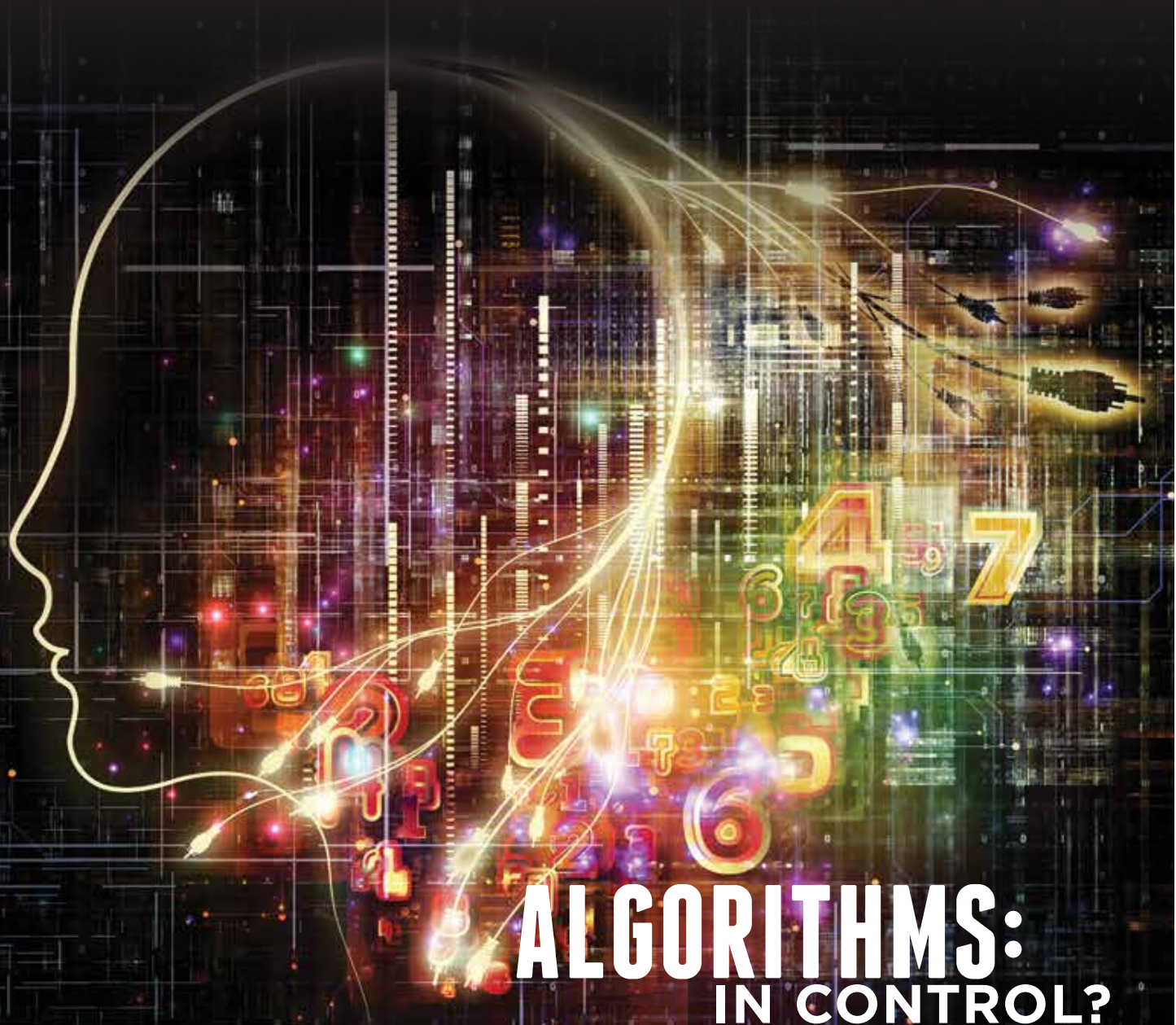
23. *Id.* ¶ 77, at 21.

24. The National Transportation Safety Board called Tesla’s Autopilot system an SAE International “Level 2 automated vehicle system.” NTSB, COLLISION BETWEEN A CAR OPERATING WITH AUTOMATED VEHICLE CONTROL SYSTEMS AND A TRACTOR-SEMITRAILER TRUCK NEAR WILLISTON, FLORIDA, MAY 7, 2016, § 1.3.3, at 9 (Sept. 12, 2017). SAE Level 2 means that the human driver is responsible for continuously monitoring the driving environment and is prepared to take over driving at any time. Another Tesla fatality occurred in late March 2018, which appeared to involve the Autopilot system as well. Tom Krishler, *Tesla: Crash Was Worsened by Missing Freeway Barrier Shield*, WASH. POST (Mar. 31, 2018), https://www.washingtonpost.com/business/technology/tesla-crash-was-worsened-by-missing-freeway-barrier-shield/2018/03/28/55fa6fe8-3290-11e8-b6bd-00841666987_story.html?utm_term=.9489d373b9b6.

IN THIS ISSUE MACHINE LEARNING • MEDICINE • AUTONOMOUS VEHICLES

THE SciTech LAWYER

VOLUME 14 ISSUE 1 | FALL 2017 | SECTION OF SCIENCE & TECHNOLOGY LAW | AMERICAN BAR ASSOCIATION



ALGORITHMS: IN CONTROL?

LISA R. LIFSHITZ, LOIS D. MERMELSTEIN, AND LARRY W. THORPE, ISSUE EDITORS

Published in The SciTech Lawyer, Volume 14, Number 1, Fall 2017. © 2017 American Bar Association. Reproduced with permission. All rights reserved. This information or any portion thereof may not be copied or disseminated in any form or by any means or stored in an electronic database or retrieval system without the express written consent of the American Bar Association.

TABLE OF CONTENTS



2 MESSAGE FROM THE CHAIR

Artificial Intelligence: Revolution or Evolution?

By Eileen Smith Ewing, Chair 2016–17

4 A SIMPLE GUIDE TO MACHINE LEARNING

“Artificial intelligence” (AI) usually refers to machine learning. Machine learning uses algorithms to perform inductive reasoning, figuring out “the rules” given the factual inputs and the results. Applying those rules to new sets of factual inputs can deduce results in new cases. Lawyers are already using machine learning to help with legal research, evaluate pleadings, perform large-scale document review, and more.

By Warren E. Agin

10 ARTIFICIAL INTELLIGENCE IN HEALTH CARE: APPLICATIONS AND LEGAL ISSUES

Big data and machine learning are enabling innovators to enhance clinical care, advance medical research, and improve efficiency, through the use of “black-box” algorithms that are too complex for their reasoning to be understood. Safety regulation, medical malpractice and product liability claims, intellectual property, and patient privacy will impact the way black-box medicine is developed and deployed.

By W. Nicholson Price II

14 AI AND MEDICINE: HOW FAST WILL ADAPTATION OCCUR?

Computers excel at working with “structured data,” such as billing codes or lab test results, but human medical judgment and doctor’s notes are much harder for a computer to analyze. In medicine, the cost of a false positive may be low, but the cost of a false negative can be catastrophic. Thus, applying AI to medicine requires small steps that can supplement and enhance—rather than replace—human decision making.

By Matthew Henson

16 B-TECH CORNER: PRENATAL GENETIC TESTING: WHERE ALGORITHMS MAY FAIL

With noninvasive prenatal genetic testing, the cost of a false positive is not low for parents who relied on the results and unfortunately terminated the pregnancies or who otherwise planned for the arrival of a child with a trisomy condition. A better understanding of the technology that makes these tests possible should lead to better laws and patient outcomes.

By Aubrey Haddach and Jeffrey Licitra

20 ARTIFICIAL INTELLIGENCE AND THE FUTURE OF LEGAL PRACTICE

Despite the alarming headlines, AI will not replace most lawyers’ jobs, at least in the short term. It will create new legal issues for lawyers, such as the liability issues of autonomous cars and the safety of medical robots, and will transform the way lawyers practice, with technology-assisted review, legal

analytics, and practice management assistants, but it will be an evolution, not a revolution.

By Gary E. Marchant

24 WHEN LAW AND ETHICS COLLIDE WITH AUTONOMOUS VEHICLES

Thought experiments can be used to study ethical issues involving autonomous vehicle (AV) algorithms. How should a manufacturer program an AV to respond to an inevitable crash, where continuing on the path will injure a large group, steering away will injure a single individual or small group, and attempting to avoid the collision may injure all? The legal and moral solutions may not be the same.

By Stephen S. Wu

28 ARTIFICIAL INTELLIGENCE AND THE LAW: MORE QUESTIONS THAN ANSWERS?

Current U.S. legislation involving AI is principally concerned with data privacy and autonomous vehicles. The European General Data Protection Regulation (GDPR) will give citizens the right to demand an account of how an adverse decision was achieved. This will require transparency in AI systems, which will raise intellectual property and privacy issues that will have to be reconciled with legislation or in the courts.

By Kay Firth-Butterfield

32 BUILDING ETHICAL ALGORITHMS

Ethical review of automated decision-making systems is a necessary prerequisite to the large-scale deployment of these systems. Several established frameworks provide ethical principles to guide organizations’ best practices around technology design and data use, and can be adapted to big data analytics, automated decision-making systems, and AI.

By Natasha Duarte

38 WHY CHANGES IN DATA SCIENCE ARE DRIVING A NEED FOR QUANTUM LAW AND POLICY, AND HOW WE GET THERE

We are living in a Newtonian age with respect to legal and policy issues for emerging technologies, content with traditional approaches and relying on the slow accretion of precedent. If we do not make the leap to quantum policy, embracing a duality where conflicting rights and ideals are balanced and encouraged to thrive at the same time, our entire ecosystem of jurisprudence and privacy rights will suffer.

By April F. Doss

43 IN MEMORIAM: CHARLES RAY “CHAS” MERRILL

The Section celebrates the life of Chas Merrill, a pioneer, intellect, and patient mentor who was a key leader in the Information Security Committee.

By Stephen S. Wu and Michael S. Baum



WHEN LAW AND ETHICS COLLIDE WITH AUTONOMOUS VEHICLES

From time to time, busy practicing lawyers face ethical issues of the kind taught in professional responsibility law school classes and continuing legal education courses. However, they do not often discuss the kinds of general ethical issues that academics and professional moral philosophers take up. Recent developments in artificial intelligence and robotics, and autonomous driving in particular, have rekindled interest in ethics throughout the world, and especially in the United States.

Autonomous vehicles (AVs) have captured the imagination of writers in popular media. Living close to the garage where Waymo (the new Google affiliate) houses its AVs in Mountain View, California, I feel like I am living in the AV capital of the world, as I frequently see AVs navigating the streets around my home in Los Altos. Nearby Tesla has deployed a driver assistance

system in its cars and intends to deploy fully automated vehicles in two years. Companies are also working on freight truck automation, and their work eventually will result in fully automated trucks.

AV manufacturers will rely on sophisticated algorithms to control AVs. Software implementing such algorithms depends on inputs from sensors, such as light detection and ranging (LiDAR), radar, cameras, and GPS. The software analyzes the AV's location, position relative to the road, and upcoming obstacles. These algorithms then determine the best path to follow and cause the AV throttle, brake, and steering to follow the planned path. A group of moral philosophers has raised ethical questions about these algorithms. In particular, this group asks how AVs should behave when accidents are about to occur. What is the moral way to design AV algorithms?

Should they try to preserve the maximum number of lives (assuming they are sophisticated enough to engage in such a calculation)? Or should they avoid doing harm to innocent pedestrians, bystanders, and passengers? Does the manufacturer owe any special ethical duties to the purchaser of the AV or the AV occupants, as opposed to occupants of other vehicles or those outside the AV? Many of the media stories raising these ethical issues rely on the work of Professor Patrick Lin of California Polytechnic State University.

Professor Lin likes to use "thought experiments" to explain ethical dilemmas. Thought experiments are "similar to everyday science experiments in which researchers create unusual conditions to isolate and test desired variables"¹ and are similar to the hypotheticals law professors use to teach legal subjects. Thought experiments can be used to study ethical issues



BY STEPHEN S. WU

involving AV algorithms. Indeed, the last administration's Department of Transportation policy on highly automated vehicles specifically mentions ethical issues in programming AVs: "Manufacturers and other entities, working cooperatively with regulators and other stakeholders (e.g., drivers, passengers and vulnerable road users), should address these situations to ensure that such ethical judgments and decisions are made consciously and intentionally."²

Of Trolleys and Autonomous Vehicles

Perhaps the most famous thought experiment is the so-called "trolley problem." As the name suggests, the trolley problem involves a runaway trolley. British philosopher Philippa Foot invented the "trolley problem" and first introduced it in 1967.³ American philosopher Judith Jarvis Thomson

expanded on the trolley problem in a 1985 *Yale Law Journal* comment,⁴ which is the more common formulation of the thought experiment: a runaway trolley is heading down the track toward five workers and will soon run over them if no intervention occurs. A spur of track leads off to the right, but there is one worker on the track. A bystander is standing by a switch. If the bystander throws the switch, the trolley will turn onto the spur, saving the five workers, but killing the single worker on the spur.⁵

If the bystander does nothing, the bystander would not be killing anyone. The bystander would merely be "allowing" the five to die. Throwing the switch would involve killing just one person. Some philosophers such as Jarvis Thomson have the view that it is better to maximize the number of lives saved in situations like this. Others such as Foot disagree, saying that it

is worse from an ethical standpoint to cause harm than it is to allow harm to happen, even if the consequences are worse. The trolley problem teases out the moral philosopher's dilemma: is it better to throw the switch and save more lives (five versus one), or is it better (for the bystander) to do nothing in order to avoid causing harm to anyone?

Professor Lin has applied the trolley problem to AVs by posing the following thought experiment:

[Y]ou are about to run over and kill five pedestrians. Your car's crash-avoidance system detects the possible accident and activates, forcibly taking control of the car from your hands. To avoid this disaster, it swerves in the only direction it can, let's say to the right. But on the right is a single pedestrian who is unfortunately killed.⁶

WHEN FACED WITH AN INEVITABLE CRASH, SHOULD AUTONOMOUS VEHICLES BE PROGRAMMED TO SAVE MORE LIVES OR AVOID CAUSING HARM?

News writers have (with or without crediting Professor Lin) repeated this and similar scenarios in numerous recent news articles.⁷ Philosophers continue to debate the question of whether it is better to save more lives or avoid doing harm. To the extent there is any consensus, a recent survey showed that philosophers favored throwing the switch in the trolley problem.⁸ Thus, if an AV manufacturer hires professional philosophers to advise it on how to design AV algorithms, they are likely to advise the manufacturer to program the AV to steer away from a large group at the cost of running over a single individual.

The Legal Trolley Problem Dilemma

As a practicing lawyer, I was curious. What would the legal consequences be if an AV manufacturer followed a philosopher's advice and tried to "do the right thing" in trolley problem situations? What would happen if the

Stephen S. Wu (ssw@svlg.com) is a shareholder in Silicon Valley Law Group in San Jose, California, and practices in the areas of information technology, security, privacy, and intellectual property compliance, litigation, transactions, and policies. He served as the 2010–2011 Chair of the ABA Section of Science & Technology Law. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1522240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF. The author is grateful for the support of the NSF, in addition to Michael Wu for making useful suggestions to improve this article, as well as his editorial assistance.

manufacturer programmed its AVs to steer away from a large group and toward a single individual or small group when they anticipate that a crash is inevitable? For the remainder of this article, I imagine a hypothetical manufacturer (Manufacturer) has implemented just such an algorithm. And I imagine that an accident occurs where the AV steers away from five people (the Five), but at the cost of striking and killing a single individual (the One). I assume that the One was an innocent bystander or pedestrian, rather than a jaywalker or someone engaging in wrongful conduct. I also assume that if the AV had attempted to avoid collision altogether, the AV may have made things worse—it may have killed all six people. I imagine that a representative of the One files a complaint against the Manufacturer.

The most common causes of action in a suit claiming a defect in a product include strict products liability, negligence, breach of warranty, and statutory violations for unfair or deceptive trade practices. With each claim, counsel for the representative of the One would contend that the feature of swerving toward the One made the AV defective. But even worse, the Manufacturer's conduct appears intentional. Indeed, the Manufacturer made a deliberate decision to cause the AV to swerve toward the One (or someone similarly situated to the One). The representative may even assert a cause of action for battery, the essence of which is harmful contact intentionally done.⁹ On its face, the representative seems to have a strong case.

The Manufacturer would not have fared any better if it programmed the AV to do nothing, allowing the AV to run over the Five. If the AV killed the Five, representatives of the Five could

file suit against the Manufacturer, contending that the Manufacturer had a safer alternative design: it could have programmed the AV to run over the One. Thus, it appears the Manufacturer is in a no-win situation.

Possible Defenses

The Manufacturer might turn to traditional defenses recognized in the law to avoid the dilemma. For instance, it could assert a necessity defense, saying that running over the One was necessary to save lives. Under the necessity doctrine, "it has long [been] recognized that '[n]ecessity often justifies an action which would otherwise constitute a trespass, as where the act is prompted by the motive of preserving life or property and reasonably appears to the actor to be necessary for that purpose.'"¹⁰ The private necessity defense thus serves as a justification for a non-governmental defendant's conduct where the defendant's act causes harm, but the defendant acted to prevent an even worse harm. However, necessity is likely to be unavailing as a defense for the Manufacturer. In its traditional form, the necessity defense justifies acts of trespass or damage to personal property, but not bodily injury.¹¹ In our hypothetical case, the AV killed the One and thus does not apply.

Another possible defense is the defense of third persons. Similar to self-defense, the Manufacturer might try to argue that its use of force against the One is justified on order to defend the Five against harm. The Restatement (Second) of Torts provides that an actor can defend any third person from wrongful injury by the use of force.¹² However, the Manufacturer's argument will fail because in our hypothetical case the One was not acting wrongfully. To the contrary, we have assumed that

the One was an innocent actor. There is no wrongful conduct for the Manufacturer to defend against, and thus the defense does not apply.

A third defense the Manufacturer could try to assert is the “sudden emergency” doctrine, also known as the “imminent peril” doctrine. “[I]f an actual or apparent emergency is found to exist the defendant is not to be held to the same quality of conduct after the onset of the emergency as under normal circumstances.”¹³ Cases involving the sudden emergency doctrine in the car accident context involve split-second decisions of drivers in a difficult position. The facts of some of these cases sound like real-world trolley problems.¹⁴ The defense recognizes that an actor in such situations cannot be held to the same standard of care as when an actor is calm in normal circumstances. However, the problem for the Manufacturer is that the Manufacturer is considering how to program an AV in the ordinary course of the design process, far from any imminent accident. The sudden emergency doctrine applies only when, *at the time of the actor’s conduct causing the accident*, the actor faced a sudden choice between two or more actions. Here, the Manufacturer’s programming decision occurred long before the accident. The Manufacturer was not facing a sudden decision. To the contrary, we have assumed that the Manufacturer undertook a careful and deliberate analysis of how to design its AV algorithms and made a choice to program the AV to steer toward the One. No sudden emergency was occurring during the design process. Accordingly, the defense does not apply.

Resolving the Liability Dilemma

Because the traditional defenses offer no protection, the Manufacturer has no easy way out of the liability dilemma. As the law currently stands, I believe the only way for the Manufacturer to limit its legal liability in the trolley problem scenario is to program its AVs to attempt to avoid collision. It should neither steer toward the One nor allow the AV to run over the Five. Rather, it should try to maximize collision avoidance.

I recognize three problems with this approach. First, we have assumed that collision avoidance may make things worse and the AV may end up hurting or killing all six people. Nonetheless, it is more legally defensible and would, as a practical matter, sit better with a jury: the Manufacturer did all it could to save everyone’s life. If the accident ended up killing all six, then at least the Manufacturer tried to save lives.

Second, my position is implicitly at odds with the trolley problem thought experiment. I am implicitly rejecting what appears to be a false choice between running over the Five or running over the One.

Finally, I recognize that my choice of collision avoidance of the “legal” solution is not the one philosophers would consider “moral.” Law and morality sometimes diverge. Conduct we consider immoral may be legal, and some conduct considered to be morally permissible may be illegal. This is one more case in which law and morality may come to different conclusions. Given the liability dilemma, the only way to immunize the Manufacturer trying to “do the right thing” and allow it to program AVs to steer toward the One is to change the law through legislation or regulations.

Trolley problems are useful starting points for analyzing the ethical issues of programming AVs, if nothing else because they spark discussion among the media and their audience. Some people reject the real-world relevance of the trolley problem, but the principles gleaned from it will aid manufacturers in deciding how to program AVs. More generally, injecting discussions of ethics raises the awareness of ethical dimensions to AV design and manufacturers’ decisions, and that is a good thing. ♦

Endnotes

1. Patrick Lin, *Why Ethics Matters for Autonomous Cars*, in AUTONOMOUS DRIVING: TECHNICAL, LEGAL AND SOCIAL ASPECTS 69, 75 (Markus Maurer et al. eds., 2016).

2. U.S. DEP’T OF TRANSP. & NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., FEDERAL AUTOMATED VEHICLES POLICY:

ACCELERATING THE NEXT REVOLUTION IN ROADWAY SAFETY 26 (2016). Likewise, the Trump administration’s latest guidance on AVs acknowledges that manufacturers should deliberate concerning ethical issues. U.S. DEP’T OF TRANSP. & NAT’L HIGHWAY TRAFFIC SAFETY ADMIN., AUTOMATED DRIVING SYSTEMS 2.0: A VISION FOR SAFETY 1 n.1 (2017).

3. Philippa Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, OXFORD REV., 1967, at 5, 8.

4. Judith Jarvis Thomson, *The Trolley Problem*, 94 YALE L.J. 1395, 1395 (1985).

5. See *id.* at 1397.

6. Lin, *supra* note 1, at 79.

7. See, e.g., *Why Self-Driving Cars Must Be Programmed to Kill*, MIT TECH. REV. (Oct. 22, 2015), <http://www.technologyreview.com/view/542626/why-self-driving-cars-must-be-programmed-to-kill/>.

8. David Bourget & David J. Chalmers, *What Do Philosophers Believe?* 16 (Nov. 30, 2013), <https://philpapers.org/archive/BOUWDP>.

9. See, e.g., 5 B.E. WITKIN, SUMMARY OF CALIFORNIA LAW § 383, at 599 (10th ed. 2005).

10. *People v. Ray*, 981 P.2d 928, 935 (Cal. 1999).

11. See RESTATEMENT (SECOND) OF TORTS § 263 (AM. LAW INST. 1965) (“One is privileged to commit an act which would otherwise be a trespass to the chattel of another or a conversion of it, if it is or is reasonably believed to be reasonable and necessary to protect the person or property of the actor, the other or a third person from serious harm.” (emphasis added)).

12. See *id.* § 76.

13. Bill Hollingsworth, Note, *The Sudden Emergency Doctrine in Florida*, 21 U. FLA. L. REV. 667, 671 (1969).

14. See, e.g., *Myhaver v. Knutson*, 942 P.2d 445, 446 (Ariz. 1997) (en banc) (involving a driver who faced a choice of driving straight into a car driving the wrong way in his lane or crossing the yellow line into oncoming traffic).

AI and Robots in the Healthcare Setting



AI & Robotics in the Healthcare Setting

Marissa Urban, Google

Steve Mutkoski, Microsoft

Christopher Hess, University of California School of Medicine, San Francisco,

Why Healthcare Needs AI: The Quadruple Aim



- **Enhancing Population Health and Improving Care Outcomes**
 - Operational and Clinical AI that can help providers make better decisions faster
 - AI systems that derive insight from a wide array of data sets associated with population health management
 - AI systems that overlay of other large data sets, including genomic data and social determinants data
- **Lowering Healthcare Costs**
 - Countries spend between roughly 6 percent and 18 percent of their gross domestic product (GDP) and administrative costs in US top 31%
 - Top 10 AI applications can create \$150 billion in annual savings for the United States healthcare economy by 2026
- **Improving Patient Access and Experience**
 - AI-enabled “Virtual Health” solutions that enable patients to obtain care on their own terms and share remotely monitored data with care providers
 - AI-enabled diagnostic aids that will enable patients to get more from visits to their primary care provider.
- **Improving Health Worker and Clinician Experience**
 - Growing administrative demands coupled increasing shortages of care workers are driving records levels of burn-out and dissatisfaction.
 - AI-enabled tools can be deployed to drastically improve clinician and healthcare team satisfaction, by removing time-consuming and often mundane tasks, while enabling the health care team to more quickly screen, diagnose, treat, and monitor patients.



Where is AI in Healthcare?

- *Operational vs Clinical AI*
- *Computer Vision, Natural Language Processing*
- *Solutions and Solution Areas for healthcare AI*



Bias

a systematic as opposed to a random distortion of a statistic as a result of sampling procedure.



Healthcare AI and Patient Privacy

Emerging Technology Runs into Timeless Values



Safety and Efficacy of AI Systems

We have a longstanding regulatory framework for medical devices, are we clear which AI systems fall within that framework and outside it?



Other Topics

- *Transparency and Trust*
- *Cultural Adoption Issues*
- *Legal Liability*

AI-Enabled Technologies and Healthcare

Steve Mutkoski

Microsoft Worldwide Healthcare
Mountain View, California

From one perspective, “Artificial Intelligence” or “AI” is not new to us, but instead just a significant step up in analytical power, due to advances in machine learning (ML)¹ and massive cloud data centers, that offers the promise to derive insight from a universe of data we have begun to accrue at greater and greater volumes. The other side of the coin is that these technology advances are so profound that they potentially exceed our human capacity in other areas, such as to understand how a system arrived at outputs or recommendations, what hidden biases it might have incorporated, and in the context of healthcare uses, where a system that seems to have made very strong predictions and been accurate in the past why it might suddenly appear to veer off course and result in injury or death to a patient. For the past year, a number of different healthcare community stakeholders have convened discussions and working groups and published white papers, in an effort to begin the important task of identify the opportunities and pitfalls associated with the use of AI in healthcare. In October 2018, two colleagues and I published an article “Realizing the Potential for AI in Precision Health”² that we hoped would serve as a frame for some of the more granular issues that we think stakeholders must address. In this note, I highlight several of the more challenging issues that have arisen in the groups with which we have engaged, with an aim toward demonstrating how much work we have ahead of us:

(1) There are some critical foundational issues around terminology and how we talk about AI writ large as well as in the context of healthcare that I believe will either ensure we embed AI-enabled technologies into our healthcare systems in the near future or cause doctors, patients, policymakers and regulators to spurn it.

(2) In many contexts, the accuracy of AI systems, their potential bias and other challenges such as “explainability” are somewhat trivial. Most of us don’t really care if the movie recommendations one video platform gives us are poor or the ads we get served on one website are meaningless, incorrect or even slightly offensive. But we will care when in the context of healthcare, AI-enabled systems don’t work as we expect them to and result in efficacy or safety concerns.

(3) The role of machines and then technology in the workforce and the impact of these developments on jobs has been a topic of debate for decades now. In terms of AI, there is a silver lining in the healthcare context in that most healthcare systems are facing increasingly acute shortages of care team workers, and AI offers an opportunity to help address those shortages. But as AI-enabled technologies become embedded in more places in our healthcare system, established rules of liability-- and in particular the apportionment of and standards for liability between technology and humans-- will need to be revisited.

(4) AI systems have complexity along a number of dimensions, including that they are essentially “self-learning.” The self-learning nature of an AI-system can either be “locked”-- once a desired state is reached with initial “training data”-- or “continuous” where the system will continue to learn based on new data that it consumes once out in the wild. Our entire conception of medical devices (and the regulation of them) is based on the concept of a static device that is developed, tested, clinically validated, marketed and used without modification. Continuous learning systems may be a technical possibility today, but in the context of healthcare, we may not be ready for them. Accordingly, the vast majority of what we see in AI-enabled systems for healthcare today and for the foreseeable future will be “locked” AI.

Walking Before We Run: How We Talk About AI Generally and in Healthcare

The way we talk about new technology can have a significant impact on how users and subjects of that technology react to it. My own view is that it does not help us as stewards of new technology when we use headlines like “The Robots are Coming” or when we spring from current capabilities to where AI capabilities might realistically be in three or five years. The notion that AI will automate or completely take over human direction or decision-making is an interesting one, but it is probably not where we should start with complex AI systems in areas like healthcare. To that end, many stakeholders have suggested that we make clear that at least today we mean to use AI-enabled systems to augment human capabilities rather than to make a process or system entirely autonomous. Indeed, part of the reason we increasingly use the term “AI” or “AI-enabled” instead of “Artificial Intelligence” is that some stakeholders are more aligned with a concept of “Augmented Intelligence.”³ It may sound pedantic to debate terminology, but the reality is that how we articulate this early wave of machine-learning driven innovation will in turn have subtle but important implications for how that technology is tested, validated, regulated, used and accepted by those who will interface with it. My hope is that with AI in the healthcare sector, by focusing initially on the notion of augmenting human capabilities, we can avoid some of the pitfalls that the transportation industry has encountered when some innovators insisted on jumping beyond terminology like “driver assist” to “self-driving.”

We must make sure we can walk before we try to run. More importantly we must make sure that patients, care providers, health payors and other stakeholders believe we can walk before we start to run. The good news on this front is that we have already seen some amazing successes with the use of ML to create AI-enabled systems that assist or augment the capabilities of care providers and drive better health outcomes and likely lower cost of care. In 2017, Microsoft received 510(k) clearance from the US FDA to market technology called Microsoft Radiomics, which is “a software-only medical device intended for use by trained radiation oncologists, dosimetrists and physicists to derive optimal organ and tumor contours for input to radiation treatment planning.” That technology leveraged more than eight years of research in computerized medical image analysis, computer vision and machine learning. It applies well tested, state-of-the-art algorithms for the assisted delineation of anatomical structures of interest in three-dimensional, clinical radiological scans.” Aside from that technical and regulatory description, what Microsoft Radiomics or “InnerEye”⁴ does as a practical matter is leverage computer vision ML capabilities to dramatically cut down on the amount of time it takes a clinician to mark up the thousands or segments of an MRI for purposes of creating a treatment or surgery plan for a cancer patient.

More recently, a number of firms have begun to share advances in AI-enabled technologies that assist with the diagnosis of diabetic retinopathy. IDx-DR is a software program that uses an artificial intelligence algorithm to analyze images of the eye taken with a retinal camera to detect certain diabetes-related eye problems.⁵ Note that IDx-DR has been described in many ways, often incorrectly. It is “the first medical device to use artificial intelligence to detect greater than a mild level of the eye disease diabetic retinopathy in adults who have diabetes” that the FDA has permitted to be marketed in the US. What IDx-DR and similar technologies promise is the ability for a patient to be more accurately “screened” at a primary care physician for diabetic retinopathy-- a capability that is found to be wanting in primary care physicians-- enabling a smaller subset of more accurately identified higher risk patients to be referred on to a specialist for further diagnosis and treatment.

There are countless other examples of existing and emerging AI-enabled technologies that we should expect to hear more about in the coming months. The two examples above rely on so-called “computer vision” machine learning.⁶ As my Microsoft Research colleagues noted in their FDA submission, computer vision-based ML is hardly new and that is a critical fact for us to understand as we explore how to safely introduce AI-enabled technologies into the healthcare sector. Image-based AI-enabled

technologies are likely to be an important tip of the spear for us and one where mature technical capabilities, a desire from clinicians for these augmented capabilities and an immediate and tangible patient benefits are likely to help us gain confidence that we can, indeed, walk.

Does It Work? How Does It Work? How Do We Know?

Harken back to 1937, when over 100 people died after taking a drug called Elixir Sulfanilamide that was freely marketed with no regulatory oversight to ensure that it did what the distributor promised and did not cause injury or death.⁷ There are a similar bundle of important regulatory oversight questions for AI-enabled systems that will be used in the healthcare system. And while it might seem at first that “there is nothing new here” in that the two examples listed above both went through time tested regulatory review, the reality is we may not have a perfect fit with existing regulatory frameworks.

One challenge will be the concept of “explainability” of the underlying algorithms, which will likely put stress on current practices for validating the underlying clinical science of a medical device (whether that validation is taking place within the entity creating the AI-enabled technology or when that technology is submitted to a regulator). The self-learning nature of machine learning means that ML can make sense of a massive amount of data in the sense that it might identify some obscure or hidden relationships between various factors, but it cannot necessarily explain to us the “why” behind those hunches and we may not be able to deduce the “why” ourselves. That’s important to understand, because without a deep understanding of the “why”, we might miss out on advanced notice of important consequences, such as that bias in the training data results in the algorithm being inaccurate for a certain ethnic, racial, age or socioeconomic group.

As we move from fairly narrow AI-enabled systems, such as one of the image-based technologies already discussed, to more complex systems that involve unstructured data (such as speech and text) from a variety of systems and sensors, the challenges around explainability multiply. Ochsner Health System announced in early 2018 that it has leveraged machine learning capabilities from EPIC and Microsoft to build a system to detect patients’ potential adverse health events more quickly and accurately, in effect allowing them to intervene with patients proactively, rather than reactively.⁸ During a 90-day pilot of the system, Ochsner reported that it recorded a 44 percent reduction in adverse events outside of the Intensive Care Unit. That is an impressive outcome for the pilot and there is no doubt that this AI-enabled system has the capability to derive insight in real time from thousands of patients, sensors and health records in ways that humans can only do after the fact. But we will need to probe more deeply into the specific results that these systems return as they operate across increasingly larger healthcare provider facilities and systems. Within that 44% reduction reported by Ochsner, are the reductions uniform across age, ethnicity and other groupings? Do we know if the system potentially diverted resources away from patients who later suffered adverse consequences?

These important questions about the safety and efficacy of AI-enabled systems, and the methods we can use to verify safety and efficacy are complicated further in the third scenario, which involves not a medical device manufacturer but a healthcare provider organization. While the first two AI-enabled systems (Microsoft Radiomics and IDx-DR) were both created with the intent they would be distributed and used by outside clinicians, these healthcare provider systems-- which are internally developed and deployed only within the organization’s facilities-- may well fall outside the jurisdiction of current medical devices regulation.

How then can the healthcare provider, and we as patients, verify that these systems will be safe and efficacious? Doubtless that these pioneering healthcare provider organizations have rigorous internal processes to address many of these questions, but over time we may need to develop standards which all organizations follow as they develop such AI-enabled technologies and deploy them for use on or with

patients. The UK has been a pioneer in this area, releasing its “*Initial code of conduct for data-driven health and care technology*” in September 2018⁹ which includes a Principle 9 “Show evidence of effectiveness for the intended use” which is intended to spur development of Standards that define what evidence should be required to demonstrate the effectiveness and economic impact of digital health innovations. More recently, the UK National Institute for Care Excellence (CARE) released “Evidence standards framework for digital health technologies”¹⁰ an initial effort to develop standards that ensure new technologies are clinically effective and offer economic value.

The Future of Work, Healthcare Worker Shortages and Liability

What’s old is new, the more things change the more they stay the same, there is an element of both as we explore the role that AI will play in the future of work in the healthcare sector. One of the more compelling and entertaining discussions of the impact of technology on the impact of jobs noted:

We’ve been trying to divine how technology will impact our jobs for centuries. Often we’ve got it badly wrong. Here are some unfounded fears and wildly optimistic predictions about the tech that has transformed the way we live, work and communicate.¹¹

What follow are some interesting historical perspectives, including a prediction in 1959 that Artificial Intelligence would take away all of our jobs. Will AI threaten the jobs of doctors, nurses and other care workers? Such fears are likely unfounded. Most, if not all, countries around the world are experiencing severe clinician shortages. And these shortages are only predicted to get worse over the next ten years. In the United States, for example, a report for the Association of American Medical Colleges predicts a shortage of physicians through 2030 under every combination of scenarios modeled. Rather than being a threat to clinicians, AI-enabled tools might well be essential to improving the efficiency of care, thereby mitigating some of the issues resulting from worsening shortages of trained and experienced clinicians.

AI in some senses won’t be a threat or a luxury in the healthcare sector, it will be a necessity. And by extension, so will working through the various regulatory and legal issues already identified, and a few others such as liability for injuries or death caused by AI-enabled systems. As noted in our October article, “there remain more questions about how accountability should be addressed than there are answers.” In particular we wondered how we should the balance of responsibility for use of suggestions provided by AI-enabled technologies between system developers, health care institutions implementing the systems, and health care professionals utilizing the systems in clinical decision making. Implicit in our questions was the fact that those entities are currently governed by differing standards of care, and so the answer to these questions can mean the difference between liability or not, monetary damages for a patient versus no monetary damages. Since the publication of our article, others have commented further on this important point, highlighting that humans providing treatment to patients are governed by the negligence standard whereas a case involving the alleged malfunction of a medical device would be subject to strict liability.

The same thing is sure to happen with AI – where plaintiffs will surely attempt to impose what amounts to absolute liability on the manufacturers of AI-enhanced medical “devices.” Just look at the experience with AI in the automotive field. Every accident involving a self-driving car becomes a big deal, with the AI systems presumed to be somehow at fault. However, the negligence standard, for both auto accidents and medical malpractice, has always been that of a “reasonable man.” That’s the crux of what will be the struggle over AI – when machines take over the functions once performed by human operators (whether drivers or doctors), are they also to be judged by a “reasonable man” negligence standard? Or will strict (really strict) liability be imposed?¹²

The author offers potential solutions, but stresses that aligning the standards for humans and AI-enabled devices will be “critically important in ensuring a successful future for AI in the medical space.” The author also notes the importance of an “alliance between physician and AI interests,” and to that it would be important to add that we should include patients and policymakers. It is possible that if we do not address this asymmetry on the issue of the standard for liability, we might see delayed or diminished deployment of AI-enabled technologies.

To Lock or Not to Lock? Are We Ready to Answer That Question?

As noted, the ML that powers an algorithm is continuous, unless the developer stops the learning process and “locks” the algorithm. At first glance we might think that once an algorithm reaches a level where we believe it is efficacious and safe, why not let it keep learning as it is deployed and encounters additional data? Wouldn't that make it more accurate and more useful? Maybe, but maybe not. It is entirely possible that unknown traits or errors could incorporate the new data in a manner that would make the algorithm less accurate or even harmful to subsequent patients. Without additional testing after ingestion of new data, we just cannot be sure. The notion of “continuous learning” AI is something that we mentioned in our earlier article is in tension with the iterative develop, test/validate and release cycles common in the medical devices sector that in turn reflects the current regulatory framework for medical devices. Anecdotal evidence we have gathered since publication of our article, including conversations with Microsoft engineers working on AI-enabled systems for use in the healthcare and colleagues in various stakeholder organizations, suggests that most AI developers today are fully occupied with safety and efficacy matters related to the development of a locked system. We have found very few active developers who believe that we have sufficient experience with testing and validation or locked AI systems to enable us to articulate requirements and standards that would be required to ensure that continuous or unlocked AI systems can be safely deployed.

Endnotes

¹ See Connected Health Initiative, Key Terminology for AI in Health, (February 2019) available at <https://actonline.org/wp-content/uploads/Artificial-Intelligence-in-Health-Appendix.pdf> (defining Machine Learning in part as “AI with an algorithm that learns and changes without being programmed when exposed to new data.”)

² https://www.americanbar.org/groups/science_technology/publications/scitech_lawyer/2019/fall/realizing-potential-ai-precision-health/

³ See Connected Health Initiative, Key Terminology for AI in Health, (February 2019) available at <https://actonline.org/wp-content/uploads/Artificial-Intelligence-in-Health-Appendix.pdf> (defining Augmented Intelligence as “An alternative conceptualization of AI advanced by a growing number of innovators and embraced by physician organizations to underscore that such systems are designed to aid humans in clinical decision-making, implementation, and administration to help scale health care.”)

⁴ *Project InnerEye: Medical Imaging AI to Empower Clinicians*, available at <https://www.microsoft.com/en-us/research/project/medical-image-analysis/>

⁵ FDA News Release, *FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems*, (April 11, 2018) available at <https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm>

⁶ See Connected Health Initiative, Key Terminology for AI in Health, (February 2019) available at <https://actonline.org/wp-content/uploads/Artificial-Intelligence-in-Health-Appendix.pdf> (defining Computer

Vision as “Algorithms that perceive and develop methods for extracting information from images. This may include object detection, segmentation of objects, tagging, and captioning of an image.”)

⁷ The Scientist, *The Elixir Tragedy, 1937*, (June 2013) available at <https://www.the-scientist.com/foundations/the-elixir-tragedy-1937-39231>

⁸ Ochsner Health System Adopts New AI Technology to Save Lives in Real-time, (February 28, 2018) available at <https://news.ochsner.org/news-releases/ochsner-health-system-adopts-new-ai-technology-to-save-lives-in-real-time>

⁹ <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology>

¹⁰ <https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies>

¹¹ Miriam Quick and Piero Zagami, *What the Distant Past Told Us about Work in the Future*, (March 12, 2018) available at <http://www.bbc.com/capital/story/20180312-historys-unfounded-fears-over-the-future-of-work>

¹² James M. Beck, The Diagnostic Artificial Intelligence Speedbump Nobody's Mentioning (November 18, 2018) available at <https://www.druganddevicelawblog.com/2018/11/the-diagnostic-artificial-intelligence-speedbump-nobodys-mentioning.html>

Artificial Intelligence: How to get it right

Putting policy into practice for safe
data-driven innovation in health and care

ABOUT NHSX

NHSX brings teams from the Department of Health and Social Care, NHS England and NHS Improvement together into one unit to drive digital transformation and lead policy, implementation and change.

NHSX is responsible for delivering the Health Secretary's Tech Vision, building on the NHS Long Term Plan by focusing on five missions:

- Reducing the burden on clinicians and staff, so they can focus on patients;
- Giving people the tools to access information and services directly;
- Ensuring clinical information can be safely accessed, wherever it is needed;
- Improving patient safety across the NHS;
- Improving NHS productivity with digital technology.

ABOUT THIS REPORT

Joshi, I., Morley, J.,(eds) (2019). *Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care*. London, United Kingdom: NHSX.

Although this report has named editors, it results from the collective effort of a great number of individuals who kindly gave up their time to contribute their thoughts, ideas and research. A full list of acknowledgements is provided at the end of the report. There are, however, several key organisations, and individuals who provided input without which this report would not have been possible. With this in mind, we would like to thank:

Tina Woods,
Collider Health

Melissa Ream, Marie-Anne Demestihis and Sile Hertz,
AHSN Network

Anna Steere,
NHSX

Dr. Sam Roberts,
Accelerated Access Collaborative

Ministerial Foreword	6
Executive Summary	10
1. Introduction	14
Definition	14
Opportunities	14
Challenges	17
2. Where Are We Now?	18
3. Developing the Governance Framework	26
Why you need Ethics & Regulation	26
A Code of Conduct	27
Principle 7: Algorithmic Explainability	28
Principle 8: Evidence for Effectiveness	34
Principle 10: Commercial Strategy	36
Self-Assurance Portal	36
Mapping the Regulation Journey	37
Overcoming Regulatory Pain Points	41
4. Clarifying Data Access and Protection	44
Navigating Data Regulation	44
Understanding Patient Data	46
Protecting the Citizen	47
Data Innovation Hubs	48
Data Collaboration at Scale	49
Data Agreements and Commercial Models	52
NHSX Data Framework	55
5. Encouraging Spread of ‘Good’ Innovation & Monitoring the Impact	56
What Does ‘Good AI’ Look Like?	56
1. Precision Medicine	56
2. Genomics	56
3. Image Recognition	57
4. Operational Efficiency	58
Tackling Barriers to Adoption	58
Measuring Impact	58
Real-world evaluation	59
6. Creating the Workforce of the Future	62

7. Developing International Best Practice Guidance	64
Global Digital Health Partnership	64
World Health Organization (WHO) & International Telecommunication Union (ITU)	65
The EQUATOR Network	70
8. Conclusion	72
Appendix: Case Studies	74
Flagship Case Studies	74
Precision Medicine	74
Genomics England	76
EMRAD	78
Non-clinical (operational) applications of AI	80
Cogstack	82
Lessons from Estonia and Finland	83
NHS-R Community	84
NHSX Mental Health	85
Optimam	85
Survey Case Studies	86
Advancing Applied Analytics	86
axial3D	87
BrainPatch	88
Chief AI	88
Concentric Health	89
eTrauma	89
First Derm	90
Forms4Health	91
Google Health	92
iRhythm Technologies	93
Kaido	93
Kortical	94
Lifelight	95
My Cognition	95
Roche Diabetes Care Platform	96
Sensyne Health	96
Sentinel	97
Storm ID	98
Veye Chest	99
References	100
Acknowledgements	106

Ministerial Foreword

We love the NHS because it's always been there for us, through some of the best moments in life and some of the worst. That's why we're so excited about the extraordinary potential of artificially intelligent systems (AIS) for healthcare.

Put simply, this technology can make the NHS even better at what it does: treating and caring for people.

This includes areas like diagnostics, using data-driven tools to complement the expert judgement of frontline staff. In the report, for example, you'll read about the East Midlands Radiology Consortium who are studying Artificial Intelligence (AI) as a 'second reader' of mammogram images, helping radiologists with an incredibly consequential decision, whether or not to recall a patient. In the near future this kind of tech could mean faster diagnosis, more accurate treatments, and ultimately more NHS patients hearing the words 'all clear'.

AIS can also help us get smarter in the way we plan the NHS and manage its resources. Take NHS Blood & Transplant, who are looking at how AI can forecast how much blood plasma a hospital needs to hold onsite on any given day. Or University College London Hospitals (UCLH) who are trialling tools that can predict the risk of missed outpatient appointments.

Most exciting of all is the possibility that AI can help with the next round of game-changing medical breakthroughs. Already, algorithms can compare tens of thousands of drug compounds in a matter of weeks instead of the years it would take a human researcher. Genomic data could radically improve our understanding of disease and help us get better at taking pre-emptive action that keeps people out of hospitals.

But while the opportunities of AI are immense so too are the challenges.

Much of the NHS is locked into ageing technology that struggles to install the latest update, never mind the latest AI tools, so we need a strong focus on fixing the basic infrastructure. That means sorting out the connectivity, standardising the data and replacing our siloed and fragmented systems with systems that can talk to each other. We also need to make sure that staff have the skills, training and support to feel confident in using or procuring emerging technology.

Just as important, as a society we need to agree the rules of the game. If we want people to trust this tech, then ethics, transparency and the founding values of the NHS have to got to run through our AI policy like letters through a stick of rock.

And while we're clear-eyed about the promise of AI we can't let ourselves be blinded by the hype (of which this field has more than its fair share). Our focus has got to be on demonstrably effective tech that can make a practical difference, at scale, right across the NHS, not just the country's most advanced teaching hospitals.

To help us deliver those changes, we've set up NHSX, a new joint team working across the NHS family to accelerate the digitisation of health and care. NHSX's job is to build the ecosystem in which healthtech innovation can flourish for the benefit of the NHS. Crucially it's also been tasked with doing this in the right way, within a standardised, ethically and socially acceptable framework.

Getting these foundations right matters hugely, which is why we are investing £250 million in the creation of the NHS AI Lab to focus on supporting innovation in an open environment where innovators, academics, clinicians and others can develop, learn, collaborate and build technologies at scale to deliver maximum impact in health and care safely and effectively.

The NHS AI Lab will be run collaboratively by NHSX and the [Accelerated Access Collaborative](#) and will encompass work programmes designed to:

- Accelerate adoption of proven AI technologies e.g. image recognition technologies including mammograms, brain scans, eye scans and heart monitoring for cancer screening.h
- Encourage the development of AI technologies for operational efficiency purposes e.g. predictive models that better estimate future needs of beds, drugs, devices or surgeries.
- Create environments to test the safety and efficacy of technologies that can be used to identify patients most at risk of diseases such as heart disease or dementia, allowing for earlier diagnosis and cheaper, more focused, personalised prevention.
- Train the NHS workforce of the future so that they can use AI systems for day-to-day tasks.
- Inspect algorithms already used by the NHS, and those being developed for the NHS, to increase the standards of AI safety, making systems fairer, more robust and ensuring patient confidentiality is protected.
- Invest in world-leading research tools and methods that help people apply ethics and regulatory requirements.

The following report sets out the foundational policy work that has been done in developing the plans for the NHS AI Lab. It also shows why we're so hopeful about the future of the NHS.



Matt Hancock,
Secretary of State



Baroness Blackwood,
Minister for Innovation

Executive Summary

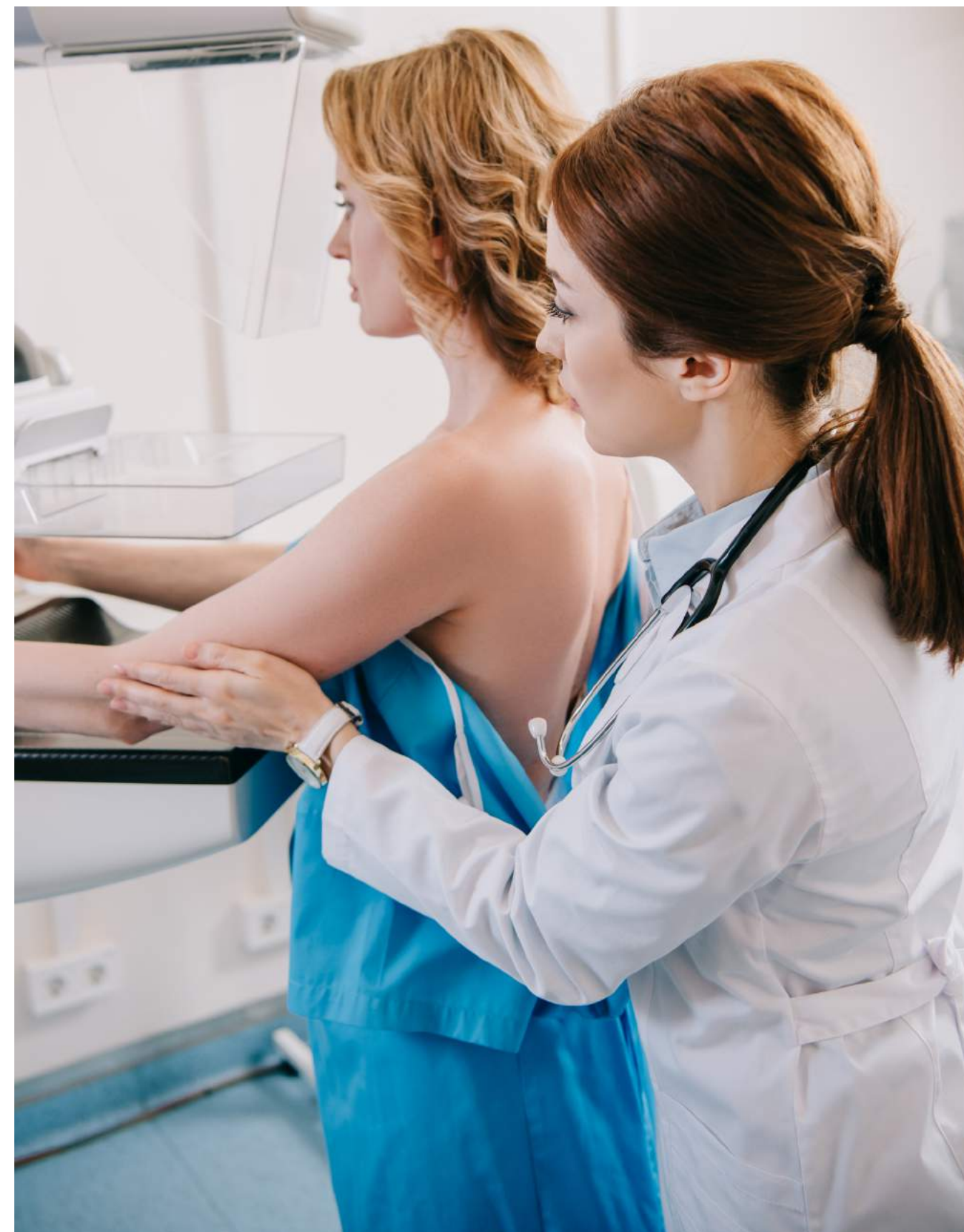
Artificial Intelligence could help personalise NHS screening and treatments for cancer, eye disease and a range of other conditions, for example, while freeing up staff time to spend with patients.

Artificial Intelligence (AI) has the potential to make a significant difference to health and care. A broad range of techniques can be used to create Artificially Intelligent Systems (AIS) to carry out or augment health and care tasks that have until now been completed by humans, or have not been possible previously; these techniques include inductive logic programming, robotic process automation, natural language processing, computer vision, neural networks and distributed artificial intelligence. These technologies present significant opportunities for keeping people healthy, improving care, saving lives and saving money for the pilot digital technologies. It could help personalised NHS screening and treatments for cancer, eye disease and a range of other conditions, for example. Furthermore, it's not just patients who can benefit. AI can also support clinicians, enabling them to make the best use of their expertise, informing their decisions and saving them time.

This report gives a considered and cohesive overview of the current state of play of data-driven technologies within the health and care system, covering everything from the local research environment to international frameworks in development. Informed by research conducted by NHSX and other partners over the past year, it outlines where in the system AI technologies can be utilised and the policy work that is, and will need to be done, to ensure this utilisation is done in a safe, effective and ethically acceptable manner. Specifically:

Chapters 1 and 2 set the scene. They provide an overview of what AI is (and importantly is not), why we believe it is important, and a detailed look at what is currently being developed by the AI ecosystem by evaluating the results of a horizon scanning exercise and our second 'State of the Nation' survey. This analysis reveals that diagnosis and screening are the most common uses of AI, with 132 different AI products identified being designed for diagnosis or screening purposes covering 70 different conditions

Chapter 3 is an in-depth look at the Governance of AI. Building on the Code of Conduct for data-driven technologies, it explores the development of a novel governance framework that emphasises both the softer ethical considerations of the "should vs should not" in the development of AI solutions as well as the more legislative regulations of "could vs could not". In particular it covers key areas such as the explainability of an algorithm, the evidence generation for efficacy of fixed algorithms, the importance of patient safety and what to consider in commercial strategies.



Chapter 4 is all about the data that fuels AI. When engaging with innovators, regulators, commissioners and citizens on AI the one topic that is guaranteed to come up is Information Governance (IG). Protecting patient data is of the utmost importance, which is why IG is crucial, but it should not be seen as a blocker to the use of data for purposes that can deliver genuine benefits to patients, clinicians and the system. This Chapter highlights how we are working collaboratively with key partners across the system (e.g. the Accelerated Access Collaborative, the Office of Life Sciences, Health Data Research UK, Genomics England, Academic Health Science Network) to clarify the rules of IG and streamline access to data for good through specific programmes such as the Digital Innovation Hubs.

Chapter 5 covers adoption and spread. Considering the sometimes negative impact the complexity of the NHS as a sociotechnical system has on the spread of important innovation, it covers the actions being taken to encourage adoption. However, given the challenges involved in the practical implementation of AI we do not want to encourage adoption for the sake of adoption, so it also covers ‘what good looks like’ and how we can monitor the impact of the introduction of AI over time so that good stays good further downstream.

Chapter 6 comes back to the people of the NHS. Building on the work of Health Education England and the Topol Review, it highlights the challenges faced by the workforce in the development, deployment and use of AI and what needs to be done in order to ensure they have the skills that they need to feel confident in using AI in clinical practice safely and effectively. Crucially it highlights how again we cannot do this alone and must work closely with national centres of data science training such as the Alan Turing Institute.

Chapter 7 goes international. Health data is not only generated in England and the AI technologies that are trained and tested on it are not developed only in England. Instead the AI ecosystem is truly international and there is, therefore, a need for international collaboration and agreement of standards, frameworks and guidance. For this reason, this chapter highlights the ongoing work of the Global Digital Health Partnership, the World Health Organisation and the EQUATOR network in developing these with us as a key partner.

Chapter 8 concludes with the NHS AI Lab. It brings together all the information included in the previous chapters to highlight why we know that the Lab is needed and why we think it will be crucial in helping us achieve our aims of:

- promoting the UK as the best place in the world to invest in healthtech.
- providing evidence of what good practice looks like to industry and commissioners.
- reassuring the public, patients and clinicians that data-driven technology is safe, effective and protects privacy.
- allowing the government to work with suppliers to guide the development of new technology so products are suitable for the health and care system in the future.
- building capability within the system with In-house expertise to prototype and develop ideas.
- making sure the NHS gets a fair deal from the commercialisation of its data resources and expertise.

1. Introduction

Dr. Indra Joshi
& Jessica Morley

DEFINITION

Despite being a well-established field of computer science research, Artificial Intelligence (AI) is difficult to define and, as such, numerous definitions exist, including:

“the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment”¹

“a cross-disciplinary approach to understanding, modelling, and replicating intelligence and cognitive processes invoking various computational, mathematical, logical, mechanical, and even biological principles and devices”²

“the science of making machines do things that would require intelligence if done by people”³

The third definition is the oldest, stemming from the field’s founding document “Proposal for the Dartmouth Summer Research Project on Artificial Intelligence” (1955). However, it is the most applicable to the uses of Artificial Intelligence for health and social care.

OPPORTUNITIES

In the context of health and care, a broad range of techniques (e.g. inductive logic programming, robotic process automation, natural language processing, computer vision, neural networks and distributed artificial intelligence such as agent based modelling⁴) are used to create Artificially Intelligent Systems (AIS) that can carry out medical tasks traditionally done by professional healthcare practitioners. The number of medical or care-related tasks that can be automated or augmented in this manner is significant. A summary of the areas of care in which such automated tasks could make a difference is presented in Figure 1.

Diagnostics	Knowledge Generation	Public Health	System Efficiency	P4 Medicine
<ul style="list-style-type: none"> • Image Recognition e.g. • Symptoms Checkers and Decision Support • Risk Stratification 	<ul style="list-style-type: none"> • Drug Discovery • Pattern Recognition • Greater knowledge of rare diseases • Greater understanding of causality 	<ul style="list-style-type: none"> • Digital epidemiology • National screening programmes 	<ul style="list-style-type: none"> • Optimisation of care pathways • Prediction of Do Not Attends • Identification of staffing requirements 	<ul style="list-style-type: none"> • Prediction of deterioration • Personalised treatments • Preventative advice

Figure 1 ⁵⁻¹²

This range of potential use cases for AI in health and care highlights the scale of the opportunity presented by AI for the health and care sector. This is why:

1. The NHS Long-Term Plan sets out the ambition to use decision support and AI to help clinicians in applying best practice, eliminate unwarranted variation across the whole pathway of care, and support patients in managing their health and condition.
2. The future of healthcare: our vision for digital, data and technology in health and care outlines the intention to use cutting-edge technologies (including AI) to support preventative, predictive and personalised care.
3. The Industrial Strategy AI Mission sets the UK the target of “using data, Artificial Intelligence and innovation to transform the prevention, early diagnosis and treatment of chronic diseases by 2030.”

We believe that the UK can be a world leader in this area for years to come - a core aim of the Office for Artificial Intelligence (OAI).



There are significant ethical and safety concerns associated with the use of AI in health and care.

CHALLENGES

As much as we believe in the power of AI to deliver significant benefits to health and care, and the wider economy, we also know that there are significant ethical and safety concerns associated with the use of AI in health and care.

If we do not think about transparency, accountability, liability, explicability, fairness, justice and bias, it is possible that increasing the use of data-driven technologies, including AI, within the health and care system could cause unintended harm.

Tackling these challenges so that the opportunities can be capitalised on, and the risks mitigated, requires taking action in five key areas:

1. **Leadership & Society:** creating a strong dialogue between industry, academia, and Government.
2. **Skills & Talent:** developing the right skills that will be needed for jobs of the future and that will contribute to building the best environment for AI development and deployment.
3. **Access to Data:** facilitating legal, fair, ethical and safe data sharing that is scalable and portable to stimulate AI technology innovation.
4. **Supporting Adoption:** driving public and private sector adoption of AI technologies that are good for society.
5. **International engagement:** securing partnerships that deliver access to scale for our ecosystem.

This report sets out current and future developments in each of these areas, and provides the rationale for why NHSX is creating the new £250 million NHS AI Lab in collaboration with the Accelerated Access Collaborative (AAC). Overall the goal is to help the system players from innovators to commissioners, to fully harness the benefits of AI technologies within safe and ethical boundaries, whilst speeding up the development, deployment and use of AI so that we can get benefits to more people - patients and staff alike - more quickly.

2. Where Are We Now?

Jessica Morley,
Marie-Anne
Demestihis,
Sile Hertz,
Ian Newington
& Mike Trenell

As a starting point, we needed to understand the baseline that we were working from. In order to develop useful frameworks and focus investment, we needed to understand what is:

- The current state of AI in the health and care system i.e. hype vs reality;
- The challenges faced by innovators in developing AI systems;
- The issues faced by policy makers and regulators in governing both the development and deployment of AI systems in health.

Two activities were carried out to get an up-to-date picture of AI solutions that are available on the market and where support is needed to accelerate their development in a safe, responsible way. The evidence base was covered from two angles:

- State of the Nation Survey which ran for four weeks between May and June of 2019 to build up a picture of critical issues surrounding ethics and regulation
- NIHR Innovation Observatory international horizon scan on the available evidence from academic publications, market authorisation and clinical trials databases

The results of the 2019 Survey and NIHR horizon scanning exercise reinforced the 2018 survey results published in [Accelerating Artificial Intelligence in Health and Care- Results from a State of the Nation Survey](#) - that AI in health is in the early stage of the [Gartner Hype Cycle](#). While significant progress has been made over the last year, just under half of the products available globally have market authorisation and just one third of AI developers in the UK believe that their product will be ready for deployment at scale in one year (as shown in Figure 2).

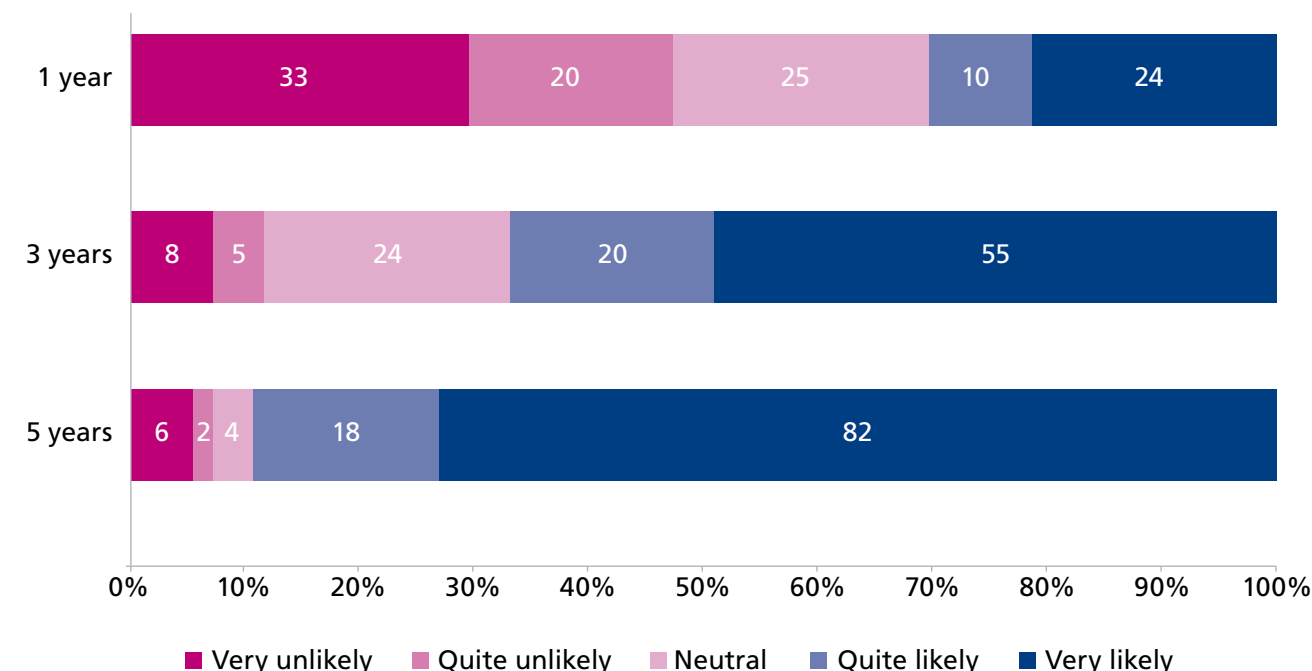
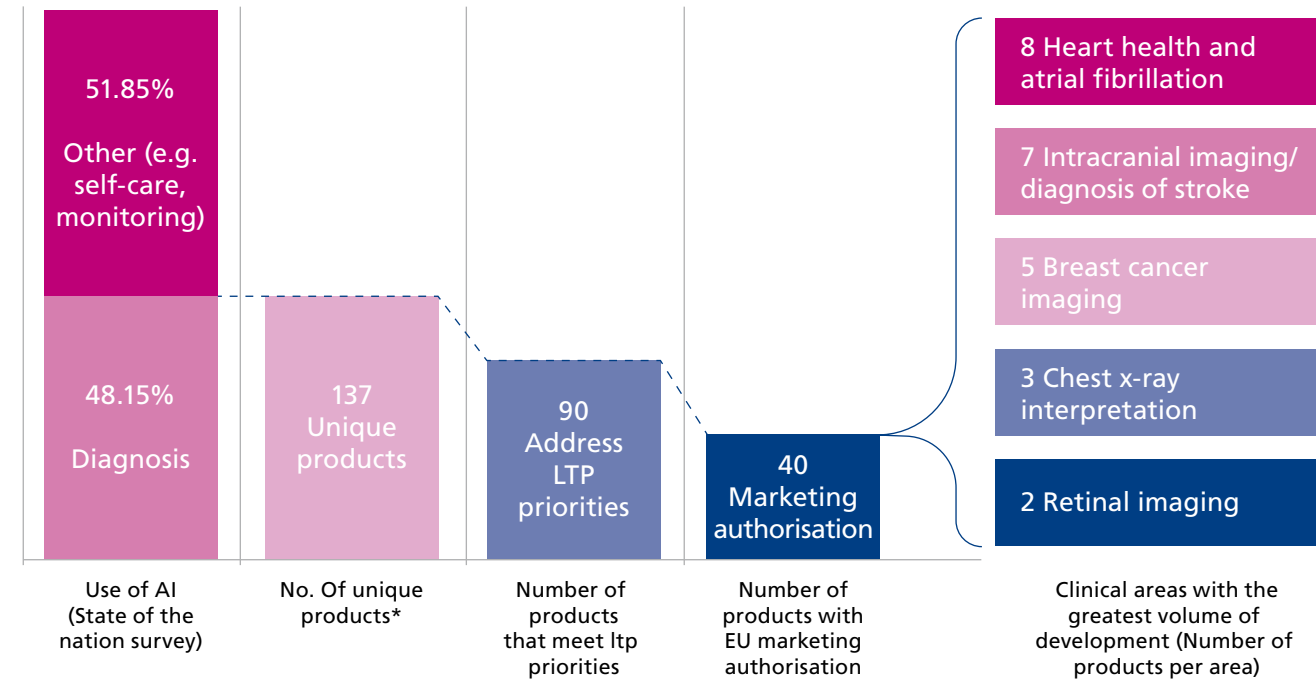


Figure 2: Proportion of products likely or very likely to be ready for at scale deployment in 1,3 or 5 years from the UK State of the Nation Report (112 products)

In addition to low market-readiness, the results also show that interest from AI developers is not yet evenly spread across all opportunity areas for AI-Health.

The results of the Horizon Scanning exercise show that diagnosis and screening are the most common uses of AI, with 132 different AI products used in diagnosis or screening covering 70 different conditions (Figure 3). Of these, 90 products addressed priorities in the Long-Term Plan and, within these,⁴⁵ had European market authorisation. Based on this analysis, interpretation of images in screening mammography, retinal imaging, X-Ray, cardiac monitoring and head CT appear to be the areas with the greatest development activity.



*Source: NIHR INNOVATION OBSERVATORY

Figure 3: Areas of greatest development in AI/data-driven technologies

AI can be used more readily in diagnostics for two main reasons:

1. Most radiology images are in a standardised digital format i.e. they provide structured input data for training purposes, compared to the unstructured and often non-digital data of health records, for example. This also means there are good data sets available for retrospective algorithm training and performance validation.
2. Image recognition machine learning techniques are more mature. The evidence so far shows that algorithms can, within constrained conditions, be used to identify the presence of malignant tumours in images of breasts^{7,8}, lungs⁹, skin¹⁰, and brain¹¹ as well as pathologies of the eye¹² to name a few.

As images are typically produced and evaluated in hospital settings by clinicians, this could explain why the survey showed 67% of all solutions are currently being developed for use by clinicians and 59% are designed to be deployed in secondary care settings.

This is higher in the case of diagnostics specifically, in which 83% of solutions are being developed for use by clinicians and 73% for use in secondary care.

The purpose of many of these diagnostic-specific solutions is to speed up the rate of diagnosis and/or to identify the patients most at risk (so they can be prioritised) as well as to help the NHS cope with staff shortages by making more effective use of the radiologists available. This is reinforced by the survey findings which show solutions are being developed to achieve quicker diagnosis (79%), faster identification of care need (63%), and better experience of health services (63%). Overall, 71% of diagnostic solutions are designed to deliver on the outcome of 'system efficiency'.

These are exciting results. However, before the NHS can capitalise on these opportunities, the ecosystem as a whole (e.g. developers, regulators, innovators, policymakers etc.) needs to consider:

- How to validate the results of individual studies – to check, for example, whether the algorithm is equally capable of recognising malignancy in mammography scans of people with different ethnicities.
- How to model the impact on individual pathways and the system as a whole. For example, we need to assess whether speeding up the rate at which people are 'diagnosed' could lead to longer anxious waits for treatment if the capacity of the system to treat is not increased as well.
- How to ensure consistently good public engagement with the concept of AI as a whole and with specific technologies

In addition, the results show that more work is needed to ensure the datasets vital to the development of life-saving AI technologies are FAIR (findable, accessible, interoperable, and reusable)¹³ and used appropriately. As whilst the results show that in almost all instances the responses to the question on 'provider of data' and 'data controller' remain the same, and there has been some

It is currently quite hit and miss whether or not developers seek ethical approval at the beginning of the development process with an almost 50/50 split between those that did and those that did not.

investment in the development of sophisticated modelling techniques. For example, 19% of solutions are being developed on algorithmically generated datasets and this is likely to increase. However, the majority (57%) are still reliant on patient data either provided by Acute Hospital Trusts (55%) or patients themselves (23%) through the use of third-party apps. Furthermore, most developers were unaware of the commercial arrangement they had in place to gain access to the data.

This shows how the current complex governance framework for AI technologies is perhaps limiting innovation and potentially risking patient safety. The survey results also reveal that it is currently quite hit and miss whether or not developers seek ethical approval at the beginning of the development process with an almost 50/50 split between those that did and those that did not. This is in part due to a lack of awareness: almost a third of respondents said they were either not developing in line with the Code of Conduct or were not sure. The main reason given for this was ‘I was unaware that it existed.’ We (NHSX) will need to ensure that in all funding applications the expectation of compliance is made clear.

Similarly, the survey indicates that half of all developers are not intending to seek CE Mark classification (ie, they are not intending their innovation to become certified as a medical device). The reason most commonly cited was that the medical device classification is not applicable. This may be the result of a general misunderstanding as it is unclear in many cases whether or not ‘algorithms’ count as medical devices. This lack of certainty may even increase with new guidance coming into force in May 2020 and May 2022. A greater degree of clarity is required regarding the regulator requirements for ‘real AI.’

The impact of this lack of clarity is obvious, with some companies developing technologies without (or at least not earlier enough) consideration of issues such as bias, discriminatory outcomes and explainability (see Figure 4).

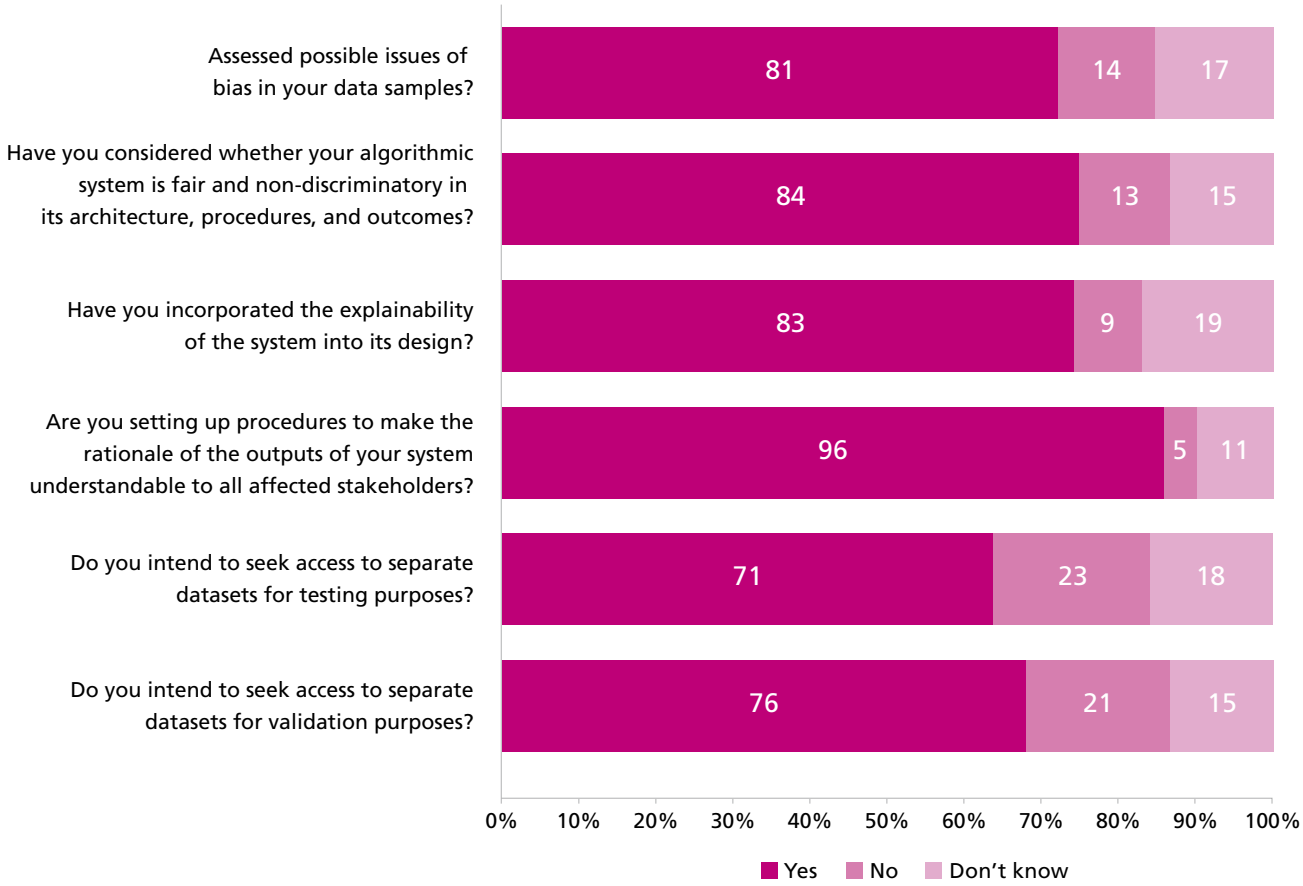
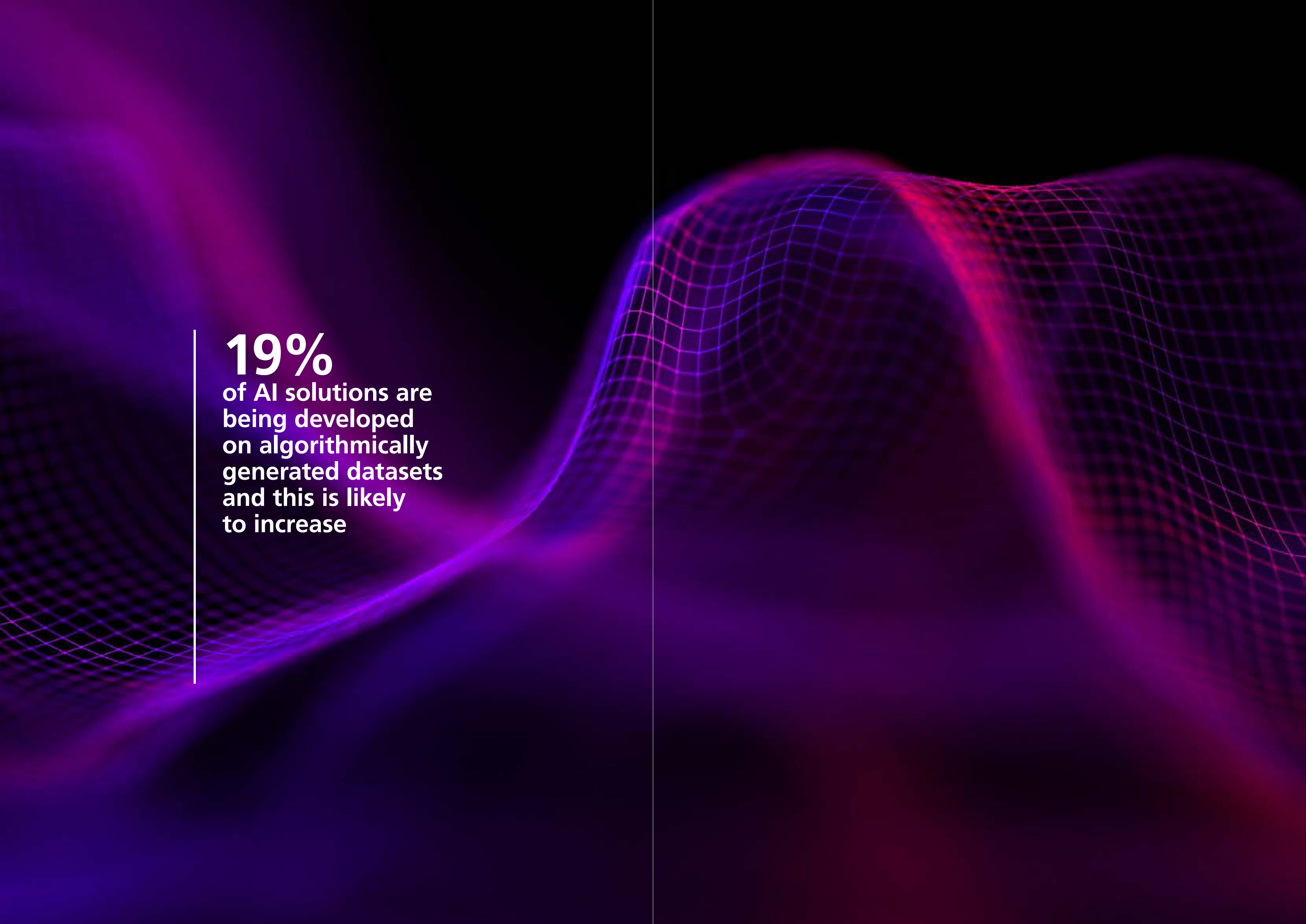


Figure 4: Consideration of ethical issues associated with algorithms during the development process

Taken together, these results provide insight into the pain points experienced by innovators which NHSX and other partners will seek to address through the NHS AI Lab by:

- Further developing the Governance (ethics and regulation) Framework
- Providing more clarity around data access and governance
- Supporting the spread of ‘good’ innovation & monitoring its impact
- Upskilling the workforce
- Developing International Best Practice Guidance



19%
of AI solutions are
being developed
on algorithmically
generated datasets
and this is likely
to increase

3. Developing the Governance Framework

Jessica Morley,
Caio C. V. Machado,
Dr. Christopher
Burr, Josh Cows,
Dr. Mariarosaria
Taddeo & Prof.
Luciano Floridi¹⁴

WHY YOU NEED ETHICS & REGULATION

In delivering on the aim of NHSX to create an ecosystem that ensures we get the use of Artificial Intelligence 'right' in health and care we need to be aware of:

a) Generic data and digital health considerations:

- i) Data Sharing & Privacy⁴⁶⁻⁵¹
- ii) Secondary uses of healthcare data^{29,52-54}
- iii) Surveillance, Nudging and Paternalism⁵⁵⁻⁵⁸
- iv) Consent¹⁵⁻¹⁸
- v) Definition of Health & Care Data¹⁹⁻²²
- vi) Ownership of Health & Care Data^{15,23-26}
- vii) Digital Divide/eHealth literacy^{27,28}
- viii) Patient Involvement^{29,30}
- ix) Patient Safety³¹
- x) Evidence of Efficacy³²⁻³⁴

b) Specific Algorithmic Considerations³⁵:

- i) Inconclusive, inscrutable or misguided evidence leading to e.g. misdiagnosis or missed diagnosis, de-personalisation of care, waste of funds or loss of trust³⁶⁻³⁸
- ii) Transformative effects and unfair outcomes leading to e.g. deskilling of healthcare practitioners, undermining of consent practices, profiling and discrimination^{22,39,40}
- iii) Loss of oversight leading to e.g. lack of clarity over liability with regards to issues of safety and effectiveness⁴¹⁻⁴⁴

To ensure that these considerations do not hinder the development or deployment of AI technologies, we need to consider the ethical, regulatory and legal framework in addition to the technical possibilities and limitations and governance mechanisms currently in place⁴⁵.

It is important to have both ethical frameworks and appropriate proportionate regulations (covered in detail below) because regulations only tell those developing, deploying or using AI what can and cannot be done whilst addressing the important safety element. This is not sufficient cover in the sensitive areas of health and care when due consideration also needs to be given to whether something should or shouldn't be done. This is why we also need soft ethics⁴⁶.

By considering the ethical implications, we can make sure that we develop frameworks that not only cover the intentions and responsibilities of different people involved in developing, deploying or using AI, but also the impacts that AI has on individuals, groups, systems, or whole populations. Ultimately, this means we can tackle any potential harms proactively rather than reactively⁴⁷.

Dr. Indra Joshi
& Jessica Morley

A CODE OF CONDUCT

The Code of Conduct for Data-Driven Health and Care Technology, initially published in September 2018 and revised in February 2019 following extensive feedback, is a core resource for anyone involved in developing, deploying and using data-driven technologies in the NHS. It provides practical 'how to' guidance on all the issues surrounding regulation and access to data.

The Code has been recognised around the world as a leading source of guidance to ensure that AI is responsibly and safely used, and addresses the need for more agile regulation- that is safe, effective and proportionate- in an environment where the pace of innovation is always going to be quicker than the ability of regulatory authorities to keep up.

The Code aims to promote the development of AI in accordance with the Nuffield Council on Bioethics' principles for data initiatives (i.e. respect for persons, respect for human rights, participation, accounting for decisions), and does this by clearly setting out the principle behaviours that the central governing organisations of the NHS expect, as follows:

- 1. Understand users, their needs and the context.
- 2. Define the outcome and how the technology will contribute to it.
- 3. Use data that is in line with appropriate guidelines for the purpose for which it is being used.
- 4. Be fair, transparent and accountable about what data is being used.
- 5. Make use of open standards.
- 6. Be transparent about the limitations of the data used.
- 7. Show what type of algorithm is being developed, or deployed, the ethical examination of how the performance will be validated, and how it will be integrated into health and care provision.
- 8. Generate evidence of effectiveness for the intended use and value for money.
- 9. Make security integral to the design.
- 10. Define the commercial strategy.

Most of these principles reflect behaviours that are already required by regulation, such as the Data Protection Act 2018, or existing NHS guidance, such as the NHS Digital Design Manual. Principles 7 and 8 and 10 are entirely new and required further supporting policy work.

Olly Buston, Dr. Matthew Fenech, Nike Strukelj, Areeq Chowdhury, Jessica Morley & Dr. Indra Joshi

Principle 7: Algorithmic Explainability

Principle 7 on ‘algorithmic explainability’ aims to tackle the ‘black box’ nature of digital healthcare applications and provide clarity to patients, users, and regulators on the functionality of an algorithm, its strengths and limitations, its methodology, and the ethical implications which arise from its use. The principle is described in detail as: ‘show what type of algorithm is being developed, or deployed, the ethical examination of how the data is used, how its performance will be validated, and how it will be integrated into health and care provision.

To help developers with this principle, NHSX has been working with Future Advocacy and other partners (including academic, industry and patient groups), to create a ‘how to’ guide for developers. The guide takes the form of a set of processes (Figure 5) that NHSX will encourage developers to undertake. The processes are divided into:

- recommendations for general processes that apply across all aspects of principle 7; and
- recommendations for specific processes that apply to certain subsections.

In both cases, the intention is to make it very clear to developers not only what is expected of them in order to develop AI for use in health and care, but also how they might go about doing it. This is because ethical and behavioural principles are necessary but not sufficient to ensure the design and practical implementation of responsible AI. The ultimate aim is to build transparency and trust.

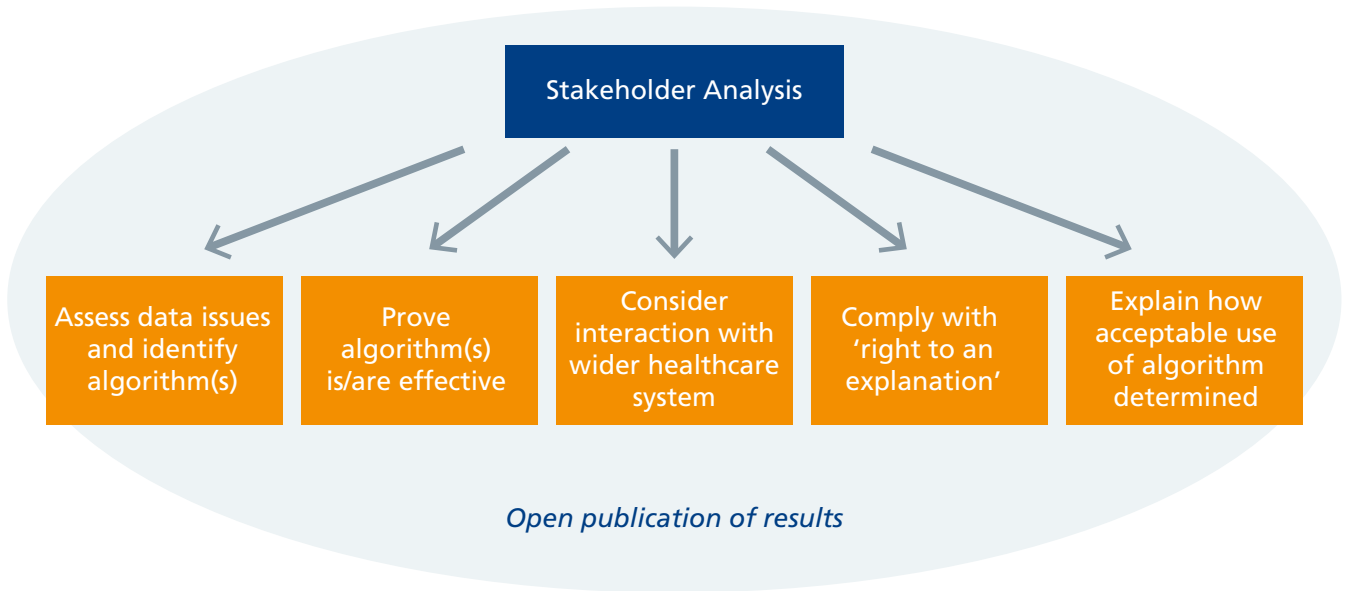


Figure 5: A schematic outlining the different components of the guidance for Principle 7

General Process: Stakeholder Analysis

Undertaking a robust and inclusive process of stakeholder analysis will help highlight and preserve relationships of importance in healthcare, ensuring that the various players in the diverse relationships making up a healthcare system are identified and involved in the development process.

This should go beyond simply identifying direct and indirect stakeholders and provide a deeper understanding of the wider cultural context (be it in the healthcare system or in wider society) in which the data-driven tool will be embedded. To ensure this output, stakeholder requirements and concerns (that is, both positively-valued and negatively-valued beliefs) should be considered through the use of value and consequence matrices (table 1). This process should be repeated at regular intervals.

a) Value matrix

	Interests/concerns relating to			
Direct and Indirect Stakeholders	Respect the dignity of individuals as persons	Care for the wellbeing of each and all	Connect with each other sensitively, openly, and inclusively	Protect the priorities of justice, social values and the public interest
Stakeholder 1				
Stakeholder 2				

b) Consequence matrix

	How does this data-driven technology affect the interests/concerns outlined in the value matrix?			
Direct and Indirect Stakeholders	Respect the dignity of individuals as persons	Care for the wellbeing of each and all	Connect with each other sensitively, openly, and inclusively	Protect the priorities of justice, social values and the public interest
Stakeholder 1				
Stakeholder 2				

Table 1: Value and consequence matrices. a) Once stakeholders are identified by developers, their concerns, wishes, values, and other views can be considered in the context of the SUM principles developed by the Alan Turing Institute ⁴⁸ b) Once these views and concerns are understood, the developer should determine how deploying their proposed data-driven technology could impact these, with a judgement applied as to whether this is a negative or positive impact. The use of colour coding (e.g. traffic light system) could then provide an at-a-glance view of the areas of greatest potential benefit and greatest concern.

Specific Processes:

1. Report on the kind of algorithm that is being developed/ deployed and how it was trained, and demonstrate that adequate care has been given to ethical considerations relating to the selection of, obtaining of, and use of data for the development of the algorithm.
 - a. Reflect on the proposed means of collecting, storing, using and sharing data, and on the proposed way that their algorithm(s) will work by using ‘Datasheets for Datasets’⁴⁹ or Open Data Institute’s ‘Data Ethics Canvas’⁵⁰
 - b. Identify what type(s) of algorithm(s) constitute the data-driven technology, and answer specific associated questions with that type of algorithm. For machine learning models, developers could adopt the ‘model card’ approach⁵¹

2. *Provide supportive evidence that the algorithm is effective:*

- a. **Submit the data-driven tools for external validation against standardised, validated datasets** (as and when these become available)
- b. **Engage with NHSX at the earliest stage of development, in order to communicate:**
 - i. The **proposed method of continuous audit**.
 - ii. The **expected inputs and outputs** against which performance will be continuously audited.
 - iii. How **these inputs and outputs were determined**.
 - iv. How **these inputs and outputs are likely to impact the different stakeholders** identified in the stakeholder analysis.
- c. **Use standard reporting frameworks**, such as those being developed by the EQUATOR Network.

3. *Demonstrate that due consideration has been given to how the algorithm will fit into the wider healthcare system, and report on potential wider resource implications of deployment of the algorithm:*

- a. **Identify:**
 - i. The **need/use case** for the data-driven technology, and the **existing care pathway(s) impacted** by the tool.
 - ii. The **associated care pathways** that interact with the target care pathway. For example, a tool designed for patients with diabetes may well have impacts on cardiovascular disease care pathways, and renal disease care pathways, as patients with diabetes are frequently seen on these pathways.
 - iii. The potential **impacts on these target and associated care pathways** of the tool.

4. *Explain the algorithm to those taking actions based on its outputs, and to those on the receiving end of such decision-making processes:*

- a. **Clarify the extent to which a decision based on an algorithmic tool is automated** and the extent that a human has been involved in the process—that is, full transparency on the use of an algorithm.
- b. Use the stakeholder analysis exercise to **clarify what is meant by the term ‘meaningful explanation’ for each stakeholder group**.
- c. **Coordinate with patient representative groups and other stakeholders to help develop ‘meaningful’ language** as part of the explanation that will be understood by patients and other stakeholders.
- d. Where explanations remain too complex for lay comprehension, developers should **support third parties that are trusted by patients (e.g. disease-specific charities) in acting as advocates for their patient groups**.

5. *Explain how the decision has been made on the acceptable use of the algorithm in the context it is being used (i.e. is there a committee, evidence or equivalent that has contributed to this decision):*

- a. **Utilise specifically-designed activities** (such as user research, talking to patient groups and representatives, citizen juries, etc) to assess thinking on the acceptable use of an algorithm. For example, nurses and clinicians should participate in the development of an algorithm that determines staff rotas.
- b. **Openly document the justification for and planning of these activities**.
- c. **Monitor user reactions** to the use of the data-driven technology, and **gauge levels of its acceptance on a recurrent basis**.

Mark Salmon,
Bernice Dillon,
Indra Joshi,
Felix Greaves
& Neelam Patel

The framework is important to encourage adoption of new technologies, including AI, as it is vital that those using them in the provision of care are confident that they work safely.

Principle 8: Evidence for Effectiveness

For principle 8, which is to ‘generate evidence of effectiveness for the intended use and value for money’, NHSX worked with the National Institute for Health and Care Excellence (NICE), Public Health England (PHE), and MedCity (the life sciences sector cluster organisation for the Greater South East of England), to create the [Evidence Standards Framework for Digital Health Technologies](#).

The framework establishes the evidence of effectiveness and economic impact required before digital health interventions can be deemed appropriate for adoption by the health and care system. In keeping with its principled proportionate approach, the framework is based on a hierarchical classification determined by the functionality (and associated risk) of the tool, which indicates the level of evidence required; for example, a more complex tool (such as one providing diagnosis) requires considerably more evidence than one simply communicating information.

The framework is important to encourage adoption of new technologies, including AI, as it is vital that those using them in the provision of care are confident that they work safely. For the NHS as a whole, it is also important that the cost and effectiveness of using a specific technology over another or a non-technological solution can be justified. In the traditional practice of evidence-based medicine this evidence is generated by randomised controlled trials. This is not always practical for digital health technologies (DHTS) or AI so NICE published the evidence standards framework for DHTs, with supporting case studies and educational materials, in March 2019.

The standards in the framework cover evidence of clinical effectiveness and economic impact and provide a common reference standard for discussions between innovators, investors and commissioners. They are designed to allow innovators to produce evidence that is better, faster, and at a lower cost and in turn; they also allow the NHS over time to commission and efficiently deploy (at scale), digital health tools that meet patient/ NHS need. Importantly, the evaluation of digital health and AI solutions can be standardised, and is a key benefit.

The framework may be used with DHTs that incorporate artificial intelligence using fixed algorithms. However, it is not designed for use with DHTs that incorporate artificial intelligence using adaptive algorithms (that is, algorithms which continually and automatically change). Separate standards (including principle 7 described in previous section) will apply to these DHTs.

What’s Next

To further build on this work, NICE is planning to set up a pilot DHT evaluation programme to establish a robust process for the national evaluation of these technologies. The aim is to enable NICE to issue positive recommendations to the NHS and care system for DHTs that offer a real benefit to patients and the NHS and social care systems.

The technologies being evaluated in the pilot will mainly be Tier 3b DHTs as defined in the NICE standards framework. These are the technologies that have measurable patient benefits, including tools used for treatment and diagnosis, as well as those influencing clinical management through active monitoring or calculation. The technologies incorporate AI to different degrees and include: a clinical decision support tool for triaging people for dementia assessment; a vital signs monitoring technology based on skin colour changes; and a technology for identifying cardiac arrhythmias.

The evaluation of the technologies will be based on the [established medical technologies guidance development process and methods](#). However, for the pilot digital technologies this will be supplemented with a technical assessment which will include examining the extent of the use of AI . The review of the clinical evidence and economic impact will align with the NICE standards framework and so a key component of the pilot will be to explore how NICE can clearly specify the data that is required to address uncertainties in the evidence as early as possible to feed this into further development of the standards.

Office of Life Sciences

Principle 10: Commercial Strategy

For Principle 10, described as ‘define a commercial strategy’, [a set of additional principles were developed by the Office for Life Sciences](#) to help the NHS realise benefits for patients and the public where the NHS shares data with researchers.

The aim of the principles is to help the NHS adapt to the ever-increasing need to share data between different parts of the healthcare system and with the research and private sectors to tackle serious healthcare problems through data-driven innovation. At the same time there is a need to put in place appropriate policies and delivery structures to ensure the NHS and patients receive a fair share of the benefits, and no more than their fair share of risk, when health data is made available for purposes beyond direct individual care.

As the technologies develop, the potential benefits and risks will shift, and so will the principles. As the frameworks iterate, it is crucial that the public feel as though they have been involved in the process. This is why NHS England and Understanding Patient Data are currently conducting research, based on public engagement and deliberation, to answer the question: what constitutes a fair partnership between the NHS and researchers, charities and industry on uses of NHS data (patient and operational)? Findings from this work will inform policy development being led by the Office for Life Sciences (OLS) - and will guide development of commercial model templates under the guidance of the National Centre for Expertise.

Jessica Morley,
Sile Herz, Marie-
Anne Demestihis,
Joseph Connor,
Hugh Hathaway &
Francesco Stefani

Self-Assurance Portal

NHSX are currently working with UCL to develop an online ‘Self-Assurance Portal’ to facilitate compliance with the Code of Conduct. The portal will help developers understand what is expected of them, prompting them to provide specific evidence for each principle. In this way NHSX hopes to not just be telling people what to do in order to develop responsible AI, but asking them to tell us how they did.

Eleonora Harwich
& Claudia Martinez

The portal is an online workbook version of the Code. Developers answer questions linked to each of the principles in the Code for each new product. For example, in relation to Principle 3, this question is asked: Is data gathered in the solution used solely for the purpose for which it is gathered? When users have provided responses to each set of maro-questions, they can see how their answers compare relative to others through visualisations.

MAPPING THE REGULATION JOURNEY

Regulation is often perceived as being a barrier to the implementation and adoption of AI in healthcare. However, a closer look at the regulatory landscape shows that there are few issues with the regulation itself. The issues lie rather in the lack of coordination between regulators and statutory bodies along the innovation pathway^[1]. In addition, the absence of a guidance and regulation navigator makes it difficult for people to figure out what they need to do and with whom they need to interact with at each stage of the process.

The journey map below (Figure 6) provides a summary of a larger scale map^[2] looking at the regulatory landscape for data-driven technologies in England, from idea generation to post-market surveillance.

Broadly speaking, there are five types of pain points in the regulatory process⁵⁹. First, in the current landscape no one body/unit is responsible for the overall process making it difficult to ensure coordination between regulators. Second, regulation can often be wrongly interpreted on the ground, particularly regarding regulation around data. Third, in some very specific instances, the regulation itself is not fit-for-purpose. The letter of the law would require people to go through such cumbersome processes that regulators follow the ‘spirit of the law’ instead. Fourth, in some cases the remit of regulators is unclear or overlapping, which means that no one is responsible for policing a specific regulatory requirement. No regulator has direct oversight over the quality of the data used to train algorithms, meaning that no one is responsible for preventing bias in algorithmic tools. Finally, there are uncertainties about how to regulate certain aspects of AI.

^[1] There are five regulators involved in the regulation of data-driven technologies in healthcare: the Care Quality Commission, the Information Commissioner’s Office, the General Medical Council, the Health Research Authority and the Medicines and Healthcare products Regulatory Agency. Another four statutory bodies: the National Data Guardians, NHS Digital, NHS England & Improvement and the National Institute for Health and Care Excellence. There is also a multitude of other bodies with a role in this field.

^[2] This larger scale map was produced thanks to a thorough literature review, 40 semi-structured interviews and three workshops with a total of 31 participants.

The data access stage causes a lot of confusion. People have difficulties in determining the legal basis for processing data and if their project should be classed as research or not.

The data access stage causes a lot of confusion. People have difficulties in determining the legal basis for processing data (i.e. direct patient care or secondary uses) and if their project should be classed as research or not. For instance, developing a piece of software using medical data should always be considered as a secondary use regardless of that software eventually being used to provide direct care to the patient. It should also be classed as research and obtain approval from the Health Research Authority (HRA)⁶⁰.

The proof-of-concept stage helps to assess the feasibility and practical implementation of data-driven technologies. At this stage, manufacturers might conduct pre-clinical studies or academic research as well as test the validity of algorithms. Manufacturers will also start generating the clinical and technical evidence required to obtain the CE marking and for getting their product commissioned by the NHS. Existing regulation and the Medicines and Healthcare Products Regulatory Agency's (MHRA) guidance on CE marking for medical devices and in-vitro diagnostic medical devices is clear and accessible⁶¹. Confusion exists, however, regarding the type and routes to obtaining evidence for products not undergoing a CE marking procedure.

The regulatory compliance stage is relatively straightforward as the MHRA's guidance is clear⁶². At this point, innovators would also have engaged with the notified bodies who carry out the conformity assessments. There are nevertheless regulatory challenges which particularly affect the regulation of AI. For example, there are no standards in place for the validation of algorithms, although, there is currently a project looking into this issue between NHS Digital and the MHRA⁶³. There is also a lack of clarity about regulating 'adaptive' algorithms and its implications for regulatory compliance.

OVERCOMING REGULATORY PAIN POINTS

Overcoming the pain points in the development pathway will be a priority for the NHS AI Lab. This will be a long-term and evolving project and guidance will need to adapt as technologies develop. There are a number of projects already underway to get the process started, and include:

- Work by the Care Quality Commission to develop principles for encouraging digital innovation as part of the 'well led' criteria of assessments.
- Work by NHS England to update the NHS Code of Confidentiality to ensure that it enables research.
- Development of HealthTech Connect by NICE:
 - Companies (health technology developers or those working on behalf of a health technology developer) can register with www.HealthTechConnect.org.uk. Data can be entered and updated about a technology as it develops. It is free of charge for companies to use.
 - The system will help companies to understand what information is needed by decision makers in the UK health and care system (such as levels of evidence), and clarify possible routes to market access.
 - The information entered will be used to identify if the technology is suitable for consideration by an organisation that offers support to health technology developers (for example, through funding, technology development, generating evidence, market access, reimbursement or adoption).
 - It will also be used to identify if the technology is suitable for evaluation by a UK health technology assessment programme.
 - Technologies that are suitable for support or evaluation will be able to access it through HealthTech Connect. This will avoid the need for companies to provide similar, separate information about the technology to the organisation or programme.



- Development of synthetic datasets by MHRA and NHS Digital (as referenced on previous page).
- Research by the Information Commissioner's Office and the Alan Turing Institute as part of Project ExplAIIn to create practical guidance to assist organisations with explaining artificial intelligence (AI) decisions to the individuals affected.
- Launch of the Care Quality Commission's regulatory sandbox for health and social care. The sandbox is running a cohort specifically for machine learning and its application to Radiology, Pathology, imaging and physiological measurement services. They will focus on developing their registration and inspection policies with industry and NHS partners. CQC are also partnering with the MHRA, British Standards Institute, NICE, and NHSx as part of this round to consider the gaps and overlaps with other regulators, and the wider issues around adopting these technologies in clinical practice.

These programmes of work have put solid foundations in place, but by creating the NHS AI Lab, and investing significantly in the development of both the regulatory frameworks themselves and the technical techniques to ensure compliance, the UK can deliver on its promise to be the best place to practice responsible AI.

4. Clarifying Data Access and Protection

*Dawn Monaghan
& Juliet Tizzard*

NAVIGATING DATA REGULATION

As highlighted in the previous section, access to health data is necessarily complex as it requires safeguards, high levels of security and data minimisation to mitigate the risk of re-identification and clear controls to ensure it is used only for appropriate purposes.

As a result, Information Governance is often cited as a problem when attempting to introduce new technologies or ways of working. This is particularly the case if the application of the rules is misunderstood or the interpretation of the rules has been turned into a complex 'black art'.

However, it's important to realise that the legal obligations of Data Protection are meant to enhance the processing of information. They can facilitate the ability to meet business needs whilst protecting individuals' privacy and confidentiality and upholding their rights.

Importantly, the principle of data minimisation is a key concept which must be embraced.

The first questions to ask are, what data is actually required and can the aims be achieved by using data other than Confidential Patient information?

- **Identifiable data** – In law it has different names - Personal Data or Confidential Patient Information - with different definitions but it includes data where an individual can be identified either directly from the dataset or in combination with other datasets.
- **Anonymous data** – Data in a form that does not identify individuals and where identification through its combination with other data is not likely to take place.
- **Synthetic data** – A synthesised, representative dataset which does not relate to any real people.

If Confidential Patient Information needs to be used then the full remit of the IG principles need to be adhered to and ethical considerations taken into account.

Applying the notion of 'Privacy by Design' at the concept stage can identify possible impacts that the proposed product or way of working may have on an individual's privacy and will help assure legal compliance and maintain trust. Completing a Data Protection Impact Assessment will help capture potential impacts, consider mitigations and enable informed risk management and proportionality judgements to be taken. It should be part of the business governance process.

In essence when considering the IG implications there is a need to be clear about:

- What data is required.
- Why is it required.
- Who will be processing it (will it be shared).
- How will it be processed (including security, storage, etc).
- Where will it be processed.

Once the above information is articulated an impact assessment can be undertaken and safeguards put in place to meet key principles and remain compliant.

Whilst lawfulness must be paramount, the requirements of not only the Data Protection Act but other relevant legislation such as the Common Law Duty of Confidentiality must be adhered to. However, other principles are just as important, and none more so than the need for transparency. It is essential that any use of data is transparent; this not only secures compliance but ensures that citizens' expectations are taken into account and managed, with public trust maintained.

When using techniques such as AI the impacts on individuals need to be fully explored, particularly if decisions may be taken without any human involvement. There are specific rules which apply to automated processing to ensure individuals understand the process and how it can affect them.

Completing a Data Protection Impact Assessment will help capture potential impacts, consider mitigations and enable informed risk management and proportionality judgements to be taken.

However, if the appropriate approvals are sought and the proper protections are in place, then it is important to facilitate access to it so that the many benefits of AI and other data-driven technologies can be unlocked. This is why we are investing in a number of important projects designed to facilitate access for the purpose of delivering better health outcomes.

Dr. Natalie Banner

UNDERSTANDING PATIENT DATA

Many people are not aware that the information within their health records has enormous research potential, nor how it could actually be used in practice or what their choices are. Unsurprisingly, there are significant public concerns especially when commercial organisations may be involved in data use.

Understanding Patient Data (UPD) exists to try and help make the uses of patient data more visible understandable and trustworthy, for patients, the public and health professionals alike. We work with charities, patient groups, academic researchers, healthcare providers, media, data custodians and policy makers to champion responsible uses of health data. Data is a complex and technical landscape, so we produce freely available resources that seek to demystify health data use. This includes guidance on jargon-free language, and animations that tell stories about patient data in an engaging way. We also bring together networks of advocates and provide a unified voice to raise concerns or challenges to policymakers about the rules, transparency and accountability of health data use.

With the advent of data-driven technologies, it will be more important than ever to find creative ways both to inform people about health data use, and to involve those who are interested or concerned in governance and decision-making at local and national levels, to ensure the systems for managing and making use of the data are worthy of the trust people are being asked to place in them.

UPD was set up as an independent initiative in 2016 partially in response to the National Data Guardian's call for a better public conversation about health data. It is primarily funded by Wellcome, a global charitable foundation dedicated to improving health.

The National Data Guardian works to ensure citizens' confidential information is kept safe and confidential, and that it is shared when appropriate to achieve better outcomes for patients.

PROTECTING THE CITIZEN

The National Data Guardian (NDG) was created in November 2014 to be an independent champion for patients and the public. It works to ensure citizens' confidential information is kept safe and confidential, and that it is shared when appropriate to achieve better outcomes for patients. The NDG does so by offering advice, guidance and encouragement to the health and care system.

The Health and Social Care (National Data Guardian) Act 2018 granted the NDG the power to issue official guidance about the use of health and adult social care data in England. This means that public bodies such as hospitals and GPs and private companies or charities that are delivering services for the NHS or publicly funded adult social care will have to take note of relevant guidance from the NDG.

The NDG wants to build trust in the use of data across the health and social care system and is guided by three main principles. First, to encourage clinicians and other members of care teams to share information that directly affects the care of the person they are treating or supporting. This can bring direct benefits to people such as joined-up care and better diagnosis and treatment. Second, to inform and provide citizens' voice in how their health and care data is used. Third, to build a dialogue with the public, commercial companies, researchers and service providers about how information should be used.

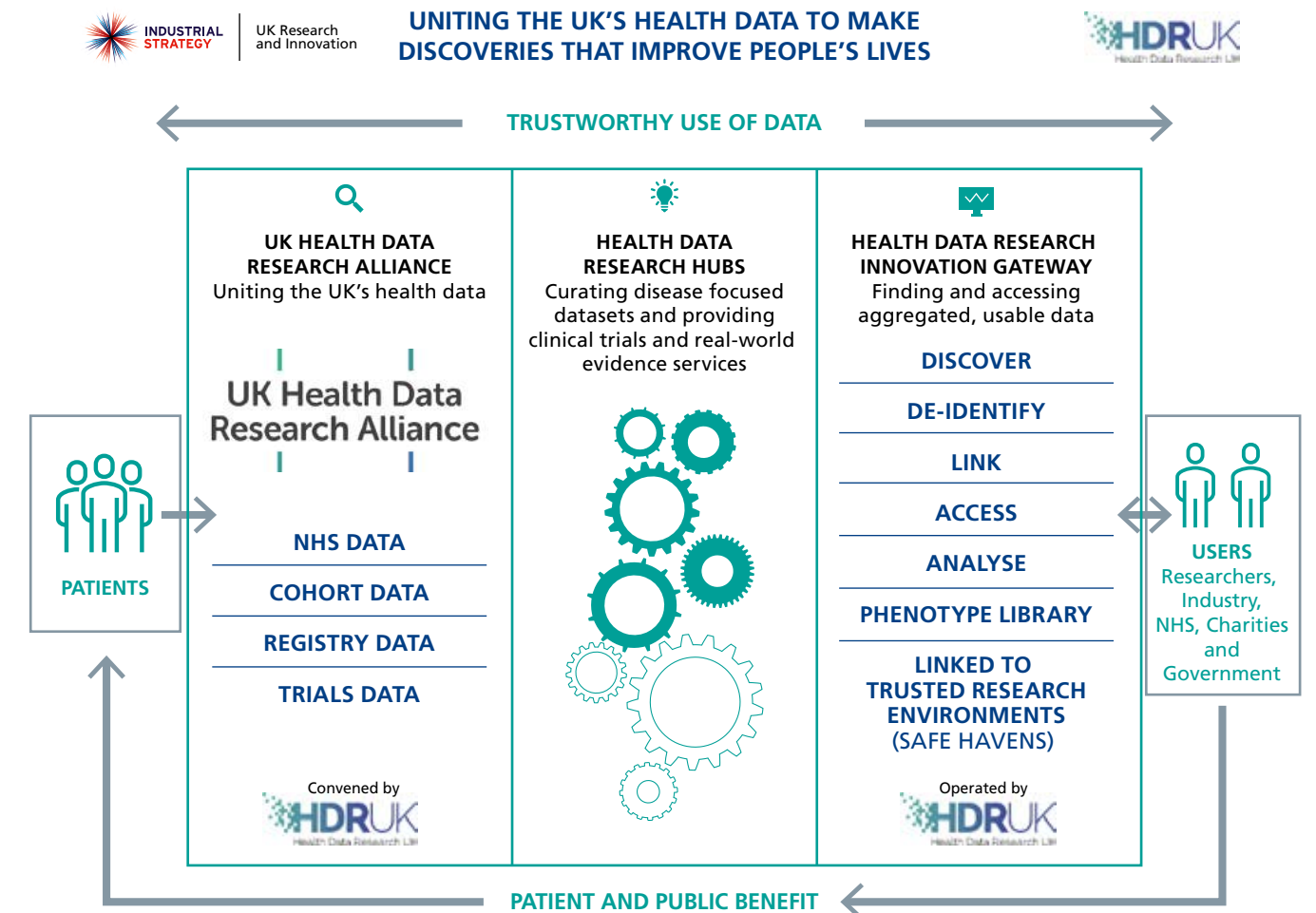
The NDG Panel is a group of experts appointed by the NDG to advise and support its work to represent the interests of patients and the public. The UK Caldicott Guardian Council is a sub-group of the NDG Panel and is responsible for protecting the confidentiality of people's health and care data and ensuring that it is used properly. All NHS organisations and local authorities that provide social services must have a Caldicott Guardian.

DATA INNOVATION HUBS

The Digital Innovation Hub Programme managed by Health Data Research UK (Health Data Research UK) aims to become a UK-wide life sciences ecosystem providing responsible and safe access to health data, technology and science, and research and innovation services to ask and answer important health and care questions.

This four-year programme is funded by the UK Research and Innovation's Industrial Strategy Challenge Fund (ISCF). The programme integrates with and reinforces Health Data Research UK's investment in health data science and talent development, to stimulate the use of health data to make lives longer and healthier. It encompasses three essential functions:


1. **UK Health Data Research Alliance** – an Alliance of data custodians committed to making an unprecedented breadth and depth of data available for research and innovation purposes for public benefit. [Find out more](#)
2. **Health Data Research Hubs** – making data available, curating data, and providing expert research services. The Hubs will be centres of expertise to get from raw and fragmented data to insight and the location to collaborate and co-create. [Find out more](#)
3. **Health Data Research Innovation Gateway** – providing discovery, accessibility, security and interoperability to surface data, support linkage, and enable health data science safely and efficiently. [Find out more](#)



The NHS AI Lab working in partnership with Health Data Research UK through the **UK Health Data Research Alliance** will help create the pipeline from development to deployment of AI systems within health and care.

DATA COLLABORATION AT SCALE

The UK Health Data Research Alliance is an independent, not-for-profit alliance of health data custodians from the UK's leading health and research organisations [members] united to establish best practice for the ethical use of UK health data for research at scale.



75%
of Global Digital Health
Partnership members
say that the body (or
bodies) responsible
for regulating digital
health technologies
is currently looking
to change their remit
and adapt regulations
appropriately

For researchers and innovators to benefit from secure access to an array of health-related data and address health problems, we need the right expertise, trusted governance, and collaboration through a UK-wide research alliance.

The UK has some of the richest health and research datasets and assets world-wide. Some of these are well organised, but only a fraction of all NHS and research data is accessible. By developing and co-ordinating the adoption of tools, techniques, conventions, technologies, and designs, the Alliance enables the use of health data in a trustworthy and ethical way for research and innovation.

For researchers and innovators to benefit from secure access to an array of health-related data and address health problems, we need the right expertise, trusted governance, and collaboration through a UK-wide research alliance. The UK Health Data Research Alliance is coordinated by Health Data Research UK and was formally launched on 7 February 2019. Other custodians of large-scale health data are warmly welcomed to join the Alliance to widen the opportunities for medical breakthroughs.

DATA AGREEMENTS AND COMMERCIAL MODELS

Data sharing agreements and 'future fit' commercial model templates will address a clear pain point for innovators. Questions on data access and sharing are complex and will have different answers in different contexts. A key challenge is answering when, by whom, why, where and how data (especially sensitive data) can be accessed¹⁶. Commercial models will engender trust, reduce negotiation time and complexity, and help innovators meet Principle 10 of the Code of Conduct.

This is why in addition to guidance provided by the Code of Conduct, The National Data Guardian, The Information Commissioner's Office, and the guiding principles in the framework for data sharing with researchers published in July 2019, NHSX is committed to developing a National Centre of Expertise to oversee the policy framework, provide specialist commercial and legal advice to NHS organisations entering data agreements, develop standard contracts and guidance, and ensure that the advantages of scale in the NHS can deliver benefits for patients and the NHS. The National Centre of Expertise will have a crucial role to play in ensuring that data sharing

agreements meet the first guiding principle that "any use of NHS data, including operational data, not available in the public domain must have an explicit aim to improve the health, welfare and/or care of patients in the NHS, or the operation of the NHS."³¹

The Centre of Expertise will sit in NHSX and its core functions will include:

- Providing hands-on commercial and legal expertise to NHS organisations – for potential agreements involving one or many NHS organisations (eg for cross-trust data agreements or those involving national datasets). This could include providing support to negotiate and execute agreements, and assessing and building capability within NHS organisations where useful. The Centre will develop and provide tailored legal advice on relevant issues (eg intellectual property, state aid).
- Providing tools and products including good practise guidance and examples, standard contracts, and methods for assessing the value of different partnership models to the NHS.
- Signposting NHS organisations to relevant expert sources of guidance and support on matters of ethics and public engagement, both within the NHS and beyond.
- Engagement and understanding the landscape – building relationships and credibility with the research and industry community, regulators, and with NHS organisations and patient organisations, including developing insight into demand for different datasets and identifying and communicating opportunities for agreements that support data-driven research and innovation.
- Developing benchmarks and scenarios to provide NHS organisations with reference points on what 'good' looks like in agreements involving their data, taking into account demand for data, market conditions and the international context, and setting clear and robust standards on transparency and reporting to underpin and support public trust.



NHSX DATA FRAMEWORK

NHSX is committed to creating an environment that will facilitate access to data, by reducing barriers and delays to data access and providing clarity to both data owners and companies on how to help the process run smoothly. A [Data Framework](#) was published in July 2019 that sets out five guiding principles to maximise benefits for patients and the public where the NHS shares data, and which will underpin successful innovation in the AI Lab.

Ongoing work by the NHSX team in the mandating of interoperability standards, by NHS England in the development of programmes such as the Local Integrated Health and Care Record Exemplars, and Health Data Research UK in providing a single point of access and facilitating the development of Digital Innovation Hubs (as well as others), will be enabled through this framework.

5. Encouraging the Spread of 'Good' Innovation & Monitoring the Impact

WHAT DOES 'GOOD AI' LOOK LIKE?

Professor Philip Beales

1. Precision Medicine

Precision medicine encompasses predictive, preventive, personalised and participatory medicine (also termed P4 medicine)¹⁷. It is moving from the traditional one-size-fits-all form of medicine to more preventative, personalised, data-driven disease management model that achieves improved patient outcomes and more cost-efficiencies. Precision medicine, as defined by the National Institute of Health (NIH), is an emerging approach for disease treatment and prevention that considers individual variability in genes, environment, and lifestyle¹⁸. The promise is that precision medicine will more accurately predict which treatment and prevention strategies will work best for a particular patient.

In the broader setting, AI is helping industry to accelerate drug development, cut costs and gain faster approvals while reducing errors. Data-driven health will also likely impact patients directly by providing access to their own personal data, improving compliance with treatments and real time monitoring for adverse events making participation in P4 medicine a reality.

*Anna Tomlinson,
Chris Wigley,
MJ Caulfield &
John Hatwell*

2. Genomics

Key to unlocking the benefits of precision medicine with AI is the use of genomic data generated by genome sequencing. Machine learning is already being used to automate genome quality control. AI has improved the ability to process genomes rapidly and to high standards and can also now help improve genome interpretation.

Genomics England, which currently contains over 100,000 genomes and over 2.5 billion clinical data points, has developed a platform to capture the substantial amounts of data from automated genome sequencing and healthcare professionals that AI needs.

Researchers are using machine learning methods to identify genetic mutation signatures associated with certain cancers, for example, helping to identify and classify sub-types of the disease and to identify targets for new treatments.

Whilst AI will undoubtedly accelerate progress, we should be careful to supervise machine learning and people are still needed to avoid misinterpretation of data in making care decisions. In order to build an optimal and sustainable future for genomic medicine, earning, building and retaining trust with patients, the public and technology and research partners, will be essential.

*Simon Harris &
Deirdre Dinneen*

3. Image Recognition

The East Midlands Radiology Consortium (EMRAD) is a pioneering digital radiology system, and is a partnership of seven NHS trusts[i] spread over 11 hospitals, covering more than five million patients. The cloud-based image-sharing system has set the national benchmark for a new model of clinical collaboration within radiology services in the NHS.

In 2018, EMRAD formed a partnership with two UK-based Artificial Intelligence (AI) companies, Faculty and Kheiron Medical Technologies, to help develop, test and - ultimately - deploy AI tools in the breast cancer screening programme in the East Midlands. It aims to improve and optimise clinical service capacity, to enhance patient care at significant scale and to increase NHS confidence in the utilisation of innovative machine learning tools. The project aims to develop and test both clinical and non-clinical (operational) AI tools.

Kheiron's Mia(TM) tool has the potential to support the clinical workforce issues in the service by acting as the second reader in the dual-read mammography workflow, while Faculty's 'Platform' software has the potential to help optimise operational processes such as clinic scheduling and staff resourcing.

James Teo &
Richard Dobson

4. Operational Efficiency

Kings College Hospital NHS Foundation Trust together with the NIHR BRC at the South London and Maudsley Hospital have developed an open-source real-time data warehousing tool called 'CogStack'. CogStack meets the acute need for a more efficient way to clinical code to improve financial and operational efficiencies for providers throughout NHS. The team at King's College Hospital NHS Foundation Trust and the South London and Maudsley Hospital tested CogStack for clinical coding in a fracture outpatient clinic setting to identify under-coding and was able to triple the depth of coding within a month (from ~10% of cases to 30% of cases having procedures recorded accurately). This translates to £1,260,000 of financial activity per annum even without the efficiency gains.

Using modern open-source natural language techniques exploited by companies such as Google and openAI, the team have developed further advanced prototype NLP algorithms for performing large-scale tagging of clinical text (Semantic EHR; MedCAT) with machine learning; this has the potential to code all clinical data in real-time with associated efficiency gains. This is already functioning and deployed in a large NHS Trust, and has been tested with a real-world NHS problem showing it has potential for substantial disruption of manual healthcare processes. CogStack has been adopted by the open source community with major contributions to the evolving codebase coming from partners including UCLH BRC and Health Data Research UK.

Asheem Singh &
Charlotte Holloway

TACKLING BARRIERS TO ADOPTION

Innovation is not just a set of technologies but an environment and a culture. The RSA conducted a research study to understand the 'human story' behind the challenge to spread technology - particularly new and complex technology - through the NHS. Over the course of June and July 2019, the RSA conducted 12 in-depth interviews (according to Chatham House rules) with professionals developing, procuring and using data-driven technologies across the country to understand what prevents people from adopting new technologies in the NHS to gain insight into:

- 1. Commissioner, clinician and patient interactions with radical technologies.
- 2. Barriers to adoption and pain-points.
- 3. Mechanisms that might help facilitate these interactions and overcome the barriers.

The results show that even in this relatively small set of conversations, there was striking convergence on what needed to be done to smooth the adoption of AI into the health system and create a genuine, human-centred culture around technological innovation in the NHS.

The macro-level factors identified as being essential to clinical adoption of AI are: patient acceptance, evidence and clinical champions. The specific recommendations are:

- Put in place a rigorous anti-bias test.
- Prove the benefits of AI by, for example deploying back-end operational solutions first.
- Model the impact on clinical workflow.
- Make provisions for the continuous upskilling of the workforce.
- Scale up our sandboxing and piloting initiatives.

Dr. Sarah Deeny,
Dr. Hannah Knight,
Dr. Geraldine Clarke,
Josh Keith & Dr.
Adam Steventon

MEASURING IMPACT

The NHS faces a huge challenge to select the right technologies for adoption, given that so many new technologies are being developed, all with uncertain impact⁵². These decisions should be informed by robust modelling of the potential impact of new technologies on patients and the health care system before they are implemented. Producing this information, or providing data that would allow the NHS and those commissioning technology to produce it, will help developers demonstrate the requirements of principle 2 of the code of conduct: "Define the outcome and how the technology will contribute to it"⁶.

It's vital that new technology addresses the most critical problems facing the healthcare system and its patients, but the technology's potential impact will depend on where in the clinical pathway it is implemented. For example, an algorithm designed to detect diseases from medical images could be used within community settings (to detect patients needing referral to specialist care) or within outpatient care (to assist specialists with diagnosis). These will have different impacts. The modelling process can borrow from simulations and assessment frameworks (for example, those created to evaluate the potential impact of vaccination or screening programmes^{53,54}), and be developed further to consider the potential impact of new AI tools. Such models should be developed and shared widely in the NHS, using open source methods where possible, and using well-established parameters for assessment where practical (such as those developed by NICE for cost effectiveness)⁵⁵.

REAL-WORLD EVALUATION

The NHS needs to track the impacts of new technology in real-world settings so that benefits can be identified and spread, and potential harms spotted quickly.

While working out if the technology works in a controlled setting is one thing, effective real-world piloting and evaluation is required to understand the wider conditions necessary to ensure technologies can deliver the benefits they promise, and how this impact might vary in different parts of the NHS.

The NHS needs to track the impacts of new technology in real-world settings so that benefits can be identified and spread, and potential harms spotted quickly. Technology companies need this information too, so they can improve their products and services over time. Several dimensions are relevant here, including impacts on patient outcomes and the demand for care, the workforce and wider system.

Monitoring the implementation of AI is challenging⁵⁷. These are complex interventions with multiple components and aims. Impact will be shaped by the context in which they are implemented, and the technologies themselves are rapidly evolving. An agile and multidimensional approach is needed, which combines both quantitative and qualitative methods (Figure 7)^{58,56}.

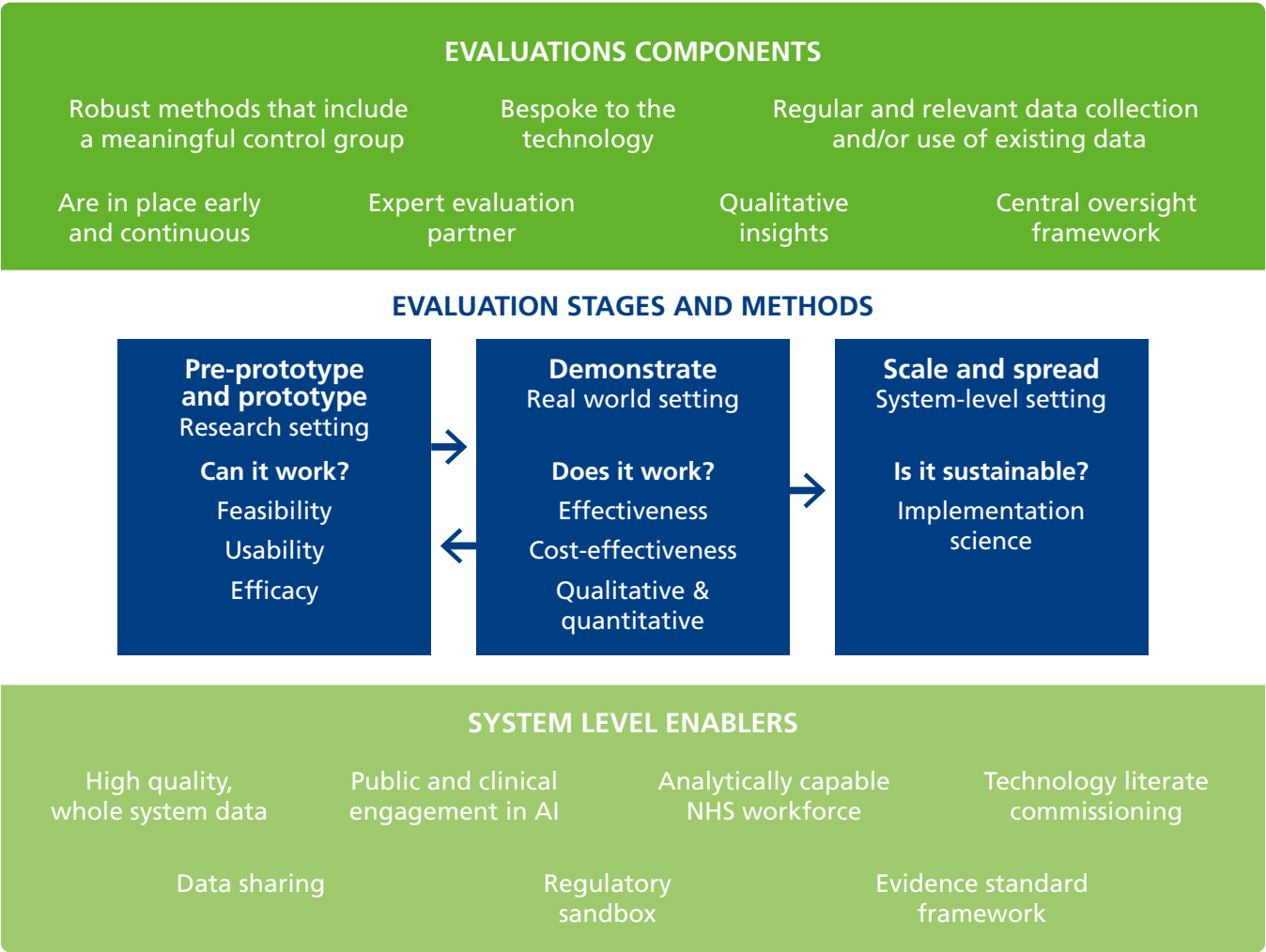


Figure 7: A framework for monitoring AI and assessing which products should be scaled for use within the NHS

The impacts of new technologies will need to be assessed against a counterfactual, which charts the outcomes of existing forms of care for similar patients. This can often be achieved using existing NHS data, as demonstrated by the Health Foundation's Improvement Analytics Unit⁵⁹. In some cases, NHS data will need to be combined with data from technology companies, which will require a data sharing agreement to be in place from the start⁶⁰.

Given the range of new technologies being developed, the NHS may need to invest in better analytical capability so that impacts can be assessed at local level (as outlined in the next chapter, on building capability and skills).

6. Creating the Workforce of the Future

Professor Chris Holmes, Professor Sir Adrian Smith, Professor Andrew Morris, Dr Adam Steventon & Ellen Coughlan

The [Topol Review](#) published in February 2019 made recommendations that will enable NHS staff to make the most of innovative technologies such as genomics, digital medicine, artificial intelligence and robotics to improve services. The aim of the recommendations is to ensure the workforce is able to fully benefit from the introduction of AI into the NHS, enabling them to make the best use of their expertise, informing their decisions and saving them time.

As a practical example of these recommendations, the Alan Turing Institute, Health Data Research UK and the Health Foundation are exploring ways to work together to foster the development of analytical skills to enable the NHS to more readily take advantage of exponential developments in technology. Advances in algorithms, including AI tools, coupled with secure, scalable, computing environments that can access curated linked data will offer substantial opportunities for the NHS to improve patient outcomes and increase operational efficiencies. Certain core skill sets will be essential to meet this potential, enabling the design, build, and deployment of advanced analytics, all the while minimising financial, patient, and ethical risk.

A primary focus will be growing critical thinking expertise at all levels of NHS seniority to evaluate the potential for AI to benefit any given situation. This will include a technical understanding of requirements for the evidence base (the level and quality of curated data needed) and the computing resource needed to support both the design-deploy stage as well as the measurement and tracking of benefit. NHS staff need to be supported to understand the ethical considerations and regulatory procedures that are needed to sign off and monitor AI solutions.



Investing in the capabilities of NHS analysts is needed to enable them to develop the appropriate skills and understanding of the data in order to apply AI solutions effectively, and to interpret and communicate the complex information produced. The core analytical skills needed include:

1. an appreciation of how AI works, what its strengths and weaknesses are, and how it complements traditional statistical approaches;
2. the ability to design reproducible studies, using open-source analytics, built on an evidence base that is sufficient to learn from and to measure success on and;
3. an understanding of the protocols and meta-data needed to ensure robust execution of ethically fair solutions.

The NHS will also benefit from the creation of nimble tech teams, working hand-in-hand with data wranglers and external partners, focussed on rapid prototyping and evaluation of potential ideas.

The NHS can benefit from existing UK governmental and charitable investment in this space. Organisations like the Alan Turing Institute, Health Data Research UK (HDRUK) and the Health Foundation (HF) have significant experience in the provision of advanced analytic tools, the linking of health data at scale, and the appraisal and improvement of analytical capability in the UK's health system^[3]. These organisations have active training and development programmes that can support professionals from different backgrounds, and at all career stages, to use analysis and AI to improve patient outcomes and the efficiency of care. They are also working in partnership with NHS organisations to gather well curated data to improve the quality, safety and efficiency of care, underpin research, and generate new insights that can continuously improve decision-making by clinicians and patients. Moreover, Turing's embedded software engineering team, skilled in the prototyping and deployment of AI solutions, can help co-design, and share experience and technical insight into scalable algorithms and systems for learning. Ultimately, example programmes such as this underpins our ambition to ensure the people who keep the NHS running feel as supported as possible when delivering care, both now and in the future so that the NHS is the best place to work. This aim is a core part of the [NHS Interim People Plan](#) which reflects the beginning of a new way of working and supporting the NHS workforce to be the best it can be.

^[3] Innovative exemplars include the Health Foundation's Advancing Applied Analytics awards show that it's possible to deliver rapid improvements in the way in which data is brought to bear on real-world problems in the health and care system.

7. Developing International Best Practice Guidance

*Clara Lubbers,
Meredith Makeham,
Rodney Ecclestone
& Dr. Indra Joshi*

The NHS AI Lab will be focused on making the UK a global leader in data and AI in the health and care space, and will be closely involved with key international initiatives as described below.

GLOBAL DIGITAL HEALTH PARTNERSHIP

The Global Digital Health Partnership (GDHP) is a collaboration of governments and territories, government agencies and the World Health Organization, formed to support the effective implementation of digital health services. There are five key workstreams: 1) interoperability, 2) cyber security, 3) policy environments, 4) clinical and consumer engagement and 5) evidence and evaluation.

The GDHP aims to establish a baseline of activity and progress across GDHP participant countries and to examine practical approaches to the ethical, regulatory and legislative frameworks needed to enable data-driven technologies (for example, machine learning and artificial intelligence) to be successfully deployed in the delivery of health and care.

Machine learning and artificial intelligence have the potential to radically transform the delivery of health and care. However, the immense potential of these approaches requires digital health infrastructure that supports interoperability within and between countries to ensure its success. In this context, the GDHP seeks to assist policy makers to ensure that they are sufficiently aware of both the potential benefits and the potential issues to ensure they create an environment that maximises the chances of successfully achieving the benefits of emerging technologies. International collaboration in this area is providing governments with evidence to accelerate their approaches to local implementations and policy development relating to digital health technologies.

To help develop this international community, we carried out a survey and asked members a series of questions about their approaches to regulating AI, with an additional request for an example AI technology being developed in their country.

The aggregate results revealed that the development of AI on an international scale is out-pacing the development of the supporting policy framework. 62% of the AI solutions given as examples by members have already been deployed, yet less than half of all members (44%) have a national or regional policy framework for the development and/or deployment of AI. Over 81% of members state that the national or regional body responsible for regulating digital health is currently not regulating adaptive algorithms in a clinical setting (none are regulating adaptive algorithms used in back-office settings), and 46% do not know whether the developers of the example provided sought ethical approval before development began. However, awareness of the need to develop this framework is growing. 75% of members confirm that the body (or bodies) responsible for regulating digital health technologies is currently looking to change their remit and adapt regulations appropriately.

*Sameer Pujari,
Clayton Hamilton,
Naomi Lee*

WORLD HEALTH ORGANIZATION (WHO) & INTERNATIONAL TELECOMMUNICATION UNION (ITU)

The World Health Organization (WHO) in partnership with the International Telecommunication Union (ITU) has established a Focus Group on Artificial Intelligence for Health (FG-AI4H)⁶¹ with the aim of identifying opportunities for international standardisation and fostering the application of AI to health issues on a global scale. The work of the focus group encompasses development of a standardised assessment framework with open benchmarks for the evaluation of AI-based methods for health, such as AI-based diagnosis, triage or treatment decisions. The initiative represents a collective effort by the international community to systematically address the adoption of AI in health through a standardised and transparent evaluation of the underlying methods used.



44%
of all GDHP members
have a national
or regional policy
framework for the
development and/or
deployment of AI



Digital technologies, and AI in particular, are seen as having a key role in delivering on national policy through mechanisms to extend the scope, transparency and accessibility of health services and health information.

WHO's actions for developing digital health are anchored in achievement of the UN Sustainable Development goals and in particular SDG target 3.8, the achievement of universal health coverage (UHC), which includes financial risk protection, access to quality essential health care services, and access to safe, effective, quality, and affordable essential medicines and vaccines for all. Digital technologies, and AI in particular, are seen as having a key role in delivering on national policy through mechanisms to extend the scope, transparency and accessibility of health services and health information, hence reaching marginalised and underserved populations, and at the same time creating efficiency gains in the operation of health systems and improving the quality of care and treatment outcomes. To further guide the development of safe and inclusive digital health services in countries, WHO has been mandated to develop a Global Digital Health Strategy, which sets out a framework of action to advance and apply digital health towards achieving the vision of health for all.

Effective AI adoption requires us to identify potential health problems to which AI interventions can be assessed and applied. Rather than attempting to specify the AI for health algorithms themselves, or standardise across a multitude of medical data formats, the AI4Health focus group is working to establish a benchmarking framework, whereby eligibility of an AI model can be assessed, and feedback provided to improve quality. This is being done by identifying common health domains (e.g. general diagnosis, specialty diagnosis, health natural language processing, general clinical encounter note data extraction and coding, Rx coding, lab coding, etc.) and for each domain to examine sourcing of test data, select current gold standard test success rates (e.g. how does a professional score on this test data), set benchmark rates for an AI system (to be acceptable for decision support, to be acceptable for autonomous operation), and define acceptable fail modes (e.g. alert human operator if below a given confidence threshold). As the work of the focus group progresses over time, several clinical and public health domains have emerged as priority areas for examination. These include cardiovascular disease risk prediction, diagnosis of bacterial infection and anti-microbial resistance (AMR), dermatology, falls among the elderly, histopathology, malaria detection and neuro-cognitive diseases.

Technical groups for Ethics and Regulation of AI have also been established under the Focus Group to examine existing approaches and provide standardised guidance for countries to follow.

Artificial Intelligence for Health also offers new ways to address the global shortage of healthcare professionals, which is a key issue decision makers face and one which threatens the sustainability of national health systems and health service delivery.

When implemented safely, AI has the potential to significantly improve and support medical diagnostics and treatment decision processes and strengthen the global public health role. It is hoped that the work of the WHO-ITU Focus Group in this area will accelerate adoption towards a safe and effective future of AI in Health.

*Professor Alastair
Denniston &
Dr. Xiaoxuan Liu*

THE EQUATOR NETWORK

The EQUATOR (Enhancing the QUALity and Transparency Of health Research) Network is an international initiative that seeks to improve the reliability and value of published health research literature by promoting transparent and accurate reporting and wider use of robust reporting guidelines.

To ensure transparency and reproducibility of any health intervention, whether it be a diagnostic test, a predictive model, an early warning system or a decision-making tool, complete reporting of study methods and results are of vital importance. The same principles apply to machine learning algorithms. Without complete and transparent reporting, it is difficult to assess the validity and generalisability of the study findings, which can result in misconceptions of overstated efficacy and utility of any health intervention. The risk is that an AI-intervention, which might not be effective or feasible in the real world, could be commissioned and implemented.



A recent systematic review⁶² of over 20,000 studies evaluating the diagnostic accuracy of deep learning diagnostic algorithms across all medical domains highlighted a number of inadequacies in study reporting, including lack of clear details about the study population or datasets, the proportion and handling of missing data and full contingency tables (which allow the derivation of key performance metrics such as true positives, true negatives, false positives and false negatives). Only 82 studies made direct comparisons between deep learning algorithms and human healthcare professional accuracy, of which only 14 studies compare algorithms and humans using the same test validation dataset in an external validation.

New reporting standards are under development to address these issues. The CONSORT-AI and SPIRIT-AI initiative^{63,64} was announced in September 2019, to develop reporting guidance for studies evaluating AI interventions in clinical trials. Whilst such guidance is primarily for ensuring transparency of reporting, they can also be helpful for assisting clinical trial design.

8. Conclusion

The NHS AI Lab aims to bring together policies, partners and programmes to develop and deploy safe, effective artificial intelligence applications. The accelerating speed and complexity of applications means, however, that we need an agile, proportionate governance framework to harness the tremendous benefits of data-driven technologies while minimising the potential risks.

The Code of Conduct described in the report is a landmark step to this end, involving a number of partners working with NHSX to ensure the public is protected while benefiting from the exponential pace of developments we are seeing – addressing both the softer ethical considerations of the “should vs should not” in the development of AI solutions as well as the more legislative regulations of “could vs could not”.

Collaboration and partnership with organisations in a mutually - supporting ecosystem will become more and more important ahead, as the true potential of AI can only be realised through access to large volumes of FAIR data (findable, accessible, interoperable, and reusable) within an open environment to optimise research, testing and deployment at scale.

An iterative, learning approach is also needed to develop the right skills and behaviours in the workforce to critically evaluate and develop AI applications that are inclusive, do not cause unintended harm and do not leave ‘anyone behind’.

The future is exciting – with opportunities that we can’t even imagine yet. People will have better access to their own data which will help them to engage more in their own health. Personalised medicine and predictive prevention will accelerate through greater access to clinical, genomic, phenotypic, behavioural and environmental datasets – where AI can spot patterns and opportunities that humans can’t. This will spur the continual development of novel breakthrough technologies and help the UK maintain and develop its position as a global leader in ethical AI.

People will have better access to their own data which will help them to engage more in their own health.



Appendix: Case Studies

FLAGSHIP CASE STUDIES

Professor Philip
Beales

Precision Medicine

Precision medicine encompasses predictive, preventive, personalised and participatory medicine (also termed P4 medicine)¹⁷. It is moving from the traditional one-size-fits-all form of medicine to more preventative, personalised, data-driven disease management model that achieves improved patient outcomes and more cost-efficiencies. Precision medicine, as defined by the National Institute of Health (NIH), is an emerging approach for disease treatment and prevention that considers individual variability in genes, environment, and lifestyle¹⁸. The promise is that precision medicine will more accurately predict which treatment and prevention strategies will work best for a particular patient.

There are, however, significant challenges to the widespread adoption of precision medicine that include a deluge of medical data, a paucity of trained specialists, and the enormous costs of drug development¹⁹. Take, for example, the 30% rise in the number of CT scans ordered in the UK between 2013 and 2016²⁰ during which period the number of radiologists only increased by 3% annually²¹. Many studies demonstrate that as radiologists are compelled to work faster their interpretation error rate rises²².

However, AI can address this by leveraging deep learning approaches to overcome the obstacles inherent in large data sets and unstructured data. In clinical settings, AI can assist clinicians to work more efficiently and make more accurate diagnoses improving the productivity of healthcare workers.

In the broader setting, AI is helping industry to accelerate drug development, cut costs and gain faster approvals while reducing errors. According to a recent research report²³, achieving the full potential of precision medicine will be impossible without applying AI and machine learning. Specifically, leveraging advanced machine learning and deep learning technology can outperform clinicians and researchers in rapidly analysing large datasets and integrating exponentially growing amounts of data from a wide variety of

novel-omics sources (e.g. genomics, transcriptomics, proteomics, metabolomics, microbiomics) into clinically actionable insights for personalised care plans and more effective population health management.

One challenge is how best to integrate precision medicine data into electronic health records (EHR). OSF HealthCare, an integrated healthcare network, has integrated CancerIQ's genetic cancer risk assessment program with Epic's EHR platform as part of a system-wide population health initiative to reduce cancer disparities and deaths²⁴. In the US, individuals can now leverage Medfusion's national health data network to consolidate their health records and help drive discovery through private, secure data sharing using LunaPBC part of LunaDNA, the first community-owned genomic and health data platform – that enables members to access their EHRs²⁵.

AI and machine learning are already transforming precision medicine delivery by driving computational phenotyping tools (www.mendelian.co) and new methods for target drug discovery (e.g. BenevolentAI). In fact there are now 100s of start-up companies using AI in drug discovery,²⁶ demonstrating the power of this emerging sector for driving innovation in healthcare.

In clinical trials, fewer than a third of all phase II compounds make it to phase III, and one-third of phase III trials fail because the trial lacks enough patients or the right kinds of patients. AI can potentially boost the success rate of clinical trials by identifying and characterising patient subpopulations best suited for specific drugs and by efficiently measuring biomarkers that reflect the effectiveness of the drug being tested. Nevertheless, we are still at the proof-of-concept stage but feasibility pilot studies are demonstrating the high potential of numerous AI techniques for improving the performance of clinical trials.

Data-driven health will also likely impact patients directly by providing access to their own personal data, improving compliance with treatments and real time monitoring for adverse events making participation in P4 medicine a reality.

*Anna Tomlinson,
Chris Wigley,
MJ Caulfield &
John Hatwell*

Genomics England

Key to unlocking the outlined benefits of precision medicine with AI, is the use of genomic data generated by genome sequencing. Genomics England has been using these immense datasets in the 100,000 Genomes Project to provide diagnoses and stratification to rare disease and cancer patients. However, they realise that this is still only scratching the surface of what genomics will be able to offer – it is fully expected that the use of AI based tools will dramatically accelerate progress in this field. As Genomics England expands its vision to sequence 5 million genomes over the coming years, these tools will become ever more important.

Machine learning is already being used to automate genome quality control. Using machine learning we can identify genomes that are outliers, in terms of their sequencing or sample characteristics. Researchers can also use these techniques to identify the genomic sex of participants and compare them with their clinical notes. AI has improved the ability to process genomes rapidly and to high standards and can also now help improve genome interpretation.

Machine-learning-based AI requires substantial amounts of data for training. To address this, Genomics England has developed a platform to capture information and actions from automated genome interpretation systems and healthcare professionals. This data is carefully stored, to make it readily available for learning. It now hold observations on over 6 million variants across more than 35,000 cases. Going forwards, Genomics England anticipate developing models which can classify variants according to their likelihood of being pathogenic. However, this is likely to require more data, given the complexity of variant-disease interactions.

The research community has also started using AI based tools within the Genomics England research environment (which currently contains over 100,000 genomes and over 2.5 billion clinical data points). For example, researchers are using machine learning methods to identify genetic mutation signatures associated with certain cancers, helping to identify and classify

sub-types of the disease and to identify targets for new treatments. This has the potential to improve cancer diagnosis by making better predictions of disease progression, and to better stratify patients for more effective and efficient treatment. This could be key to delivering on the promises of the long-term plan to diagnose three-quarters of all cancers at stage 1 or 2 by 2028²⁷. Similar work is ongoing with rare disease, to identify driver mutations of particular conditions. Developing models will be able to better predict the severity of outcomes based on genomic and clinical data.

Whilst AI can undoubtedly increase the scale and speed of data collection and analysis, and accelerate progress, we should be careful to supervise machine learning. People still have a unique ability to critically evaluate and moderate learnings, which is vital to maintaining integrity and direction, and to avoid misinterpretation of data in making care decisions.

Genomics England, recognises the very privileged position it holds to have access to such enormous amounts of data, and to have the opportunity to work with new techniques such as AI to take genomic medicine to the next level. With this privilege comes huge responsibility. As outlined in the introduction patient data and related research demand the highest levels of privacy and protection, and ensuring this is a priority for the whole AI ecosystem. Similarly, safety is crucial and it is vital that the system is able to guarantee the integrity in the diagnoses and treatment plans which are delivered to patients, whether these are facilitated by AI, or more traditional methods. In order to build an optimal and sustainable future for genomic medicine, earning, building and retaining trust with patients, the public and technology and research partners, will be essential.

Despite tremendous progress in recent years, there is significant opportunity for AI and machine learning tools to make further advances in genomics, presenting an unparalleled opportunity to transform the application of genomic medicine in healthcare across the world.

Simon Harris &
Dierdre Dineen

^[4] Chesterfield Royal Hospital NHS Foundation Trust, Kettering General Hospital NHS Foundation Trust, Northampton General Hospital NHS Trust, Nottingham University Hospitals NHS Trust (host organisation), Sherwood Forest Hospitals NHS Foundation Trust, United Lincolnshire Hospitals NHS Trust, and University Hospitals of Derby and Burton NHS Foundation Trust.

EMRAD

The East Midlands Radiology Consortium (EMRAD) is a partnership of seven NHS trusts^[4] spread over 11 hospitals, covering more than five million patients. EMRAD launched in 2013 with the objective to create a new, common digital radiology system. Its work was supported with the award of ‘Vanguard’ status by NHS England’s New Care Models programme which ran from 2016 to 2018.

This pioneering work saw the East Midlands become the first health community in the UK where NHS hospitals could quickly and easily share diagnostic images such as x-rays and scans. The cloud-based image-sharing system has set the national benchmark for a new model of clinical collaboration within radiology services in the NHS.

In 2018, EMRAD formed a partnership with two UK-based Artificial Intelligence (AI) companies, Faculty and Kheiron Medical Technologies, to help develop, test and - ultimately - deploy AI tools in the breast cancer screening programme in the East Midlands. The project is one of seven ‘wave two’ NHS Test Beds, and is administered by NHS England and the Office for Life Sciences. The Test Bed project is focused on Capacity, Care, and Confidence: it aims to improve and optimise clinical service capacity, to enhance patient care at significant scale and to increase NHS confidence in the utilisation of innovative machine learning tools. The project aims to develop and test both clinical and non-clinical (operational) AI tools. Kheiron’s ‘Mia(™)’ tool has the potential to support the clinical workforce issues in the service by acting as the second reader in the dual-read mammography workflow, while Faculty’s ‘Platform’ software has the potential to help optimise operational processes such as clinic scheduling and staff resourcing.

Clinical applications of AI

At the June EMRAD AI Project Board, the Lincolnshire BSP lead described the pressures as “tremendous”, with “no end in sight”. Deep learning is increasingly being proposed as a solution to the breast cancer screening workforce crisis. As part of their role in the Test Bed, Kheiron are conducting a large-scale retrospective

study on mammograms from two of the NHS sites within the EMRAD Consortium. The aim is to test the generalisability of Mia™, their novel deep learning mammography software. They hope to conclude that their generalisable model is suitable for consideration as an independent reader in double-read screening programmes (see Figure 1). This would have significant implications for the future of the breast screening workforce throughout the UK.

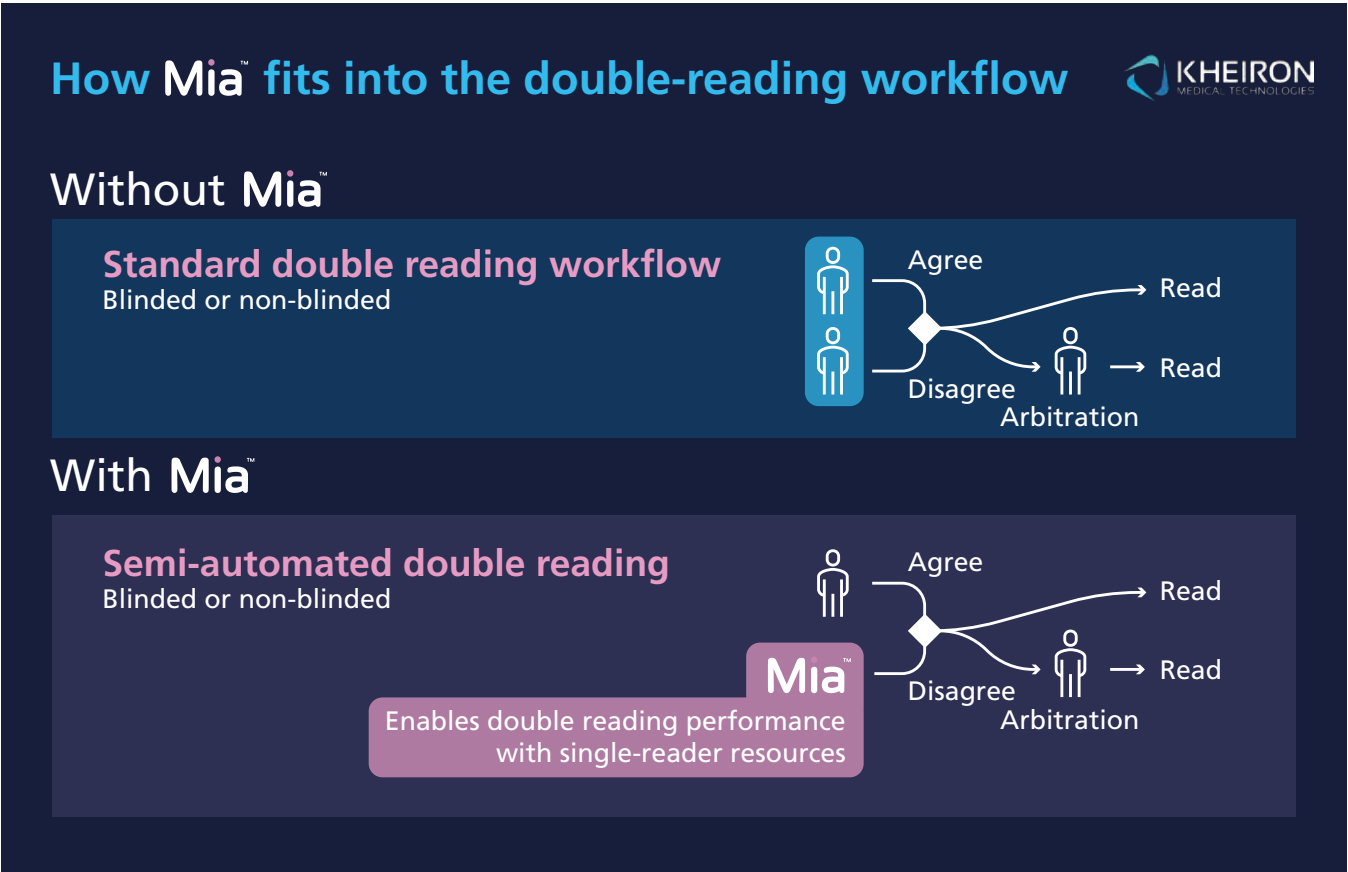


Figure 8: How Mia™ fits into the double-reading breast screening workflow

Non-clinical (operational) applications of AI

As well as using AI to address a clinical task, the Test Bed project is seeking to apply AI tools and techniques to the operational and administrative aspects of the breast screening programme, considering how AI can help to run the service in the most efficient and effective way possible.

Faculty’s role in the Test Bed is to test the potential application in the breast screening programme of process optimisation tools and techniques developed on their ‘Platform’ software and pioneered in complex operational environments such as airlines, railways and high-street retailers. The aim is to make the best possible use of scarce resources like radiologists’ time and expensive machinery, and to reduce stress on the clinical and administrative workforce delivering the programme (see Figure 9).

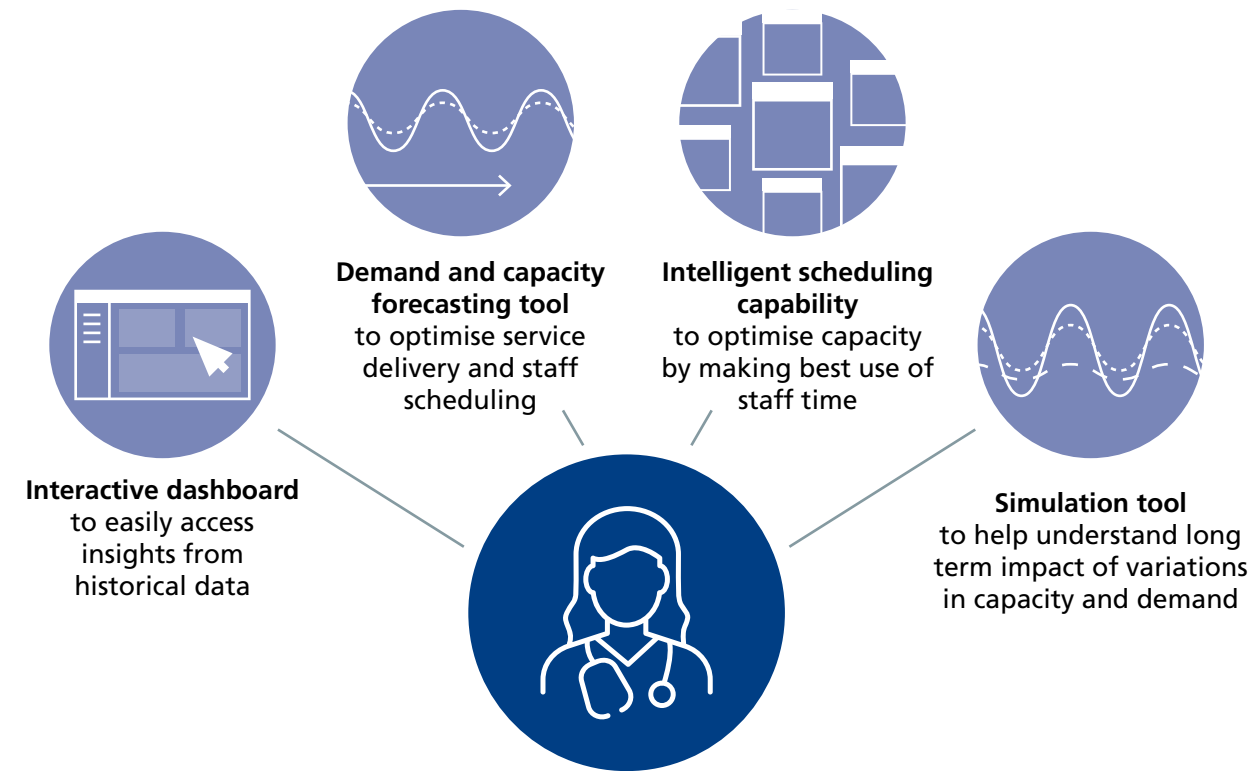


Figure 9: The AI-powered planning and scheduling tool

The Test Bed is working within the current service definition: it aims to deliver exactly the same breast screening programme in a more efficient way. For example, we will be training and testing an AI-powered intelligent capacity and demand planner. This should help screening round managers to more easily and accurately identify likely pinch-points where demand exceeds the service’s capacity, to identify ways to mitigate these pressures, and to simulate and estimate the likely knock-on impact of – for example – unplanned machinery or site down-time, or workforce changes, on the maintenance of the round length. We will also be testing a dashboard which aims to help programme managers identify their most and least efficiently used clinics, helping to best target constrained staff resources.

Discovery work for the Test Bed also identified opportunities to use AI tools to accurately predict which screening clients would and would not attend their screening appointment, potentially allowing for more accurate over-booking of clinics without detriment to the service levels individual screening clients receive. Similar techniques could be used to identify the screening appointment slot which an individual client is most likely to accept - boosting first-time acceptance rates, reducing DNAs, and reducing the administrative workload which is currently created by clients phoning to re-book their appointment for a more convenient time. Techniques developed in the marketing sector could even be used to test and optimise the communications sent to screening clients to maximise uptake and attendance rates.

Beyond the core screening clinics, AI tools may also be able to provide an early warning of increases in demand for assessment clinics, helping service managers to proactively schedule extra clinics before a spike in demand. Or these tools could predict the range of diagnostic tests which an attendee at a symptomatic clinic is likely to need, helping to schedule slots within a clinic to reduce clients’ waiting times.

If the structure and specification of the breast screening programme were to evolve in the future, this could open up further opportunities to make use of AI tools. For example, service capacity across a region is currently split into silos based on the allocation of GP practices to specific screening sites. If, in future, a 'next test due date' approach were adopted, this could potentially allow clients to be called for screening at a location that optimised their travel time and the available service capacity.

Realising all these opportunities will depend significantly on the availability and accessibility of the data which is required to train AI tools. Finding ways to securely and safely share appropriately de-identified client-level data for the purposes of service improvement will be vital. Building functional, modern connectivity into the core breast screening programme software will also be essential: programme managers already have to contend with a variety of IT systems, so core service delivery systems will need to be able to connect seamlessly with new AI tools if those tools are to be deployed effectively in the real world.

James Teo

Cogstack

King's College Hospital NHS Foundation Trust, together with the South London and Maudsley Hospital, have developed an open-source real-time data warehousing tool called 'Cogstack' that improves operational efficiency significantly.

Cogstack meets the acute need for a more efficient way to clinically code to improve financial and operational efficiencies for providers throughout NHS.

Reading health text for clinical coding underpins operational, financial and forward planning activity throughout the NHS, but there is a significant source of data quality issues in the NHS (Capita, 2014). This is because Clinical Coding as a domain is under significant pressure due to the expanding volume of data in the past decade, resulting in incomplete data capture, and manpower shortages of skilled clinical coders. Big Data techniques may be a solution – search technology, web analytical techniques like natural language processing (NLP) and semantic artificial intelligence (AI) tools that can read and interpret electronic free text at scale.

These advances in NLP are accelerating, and Microsoft, Google and Amazon have now all expressed publicly that they are exploring the NLP and search technology use-cases in healthcare. This has the potential to accelerate the efficiency of clinical coders in terms of depth of coding as well as release staff for more complex tasks.

The team at King's College Hospital NHS Foundation Trust and the South London and Maudsley Hospital tested Cogstack for clinical coding in a fracture outpatient clinic setting to identify under-coding and was able to triple the depth of coding within a month (from ~10% of cases to 30% of cases having procedures recorded accurately). This translates to £1,260,000 of financial activity per annum even without the efficiency gains. Using similar methods to Google, Apple Siri and Amazon Alexa, the team have developed further advanced prototype NLP algorithms for performing large-scale tagging of clinical text (Semantic EHR; MedCAT) with machine learning; this has the potential to code all clinical data in real-time with associated efficiency gains.

Cogstack is already functioning and deployed in a large NHS Trust, and has been tested with a real-world NHS problem showing it has potential for substantial disruption of manual healthcare processes. Further accelerated development is being proposed to expand the use of such NHS-grown open-source technologies in the NHS. Near-term benefits would be to improve efficiency of business intelligence and operational tasks initially, and allow release of human resource for more complex tasks. Subsequent medium-term and long-term benefits would emerge as the AI systems develop an understanding of medical language, revolutionising how clinical staff interact with computers.

*Tina Woods &
Melissa Ream*

Lessons from Estonia and Finland

The AHSN Network AI programme took a group of senior stakeholders on a Study Tour to Estonia and Finland in April 2019 to take learnings back to the UK on creating a citizen-led open health & care data ecosystem.

Why these two countries?

Estonia is considered the most advanced digital society in the world and has built an efficient, secure and transparent ecosystem that saves time and money (estimated to be 2% of GDP per year). Estonia's e-citizen programme provides innovative e-solutions via a secure citizen ID and has demonstrated that an open and transparent attitude provides a good foundation for trust between the citizen and the state, and gives more control to the real owner of the data, the citizen.

Finland is harnessing the power of data to benefit its citizens, having recently passed legislation on secondary use of data. It endorses ethical use of data, through its work on IHAN (International Human Account Network)- comparable to the IBAN standard used in banking but serving the 'human driven data economy' instead- and development of European guidelines for ethical use of data. SITRA, the Finnish innovation agency, is working with Estonia on two data exchange layer pilots to test the Estonia's X-Road solution to provide social and healthcare services

The key learnings from this Study Tour to take back into the UK scenarios are as follows:

- Engaging citizens and building their trust in government has been central to the success of digital innovation in Finland and Estonia.
- The development of a data infrastructure to support effective data exchange, democratise access to data and leverage integrated datasets has been fundamental to realising the potential of data-driven technologies.
- An open approach to collaboration has been a key driver for change.

NHS-R Community

<https://nhsrcommunity.com/>

R is an open source tool for data-science and statistical analysis, employed by many different organisations including BBC News, Heathrow Ltd, GSK, The FT and BT for their data science needs.

Now, the NHS-R Community, established in 2017, from a standing start with support from The Health Foundation's Advancing Applied Analytics Award, is dedicated to promoting the learning, application and use of R in the NHS. By providing problem-oriented workshops across the country and a dedicated website with hundreds of subscribers the NHS-R Community promotes a positive attitude towards peer reviewing and learning. Helping to share efficiencies for members who have cut time spent on common tasks from weeks to hours and stimulate a wider conversation about what analysts can contribute to health care. Currently the NHS-R Community is working on producing an open library of validated, off-the-shelf solutions for others to adapt and use in healthcare.

NHSX Mental Health

<https://nhsx.github.io/Mental-Health/>

NHSX have partnered with NHS England to work on meaningful challenges within the mental health space. The first is to look into children and young people's mental health, specifically around preparation for future appointments, waiting times and non-attendance rates do not attend/was not brought rates.

The team includes subject matter experts alongside NHSX colleagues who bring user-centered design practices and are dedicated to not only working on a significant problem but also raising capability of colleagues through exposing new ways of working.

The team is researching with users, working in the open and proactive in sharing what they have found. They will discover pain points in the children and young people's mental health services, outline future opportunities and constraints and recommend what to do next.

Optimam

<https://medphys.royalsurrey.nhs.uk/omidb/>

Intelligent diagnostic support has advanced significantly in the areas of image recognition. Not only is image recognition a relatively mature area of artificial intelligence, but the structured

digital datasets that are now common in radiology has facilitated the rapid training of algorithms to detect a number of pathologies. Despite this, there is a need for large AI research databases to ensure that models trained are robust and do not overfit their training data, or reproduce bias present in small datasets.

For this reason, over the last 10 years the Royal County Surrey Hospital has developed the knowledge, processes and bespoke tools to create large-scale medical image research databases, primarily aimed for use in AI and in the area of Breast Screening, Nuclear Medicine, MRI, Tomosynthesis and radiogenomics. The flagship database, OPTIMAM, focuses on 2D Full-Field Digital Mammography (FFDM) screening images, focussing on over-capacity breast screening programmes nationally and internationally. With 2.5 million de-identified images with associated clinical data and cancer location information remote PACS software). This cloud-based database is currently shared with over 50 groups and companies working on AI. With complex sharing tools to manage the sharing of huge volumes of data, including comprehensive tracking and audit logs it is easy to manage data governance and ensure that training and test sets, do not overlap.

SURVEY CASE STUDIES

Advancing Applied Analytics

<https://www.health.org.uk/funding-and-partnerships/programmes/advancing-applied-analytics>

Established in 2017, the Advancing Applied Analytics awards have awarded approximately £2.5m to 33 projects across the UK that seek to apply innovative approaches to building capability in the health and care system over 15 months. An additional £1.5m has been committed by the Health Foundation to fund up to 24 more projects over the coming years.

The programme funds projects that seek to address the capability deficiencies identified in the Foundation's Understanding analytical capability in health and care report. The projects thus far have taken heterogeneous approaches, ranging from

apprenticeships and training programmes to develop models and tools for clinicians and managers to better plan and deliver care.

An evaluation of the awards found that each project saw improvements in analytical capability that would not have been seen without the seed funding including the use of novel tools (deep learning, systems thinking), development of specific analytics techniques, learning new software skills (R, systems modelling), creation of new development opportunities, formation of new networks and collaborations that enabled analytics

These awards have demonstrated that small seed funding is an effective way to promote innovation and catalyse analytical capability in the NHS, ultimately improving health care and wellbeing across the UK.

axial3D

<https://www.axial3d.com/>

Each person is unique, and each operation will be too, more-so in the cases of complex or novel procedures. A complex surgical procedure involving variant anatomy may be the first of its type seen by a surgeon, even in a specialist centre, and providing surgeons with the ability to anticipate and manage for the operation in advance can have a great impact on outcomes.

axial3D is helping to transform surgery by providing previously unavailable insights to clinicians for preoperative planning across the world. axial3D produces highly accurate 3D printed models of specific parts of the patient's own anatomy. Based around patient scans, axial3D uses machine learning-based segmentation algorithms to automatically produce patient-specific 3D anatomical models which can be created and delivered in 24-48 hours. These anatomical models can be used by surgeons to plan for complex procedures, gain patient consent and minimise inter-clinician variability.

This results in better clinical outcomes for patients, enhanced insights for surgeons, and reduced time and costs for the surgery. Patients are also much better informed about their condition ahead of the proposed surgery when a 3D printed model is used in gaining consent.

BrainPatch

<http://www.brainpatch.ai/>

Transcranial electrotherapy uses small pulsed currents of electricity, delivered in a non-invasive manner, to influence cortical brain activity. It is hoped that through modulating cortical activity and promoting certain patterns, this technique can be used to influence learning, mood and other brain functions.

BrainPatch is a neurotechnology start-up working on a solution based on low current non-invasive brain stimulation, already shown to have positive effects in mental health, depression and diseases such as epilepsy and Alzheimer's. By using AI to optimise the stimulation protocols to each individual it may be possible to achieve better efficacy with no additional risk for a variety of medical and non-medical everyday applications. Brain Patch intends to use smart-contracts, allowing developers and consumers to prioritise their own goals, making use of their hard-and soft-ware to fulfill them.

Chief AI

<https://chief.ai/>

Whilst AI poses to transform healthcare, the digital receptiveness of existing healthcare bodies stands to challenge its rapid and effective adoption. Current structures of access and procurement are not designed for digital processes, further slowing down the speed of adaptation and innovation in this area.

Chief AI is commissioned to create an intelligent marketplace where third party AI algorithms can be provisioned on demand, leading to efficiency and convergence in the uptake of AI within the healthcare ecosystem. Moreover, Chief AI's own products provide novel diagnostics and drug discovery services powered by artificial intelligence and machine learning. Trained on world class medical datasets from around the world, Chief AI creates efficiency in the global management of non-communicable diseases such as cancer by developing and deploying medical diagnostic models, informing and customising oncology treatments, and identifying novel biomarkers for precision medicine.

Concentric Health

<https://concentric.health/>

Ensuring that patients understand the reasons and consequences for decisions made about their health forms the backbone, of an auditable and accountable healthcare system, as well as being a necessary pre-requisite to empowering people to make better decisions about their health.

Welsh health-tech startup, Concentric Health's, vision is to ensure all decisions are informed by patient outcomes and shared by patient and clinician. Their fully auditable digital consent platform, uses data to personalise outcomes, design to simplify complexity and ensures that patients are informed and engaged throughout their journey.

eTrauma

<https://openmedical.co.uk/etrauma>

In busy, fast paced, clinical environments it can be difficult to keep on top of the administrative task of ensuring every patient is met with the highest standard of care. Around the whiteboards, paper-notes and ward lists care has to be delivered in a standard, accountable way, ensuring that no one is missed and the service works efficiently, ensuring that the max number of people can be seen. No-where is this more true than in the emergency and acute services.

eTrauma is a bespoke cloud-based patient management system that has been used to manage over 500,000 patients in 15 acute trusts. The platform is hosted within the N3 / HSCN network and is used as an end-to-end clinical management system for Trauma and Orthopaedic departments. Our use of cloud-based, learning natural language processing algorithms, and structured granular data sets has helped to significantly improve data quality in patients with acute traumatic orthopaedic injuries. Leveraging this detailed data, we have partnered with NHS trusts to develop new models of care such as a revolutionary digital fracture liaison service, which automatically detects patients who may benefit from secondary osteoporosis prevention at the time of initial referral, thus improving both efficiency and capture of patients that may benefit from preventative treatment strategies.

In addition, by facilitating service redesign and improved efficiency, our systems have helped NHS trusts reduce referral to assessment times for orthopaedic fracture patients from 2 weeks to 2 days, injury to theatre times by up to 7 days, and saving over 1,000 clinic appointments in a single NHS trust alone.

First Derm

<https://www.firstderm.com/>

Skin diseases are amongst the most common diseases encountered by health professionals, with 13 million skin related GP consultations per year. Despite this, they are relatively overlooked in medical education, leaving many GPs unequipped to deal with the quantity and diversity of presentations they may see. These ailments have the potential to be life-threatening, or even psychologically harmful, given their highly visible nature, and the goal of providing high quality care to all has become harder in the context of an ageing population with dwindling resources and staffing shortages.

First Derm aims to Transition from a human powered dermatology service to a faster and more accurate artificial intelligence (AI) algorithm driven service empowering patients to access high quality care through their phones. Over the last 5 years First Derm has collected a unique dataset of over 350k+ "amateur" smartphone dermatology images and we are adding another 15 000 images per month. The AI API in dermatology (<https://www.firstderm.com/ai-dermatology/>) can screen 43 common skin diseases, from STDs to skin cancer. The algorithms are now 50% accurate on one disease and 80% correct including top 3 skin diseases (differential answers). The AI is a good fit for an anonymous quick screening of a skin concern that will let you know what next steps to take and as a decision support tool for virtual care clinics. The service is in 7 different languages and has users from over 160 and is currently being integrated into online STD testing websites, messaging apps such as Microsoft Teams and WeChat, and online health services like abi.ai

Forms4Health

<https://forms4health.com/>

Many healthcare providers operate with paper notes, records, and forms of data collection. This format is not only expensive, environmentally damaging and incompatible with rapid iterative design at a systems level, but also requires time and work to manage, store and analyse.

Forms4Health is an intuitive, easy to integrate, electronic smart forms platform, facilitating paperless working across health and social care. We believe clinical software needs to reflect the variety and dynamic nature of the healthcare sector. The design of forms4health is responsive, versatile and interoperable to the complex digital needs of the Health and Social Care profession, allowing an organisation to quickly and easily design their own smart forms. These may be used to create forms for any purpose, supporting workflows across all departments can be used by staff to produce complex resources with very little training including a range of sophisticated features, from branching logic (conditional questions), calculations, to clinical decision support.

In addition to removing the need for paper forms, reducing costs, and saving time, the data gathered through Forms4Health can be easily analysed to provide valuable business intelligence, from evaluating patient wait times, to assessing resource allocation. Benefits realised for our customers include improving care outcomes through accurate up to date information recorded at the point of care, supporting patient engagement, enhancing information sharing across internal and external departments and care settings, connecting existing systems to deliver care, improving data quality through standardisation, ease of completion and consistent inputs and saving clinical and patient time driving efficiency.

Google Health

A turning point has been reached in AI for healthcare: not only is more and better research emerging all the time, we are closer than ever before to applying this research to real-world clinical practice and to proving out benefits for patients and clinicians. That's why we've formed Google Health: to bring cutting-edge AI research into clinical practice. Google Health brings together many teams from across the organisation, including the team that was formerly DeepMind Health.

One example is our work with Moorfields Eye Hospital. The DeepMind Health team began working with Moorfields in 2016 to apply machine learning approaches to diagnosing common eye disease from routine scans. Initial research from the partnership showed that the AI system could triage urgent OCT scans as accurately as leading retina specialists. Crucially, it didn't miss a single urgent case. This research has now been transferred over to Google Health, as of September 2019, and we are now working with Moorfields to validate this research on a wider population and ultimately to deploy the system into clinical practice.

Streams, our mobile medical assistant app for clinicians, is another interesting example. It was developed by the DeepMind Health team in collaboration with clinicians at the Royal Free London NHS Foundation Trust to help identify patients at risk of acute kidney injury (AKI). The app uses the existing national AKI algorithm to flag patient deterioration and supports the review of clinical information at the bedside. An independent evaluation showed that the app speeds up detection and treatment and prevents missed cases. Thanks to the app, clinicians were able to respond to urgent AKI cases in 14 minutes or less - a process which, using existing systems, might otherwise have taken many hours. The Trust, as data controller, remains in control over personal data at all times and the team can only process it in line with their instructions. Although it does not yet use AI, Streams demonstrates the potential to surface predictive insights at the moment when clinicians need them--with life-saving consequences. The app, and the team working on it, also transitioned over to Google Health in September 2019.

iRhythm Technologies

<https://www.irhythmtech.com/>

Detecting cardiac dysrhythmias can be difficult in practice; with a wide variety of problems, ranging from innocuous ectopic and skipped beats, to life threatening arrhythmias. To differentiate in practice, a record of the event is needed, which can be hard to gather, often requiring expensive and repeated testing, often without fruition.

iRhythm Technologies is a leading digital health care company redefining the way cardiac arrhythmias are clinically diagnosed. Our Zio service provides uninterrupted, comprehensive monitoring to accurately and efficiently detect and diagnose heart conditions, such as arrhythmias, at the first time of asking. The company combines wearable biosensor devices worn for up to 14 days and cloud-based data analytics with powerful proprietary algorithms that distil data from millions of heartbeats into clinically actionable information. Published studies have shown improvements in arrhythmia detection and characterisation have the potential to transform clinical management of patients – eliminating the requirement for numerous patient visits due to indeterminate repeat tests and significantly improving the patient experience.

Kaido

<https://kaido.org/>

The Physical and Mental Health of employees is an increasingly important strategic priority for the modern employer due to its link with productivity and engagement. In the NHS, this has a direct correlation with the quality and standard of care that is delivered to patients.

Through a combination of fun and engaging company wide challenges, tailored health and wellbeing content and bespoke management reporting, Kaido empowers employees to take control of their wellbeing, whilst supporting NHS employers in

building a happier, healthier and more positive place to work. Kaido is already having a profound impact across the NHS working with employers such as The Royal Marsden Hospital NHS Foundation Trust, Great Ormond Street Hospital NHS Foundation Trust and North West Ambulance service and to date NHS employees have completed a massive 30,000,000 minutes of physical activity on the platform, and report a decrease in work-related stress (37%), an improved mood (40%) and increased workplace motivation (51%).

Barbara Johnson

Kortical

<https://kortical.com/case-studies/nhs>

Blood products have a short shelf life; in fact, blood platelets only last only 7 days. Ensuring all hospitals have a supply of the different blood types, at all times, is a complex problem, which involves understanding supply, manufacturing, distribution, stock holding, logistics and hospital demand levels, and necessitates an amount of waste. Kortical is using their platform to build Artificial Intelligence (AI)/Machine Learning (ML) models to predict demand and supply levels, allowing for patient needs to be met more precisely, reducing ad-hoc transport costs and improving efficiencies across the board.

Kortical have built an AI-powered platform to predict supply and demand for platelets for all the hospitals in England, which also takes into account diverse data such as weather, bank holidays, and understands that different trusts require different types of blood products, given the large regional differences.

The AI predicts supply, which varies with who comes in, and which blood type and volume of platelets they donate, as well as demand, 1 to 7 days out, and finally the platform optimises for logistics. All this to ensure that there are the right levels in the right place and demand is met with a little contingency on top to keep the nation covered. Testing against historical data the AI platform is performing better than previous demand and supply forecasting methods by over 10% and in the first quarter of 2020 Kortical will be able to release the figures in production, demonstrating their ability to reduce cost and waste without compromising patient care.

Lifelight

<http://www.lifelight.care/>

Vital signs monitoring forms the basis of a clinician's view of a patient, informing them over time about their baseline health and trajectory of change. Previously this information has been available through regular observations in a healthcare context, taking valuable staff time; however, patients have not been empowered to gather their own observations in a clinically valid way, and as such in the community this perspective is largely unavailable.

Lifelight is a unique software technology that measures blood pressure, heart rate, respiration and oxygen saturation in just 40 seconds simply by a patient looking into the camera on a standard smartphone or tablet. No cuff, wearables or contact needed. Lifelight's computer-vision, machine learning, algorithms are trained on data from an 8,500 patient clinical study at Portsmouth Hospitals Trust which recorded over 1 million heartbeats. Validated to clinical grade, Lifelight allows rapid triage in ED and urgent care, fast, contactless ward observations and ultimately, the ability for any patient at home to measure their own vital signs simply with their smartphone as part of 111 calls, remote consultations and outpatient appointments.

My Cognition

<https://mycognition.com/>

Poor cognitive fitness is a major component and a risk factor in mental illness. Long-term conditions have a negative impact on patients' mental health, their everyday activities, social relationships and self-management. 30-40% of people with a chronic condition have a mental health problem and conversely 46% of people with a mental health problem have a long-term physical health problem. Either way 50% - 90% of both populations will experience clinically defined cognitive impairment, which is impacting their healthcare outcomes and how well they manage their condition, this becomes more disabling as we age.

MyCognition App strengthens patient cognitive health, protects against cognitive decline, develops mental resilience and prevents and treats mental illness. Occupationally it promotes independence, improves quality of life and ability to self-manage physical health. The App measures, monitors and tracks cognitive health with its clinically validated 15-min digital cognitive assessment MyCQ, and then corrects cognitive deficits with a personalised training video game, AquaSnap. Healthy habits are encouraged with its positive behavioural and lifestyle change programme.

Roche Diabetes Care Platform

<https://www.diabetescareplatform.com/>

Nearly 1 in 10 people worldwide have diabetes, and the proportion is rising. This chronic condition covers a large number of causes; however, most common is Type 2 diabetes, closely linked to diet, weight and lifestyle. Despite its high prevalence, this condition has a high impact on morbidity, being a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation worldwide. The management of diabetes is complex and requires high levels of clinical coordination as well as patient understanding and engagement to ensure that short and long term complications do not ensue.

Roche Diabetes Care Platform is an online modular diabetes management platform that quickly connects you to actionable insights for a more personalized approach to diabetes care. The streamlined workflow of the platform allows for fast insights into key health and diabetes data, contributing to a clear and holistic picture of patient health. Moreover the platform aims to engage patients and promote better understanding and positive behaviour change.

Sensyne Health

<https://www.sensynehealth.com/>

The wealth of data generated in the healthcare system today poses many opportunities for improved care and increased efficiencies.

Sensyne health creates value from accelerating the discovery and development of new medicines and improving patient care through the analysis of real-world evidence from large databases of anonymised patient data in collaboration with NHS Trusts. Currently with four validated products in employment, Sensyne health works closely with clinical partners to ensure that products designed are fit for purpose in a clinical environment, making use of and working towards well defined clinically validated metrics. GDM-Health and EDGE empower patients to manage gestational diabetes and COPD more effectively in the community, whilst also facilitating the collection and analysis of relevant data. Together these applications save money and clinician time through helping patients avoid urgent and emergency hospital admission by helping to improve their understanding and management of their conditions, as well as providing simple, easy and effective digital contact to their clinicians.

SEND makes use of routine vital signs monitoring to calculate a dynamic risk score for patients, predicting likelihood of deterioration and allowing clinicians to monitor and manage a population of patients effectively, allocating resources based on clinical need. Moreover the analytic capacity of their products is employed in research and has to date identified a distinct phenotype of patients experiencing blood-pressure rises in the early morning, placing them at higher risk of an adverse event.

Sentinel

<https://www.sentinel sensor.co.uk/>

"On average patients are monitored only once every 8 hours on general wards and significantly less when discharged home. The UK NHS discharges 16.3m patients p.a. of which 14% will be re-admitted as an emergency admission within 30 days of discharge. This has a significant increased cost impact and results in increased mortality. These figures are broadly similar across the E.U."

As a consortium of university hospitals concluded, "Patients die because signs of deterioration are missed. There is a huge unfulfilled need for better monitoring of vital signs to identify high-risk patients who are on general hospital wards or at home."

The Universities calculated that they will save £4.3m for every 30k patients discharged if high-risk patients are continuously monitored in hospitals and at home, post-discharge.

Sentinel provides a continuous wearable monitor, currently undergoing MHRA conformity assessments, for known patterns in the patients' vital-signs every 60 seconds, 24x7. In doing so, Sentinel can detect if new conditions are developing or if the patient is deteriorating before the situation develops into an emergency, thus allowing for faster and more effective treatment.

Storm ID

<https://stormid.com/>

The management of chronic diseases occupies an ever larger proportion of hospital productivity. These complex and often co-morbid conditions provide an opportunity to develop new models of care allowing for intelligent predictive and preventative management in the community, ensuring that hospital resources are used to their maximum efficiency.

In the case of Chronic Obstructive Pulmonary Disorder (COPD), the second most common cause of emergency hospital admissions in the UK, accounting for 1 in 8 of all hospital admissions, the need is well defined. This chronic respiratory condition affects 1.2 million people in the UK and is forecast by the WHO to become the third leading cause of death worldwide by 2030. With the cost of emergency admission and management high, this new service combines patient generated health data, with clinical details pulled dynamically from electronic health records to present a real time view of COPD patients to care teams. Using predictive analytics, the service aims to stratify those patients at highest risk of exacerbation to prioritise those for intervention and prevent an emergency hospital admission. The service offers patients access to their personalised self management plan, details of standard and emergency medication and two way secure messaging with care team.

Storm ID have also launched new outpatient digital appointments services across NHS Trusts in Scotland which moves routine outpatient care to an asynchronous, virtual model that can be delivered up to 40% more efficiently. We are collaborating with NHS on new digital care pathways with integrated machine learning for Stroke, Sleep Apnoea and have a Retinopathy service in early planning stages. We believe that virtual care models will transform the routine treatment of chronic conditions to a management by exception care model where machine learning and rules based models will identify those patients in need of intervention.

Veye Chest

<https://www.lify.io/3p-products/veye-chest>

Lung cancer is the most common cause of cancer in men, and the third most common cause of cancer in the UK. Given that three quarters of lung cancers in the UK are detected at a late stage, lung cancer is often fatal and contributes to a large national and global burden of mortality.

Aidence has developed Veye Chest to aid early diagnosis of lung cancer on chest-CT scans. Their software automatically detects and evaluates pulmonary nodules, compares the patient's current scan to prior imaging, and produces an automated report for the radiologist. Their machine learning algorithms were trained using 45,000 chest-CT scans (NLST dataset) and a subsequent clinical validation study of VeyeChest was performed by the University of Edinburgh. Since receiving CE-certification, VeyeChest has been running in clinical practice in more than 10 sites across the UK and wider Europe.

Aidence's goals are to help improve efficiency in radiology departments, accuracy in early lung cancer detection, and to contribute to the NHS achieving successful and sustainable targeted lung health check programmes.

References

1. Russell, S. J., Norvig, P., Davis, E. & Edwards, D. *Artificial intelligence: a modern approach*. (Pearson, 2016).
2. *The Cambridge handbook of artificial intelligence*. (Cambridge University Press, 2014).
3. Minsky, M. L. *Semantic information processing*. (The MIT Press, 1968).
4. Corea, F. AI knowledge map: how to classify AI technologies, a sketch of a new AI technology landscape. *Medium* https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020. (2018).
5. Reddy, S., Fox, J. & Purohit, M. P. Artificial intelligence-enabled healthcare delivery. *J. R. Soc. Med.* **112**, 22–28 (2019).
6. Department of Health and Social Care. Code of conduct for data-driven health and care technology. *GOV.UK* <https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology> (2019).
7. Houssami, N., Kirkpatrick-Jones, G., Noguchi, N. & Lee, C. I. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev. Med. Devices* **16**, 351–362 (2019).
8. Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J. & Séroussi, B. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artif. Intell. Med.* **94**, 42–53 (2019).
9. Gupta, N., Gupta, D., Khanna, A., Rebouças Filho, P. P. & de Albuquerque, V. H. C. Evolutionary algorithms for automatic lung disease detection. *Measurement* **140**, 590–608 (2019).
10. Basturk, A., Yuksei, M. E., Badem, H. & Caliskan, A. Deep neural network based diagnosis system for melanoma skin cancer. in (2017). doi:10.1109/SIU.2017.7960563.
11. Albadawy, E. A., Saha, A. & Mazurowski, M. A. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing: Impact. *Med. Phys.* **45**, 1150–1158 (2018).
12. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
13. Kemper, J. & Kolkman, D. Transparent to whom? No algorithmic accountability without a critical audience. *Inf. Commun. Soc.* 1–16 (2018) doi:10.1080/1369118X.2018.1477967.
14. Machado, C., Burr, C., Morley, J., Taddeo, M. & Floridi, L. The Debate on the Ethics of AI in Health Care: a Reconstruction and Critical Review. *Forthcoming*.
15. Millett, S. & O'Leary, P. Revisiting consent for health information databanks. *Res. Ethics* **11**, 151–163 (2015).
16. Ploug, T. & Holm, S. Meta consent - A flexible solution to the problem of secondary use of health data. *Bioethics* **30**, 721–732 (2016).
17. Mann, S. P., Savulescu, J. & Sahakian, B. J. Facilitating the ethical use of health data for the benefit of society: Electronic health records, consent and the duty of easy rescue. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **374**, (2016).
18. Balthazar, P., Harri, P., Prater, A. & Safdar, N. M. Protecting Your Patients' Interests in the Era of Big Data, Artificial Intelligence, and Predictive Analytics. *J. Am. Coll. Radiol. JACR* **15**, 580–586 (2018).
19. Floridi, L. et al. Key Ethical Challenges in the European Medical Information Framework. *Minds Mach.* 1–17 (2018) doi:10.1007/s11023-018-9467-4.
20. Voigt, K. Social Justice, Equality and Primary Care: (How) Can 'Big Data' Help? *Philos. Technol.* **32**, 57–68 (2019).
21. Holzinger, A., Haibe-Kains, B. & Jurisica, I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *Eur. J. Nucl. Med. Mol. Imaging* (2019) doi:10.1007/s00259-019-04382-9.
22. Kleinpeter, E. Four Ethical Issues of "E-Health". *IRBM* **38**, 245–249 (2017).
23. Chiauuzzi, E. & Newell, A. Mental health apps in psychiatric treatment: A patient perspective on real world technology usage. *J. Med. Internet Res.* **21**, (2019).
24. Krutzinna, J., Taddeo, M. & Floridi, L. Enabling Posthumous Medical Data Donation: An Appeal for the Ethical Utilisation of Personal Health Data. *Sci. Eng. Ethics* (2018) doi:10.1007/s11948-018-0067-8.
25. Shaw, D. M., Gross, J. V. & Erren, T. C. Data donation after death: A proposal to prevent the waste of medical research data. *EMBO Rep.* **17**, 14–17 (2016).
26. Sterckx, S., Rakic, V., Cockbain, J. & Borry, P. "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. *Med. Health Care Philos.* **19**, 177–190 (2016).

27. Celi, L. A. *et al.* Bridging the health data divide. *J. Med. Internet Res.* **18**, (2016).
28. Kuek, A. & Hakkennes, S. Healthcare staff digital literacy levels and their attitudes towards information systems. *Health Informatics J.* 146045821983961 (2019) doi:10.1177/1460458219839613.
29. Aitken, M. *et al.* Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. *Int. J. Popul. Data Sci.* **4**, (2019).
30. Page, S. A., Manhas, K. P. & Muruve, D. A. A survey of patient perspectives on the research use of health information and biospecimens. *BMC Med. Ethics* **17**, (2016).
31. Barras, C. Mental health apps lean on bots and unlicensed therapists. *Nat. Med.* (2019) doi:10.1038/d41591-019-00009-6.
32. Ferretti, A., Ronchi, E. & Vayena, E. From principles to practice: benchmarking government guidance on health apps. *Lancet Digit. Health* **1**, e55–e57 (2019).
33. Henson, P., David, G., Albright, K. & Torous, J. Deriving a practical framework for the evaluation of health apps. *Lancet Digit. Health* **1**, e52–e54 (2019).
34. Larsen, M. E. *et al.* Using science to sell apps: Evaluation of mental health app store quality claims. *Npj Digit. Med.* **2**, 18 (2019).
35. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data Soc.* **3**, 205395171667967 (2016).
36. Hailu, R. Fitbits and other wearables may not accurately track heart rates in people of color. *STAT* (2019).
37. Wachter, R. M. *The digital doctor: hope, hype, and harm at the dawn of medicine's computer age.* (McGraw-Hill Education, 2015).
38. Liu, C. *et al.* Using Artificial Intelligence (Watson for Oncology) for Treatment Recommendations Amongst Chinese Patients with Lung Cancer: Feasibility Study. *J. Med. Internet Res.* **20**, e11087 (2018).
39. Garattini, C., Raffle, J., Aisyah, D. N., Sartain, F. & Kozlakidis, Z. Big Data Analytics, Infectious Diseases and Associated Ethical Impacts. *Philos. Technol.* **32**, 69–85 (2019).
40. Maher, N. A. *et al.* Passive data collection and use in healthcare: A systematic review of ethical issues. *Int. J. Med. Inf.* **129**, 242–247 (2019).
41. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).

42. Racine, E., Boehlen, W. & Sample, M. Healthcare uses of artificial intelligence: Challenges and opportunities for growth. *Healthc. Manage. Forum* (2019) doi:10.1177/0840470419843831.
43. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: Addressing ethical challenges. *PLoS Med.* **15**, e1002689 (2018).
44. Challen, R. *et al.* Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).
45. Morley, J. & Joshi, I. Developing effective policy to support Artificial Intelligence in Health and Care. *Eurohealth* **25**, (2019).
46. Floridi, L. Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philos. Transact. A Math. Phys. Eng. Sci.* **376**, (2018).
47. Floridi, L. *et al.* AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **28**, 689–707 (2018).
48. Leslie, D. *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.* <https://zenodo.org/record/3240529> (2019) doi:10.5281/zenodo.3240529.
49. Gebru, T. *et al.* Datasheets for Datasets. *ArXiv180309010 Cs* (2018).
50. ODI. Data Ethics Canvas User Guide. https://docs.google.com/document/d/1MkvoAP86CwimbBD0dxySVCO0zeVOput_bu1A6kHV73M/edit.
51. Mitchell, M. *et al.* Model Cards for Model Reporting. *Proc. Conf. Fairness Account. Transpar.* - FAT 19 220–229 (2019) doi:10.1145/3287560.3287596.
52. Steventon, A., Deeny, S. R., Keith, J. & Wolters, A. T. New AI laboratory for the NHS. *BMJ* **l5434** (2019) doi:10.1136/bmj.l5434.
53. Department of Health and Social Care. UK National Screening Committee. <https://www.gov.uk/government/groups/uk-national-screening-committee-uk-nsc> (2019).
54. Department of Health and Social Care. Joint Committee on Vaccination and Immunisation. <https://www.gov.uk/government/groups/joint-committee-on-vaccination-and-immunisation>.
55. NICE. *The Guidelines Manual.* <https://www.nice.org.uk/process/pmg6/chapter/assessing-cost-effectiveness> (2012).

56. Murray, E. et al. Evaluating Digital Health Interventions. *Am. J. Prev. Med.* 51, 843–851 (2016).
57. Clarke, G. M., Conti, S., Wolters, A. T. & Steventon, A. Evaluating the impact of healthcare interventions using routine data. *BMJ* l2239 (2019) doi:10.1136/bmj.l2239.
58. WHO. *Monitoring and evaluating digital health interventions A practical guide to conducting research and assessment*. <https://www.who.int/reproductivehealth/publications/mhealth/digital-health-interventions/en/> (2016).
59. The Health Foundation. Improvement Analytics Unit. <https://www.health.org.uk/funding-and-partnerships/our-partnerships/improvement-analytics-unit>.
60. Steventon, A. How to enable the uptake of technology in the NHS? <https://www.health.org.uk/blogs/how-to-enable-the-uptake-of-technology-in-the-nhs> (2018).
61. WHO. Focus Group on 'Artificial Intelligence for Health'. <https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx>.
62. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* 1, e271–e297 (2019).
63. Liu, X., Faes, L., Calvert, M. J. & Denniston, A. K. Extension of the CONSORT and SPIRIT statements. *The Lancet* 394, 1225 (2019).
64. The CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med.* (2019) doi:10.1038/s41591-019-0603-3.

Acknowledgements

We would like to thank each individual person who contributed to the writing, editing, designing and publishing of this report:

Asheem Singh	Melissa Ream	Lizzie Barclay	Tom Stocker
Charlotte Holloway	Professor Chris Holmes	Barbara Johnson	Joseph Connor
Kayshani Gibbon,	Professor Sir Adrian Smith	Michael Morgan-Curran	Colin Pattinson
Eleonora Harwich		Juliet Tizzard	Sana Khareghani
Claudia Martinez	Tina Woods	Clara Lubbers	Edward Teather
Olly Buston	Marie-Anne Demesthias	Meredith Makeham	Farzana Dudhwala
Dr. Matthew Feneck	Dr. Adrian Price	Rodney Ecclestone	Jay Ashton-Butler
Nika Strukelj	Laurence Thorne	Sameer Pujari	Rich Westman
Areeq Chowdhury	Simon Harris	Clayton Hamilton	Laurence Pearce
Caio Machado	Dave Hurhangee	Michael Macdonnel	Tom Hardie
Dr. Christopher Burr	Nickolai Vysokov	Alexa Knight	Tim Horton
Joshua Cows	James Teo	Charlotte Chorley	Dr. Hannah Knight
Dr. Mariarosaria Taddeo	Mark Halling-Brown	Dr. Geraldine Clarke	Abdullah Mahmood
Professor Luciano Floridi	Joseph Waller	Dr. Adam Steventon	Professor Alastair Denniston
Anna Steere	Alexander Börve	Josh Keith	Dr. Xiaoxuan Liu
Sile Hertz	Dafydd Loughran	Dawn Monaghan	Professor Philip Beales
Hugh Hathaway	Paul McGinness	Dr. Sarah Deeny	Anna Tomlinson
Dawn Monaghan	Niall Haslam	Caroline Cake	MJ Caulfield
Dr. Natalie Banner	Kirsty Atkinson	John Hatwell	John Hatwell
Clara Lubbers	Waqar Ali Shah	Professor Philip Beale	Ian Newington
Mark Salmon	Piyush Mahaptara	Chris Wigley	
Bernice Dillon	Conn O'Neil	Paul Sansom	



THE BLUE RIDGE ACADEMIC HEALTH GROUP

*Separating Fact
from Fiction:
Recommendations for
Academic Health Centers
on Artificial and
Augmented Intelligence*

Members and participants

(June 2018 meeting)

MEETING CO-CHAIRS

Jonathan S. Lewin, MD (current Co-Chair)
Executive VP for Health Affairs, Emory University
Executive Director, Woodruff Health Sciences Center
President, CEO, and Board Chairman, Emory Healthcare

Jeffrey R. Balser, MD, PhD (current Co-Chair)
President and CEO, Vanderbilt University Medical Center
Dean, Vanderbilt University School of Medicine

OTHER MEMBERS

S. Wright Caughman, MD
Professor, Emory School of Medicine and Rollins School of
Public Health, Emory University
Emeritus Executive VP for Health Affairs, Emory University

Michael V. Drake, MD
President, Ohio State University

Julie Freischlag, MD (as of September 2018)
Dean, Wake Forest University School of Medicine
CEO, Wake Forest Baptist Medical Center

Michael M.E. Johns, MD
Professor, Emory School of Medicine and Rollins School of
Public Health; Emory Emeritus Executive VP for Health
Affairs; Emeritus President, CEO, and Board Chairman,
Emory Healthcare

Darrell G. Kirch, MD
President, Association of American Medical Colleges

Steven Lipstein
Former President and CEO, BJC Health Care (retired)
Member, Emory University Board of Trustees

Lloyd B. Minor, MD
Dean, Stanford University School of Medicine

Elizabeth G. Nabel, MD (as of September 2018)
President, Brigham Health

Mary D. Naylor, PhD
Marian S. Ware Professor in Gerontology
Director, NewCourtland Center for Transitions & Health
University of Pennsylvania School of Nursing

Daniel K. Podolsky, MD (as of September 2018)
President, University of Texas Southwestern Medical Center

Kenneth S. Polonsky, MD
Executive VP for Medical Affairs; Dean, Biological Sciences
Division and School of Medicine, University of Chicago

Claire Pomeroy, MD, MBA
President, Albert and Mary Lasker Foundation

Marschall S. Runge, MD, PhD
Executive VP for Medical Affairs, University of Michigan

Fred Sanfilippo, MD, PhD
Director, Healthcare Innovation Program, Emory University

Richard P. Shannon, MD
Executive VP for Health Affairs
University of Virginia Health System

Irene M. Thompson
Vice Chair, AMC Networks Board of Managers, Vizient, Inc.

SENIOR MEMBERS

William R. Brody, MD, PhD
Professor Emeritus
Salk Institute for Biological Studies

Don E. Detmer, MD, MA
Professor of Medical Education
University of Virginia

Michael A. Geheb, MD
Senior Consulting Director, IBM Watson Health
Board of Directors, Wayne State University Physicians Group

William N. Kelley, MD
Professor of Medicine, Perelman School of Medicine,
University of Pennsylvania
Trustee Emeritus, Emory University

Arthur H. Rubenstein, MBBCh
Professor of Medicine, Perelman School of Medicine
University of Pennsylvania

John D. Stobo, MD
Executive VP, UC Health, University of California

Bruce C. Vladeck, PhD
Former CMS Administrator and Health Policy Adviser

FEATURED PRESENTERS

Gari D. Clifford, PhD
Interim Chair, Biomedical Informatics
Emory School of Medicine

Kyu Rhee, MD, MPP
Vice President and Chief Health Officer, IBM Corporation

William B. Rouse, PhD
Chair, Economics of Engineering
Stevens Institute of Technology

Khan Siddiqui, MD
Chief Technology Officer & Chief Medical Officer
higi SH Holdings, Inc.

FACILITATORS

Steve Levin
Director, The Chartis Group

Ryan Bertram
Associate Principal, The Chartis Group

Alexandra Schumm, Principal and VP for Research
The Chartis Group

GUESTS

John G. Rice
Vice Chairman, General Electric (retired)

William W. Stead, MD
Chief Strategy Officer
Vanderbilt University Medical Center

STAFF

Anita Bray
Project Coordinator, Woodruff Health Sciences Center,
Emory University

Gary L. Teal
Vice President, Woodruff Health Sciences Center
Emory University

REPORTER AND EDITOR

Alexandra Schumm (See Facilitators)

EDITORIAL AND DESIGN CONSULTANTS

Karon Schindler, Peta Westmaas
Woodruff Health Sciences Center, Emory University

Contents

Report 23. Separating Fact from Fiction: Recommendations for Academic Health Centers on Artificial and Augmented Intelligence

Introduction: The Digital Transformation in Health Care 2

I. Vision of the Future. 2

II. A Principled Approach to Artificial Intelligence. 5

III. The Promise of Artificial Intelligence/Machine Learning: Applications and Implications Across the Tripartite Mission. 6

IV. A Call to Action for Academic Health Systems 17

References 19

Appendix 23

About the Blue Ridge Academic Health Group 24

Previous Blue Ridge Reports 25

Mission *The Blue Ridge Academic Health Group seeks to take a societal view of health and health care needs and to identify recommendations for academic health centers (AHCs) to help create greater value for society. The Blue Ridge Group also recommends public policies to enable AHCs to accomplish these ends.*

Reproductions of this document may be made with written permission of Emory University’s Woodruff Health Sciences Center by contacting Anita Bray, Emory School of Medicine, James B. Williams Medical Education Building, 100 Woodruff Circle, Suite A367, Atlanta, GA, 30322. Phone: 404-712-3510. Email: abray@emory.edu.

Recommendations and opinions expressed in this report represent those of the Blue Ridge Academic Health Group and are not official positions of Emory University. This report is not intended to be relied on as a substitute for specific legal and business advice. Copyright 2019 by Emory University.

Introduction: The Digital Transformation in Health Care

Countless papers, including the Winter 2017-2018 report,¹ of the Blue Ridge Academic Health Group (BRAHG) acknowledge that the promise of electronic health records (EHRs) has yet to be fully realized. While we are learning to use and optimize EHRs, we may be exposed to new hurdles and concerns such as clinician burnout due, in part, to the additional clerical burden associated with most EHRs. However, we need to think much broader than EHRs, as a true digital transformation is under way in health care that goes far beyond this tool.

Today's health care professionals are bombarded by data from multiple sources—diagnostic, claims, financial, psychosocial, epidemiologic, biometric, genomic, consumer-generated—at ever-increasing rates. Each person is estimated to generate enough health data to fill 300 million books in a lifetime²—equivalent to filling the Library of Congress nearly 20 times over. In fact, health care data is expected to double every 73 days through 2020. **Even if all this data were clean and harmonized,** its sheer quantity necessitates new tools to make sense of the ever-expanding data universe. This new requirement goes beyond big data analytics; it requires a fundamental shift in how we generate, curate, clean, and interpret data. Enter artificial intelligence (AI), which can be used to enable human “augmented intelligence.” For definitions, see the Definitions sidebar.

The purpose of this report is to clearly define artificial and augmented intelligence, explore its potential impact on health and health care, identify and react to challenges and risks associated with AI-based technologies in health care, and present a call to action for academic health systems (AHSs).

I. Vision of the Future

To help illustrate the transformative power of AI, consider the following vignette in which an AI-powered cognitive assistant plays an important role in a simple primary care visit:

Dr. Edward Cummings, a primary care physician, was wrapping up his clinical day. He sat at his desk, facing a large display connected to his laptop. Millie, Dr. Cummings' cognitive assistant, was on the upper left of the display with a Skype-like visage.

EC: “Pretty routine day, Millie. Do you agree?”
Millie: “Perhaps, except for Mr. Burdell.”
George Burdell had presented with symptoms of a sore throat and right ear pain for the last three weeks.
Millie: “You examined and found his ear drums and canal normal but also found an ulcer at the base of the right tonsil.”
EC: “Right, and I referred him to a local otolaryngologist—head and neck surgeon—for evaluation.”
Millie: “With that constellation of symptoms and signs, this is almost certainly a head and neck cancer.”
EC: “Yes, I thought of that, why do you bring it up?”
Millie: “Well, I took the liberty of reviewing the literature on the subject, and it turns out that survival is significantly improved if the patient is treated and evaluated at a major center or an NCI [National Cancer Institute] comprehensive cancer center.”
EC: “Well, if that is the case, perhaps we should have him evaluated down in Atlanta—can you set up that referral please?”
Millie: “Yes, I would be happy to. While we are discussing it, did you know Mr. Burdell has been your patient for many years? You last saw him four years ago for what turned out to be a sinus infection.”
EC: “I don't recall that, but I am sure you are right.”
Millie: “He has never had a routine physical. It seems that he only shows up when he has an obvious health problem.”
EC: “Lots of people are like that.”
Millie: “Yes, but Mr. Burdell may be facing something serious, and he does not seem like the type of person who takes care of himself.”
EC: “Could be. We will worry about that after we see what the folks in Atlanta have to say.”
Millie: “OK. I will send you the read-aheads for tomorrow.”
EC: “Great. You know how to find me if anything comes up.”
Millie: “Of course, good night.”

Dr. Henry Ramsey is an otolaryngologist (a head and neck surgeon). Ramsey and his cognitive assistant, Peggy, review his day's work, including the examination of Burdell.
Peggy: “As you know, you found a 2.5-cm ulcerated mass at the base of the right tonsil and a 2-cm cervical

Definitions

Artificial and augmented intelligence are often confused or conflated with other related terms. Following are definitions for reference:

■ Artificial intelligence

- The ability for a program to make predictions or decisions or take actions based on insights developed by machine learning algorithms.

■ Augmented intelligence

- Normal human intelligence is supplemented through the use of artificial intelligence. It is a complement to, not a replacement of, human intelligence.

■ Algorithms

- Math formulas and/or programming commands that inform a regular non-intelligent computer on how to solve problems. Algorithms are rules that teach computers how to figure things out on their own.
- “Care pathways” also can be referred to as algorithms when they include a flow chart format of decisions and actions in cases where a patient with a particular condition or set of conditions is treated in a sequence of steps.

■ Machine learning

- A method of data analysis that automates analytical model building, where computers can be supplied with massive amounts of data and can “learn” how to complete a specific task or make a prediction or decision based on patterns it identifies in that data. Traditional machine learning requires labeled data, where humans provide the set of initial rules, data features (individual characteristics or variables—input), and data labels (output), then the computer learns by applying those to a dataset. The resulting algorithm is refined as the computer is provided with and analyzes new datasets. Humans correct any errors the machine makes.

■ Artificial neural network

- A computer analysis approach modeled on the layered, dense, and interconnected structure of neurons in a human brain. While the concept of artificial neural networks has been around for some time, increasingly powerful computers and advanced mathematical formulas now enable algorithms to be developed based on many more layers of data in much more complex neural networks than in the past.

■ Deep learning

- Deep learning is a form of machine learning, where artificial neural networks are used because of the multiple layers of complex data involved. The computer learns to recognize patterns in the layers of data and make decisions about the data—for example, to analyze an image or sound and determine what it is or means.

■ Supervised and unsupervised learning

- Supervised learning: A type of machine learning in which human data labeling and supervision are an integral part of the machine learning process on an ongoing basis. In supervised learning, there is a clear outcome to the machine's data mining, and its target function is to achieve this outcome, nothing more.
- Unsupervised Learning: A type of machine learning in which human input and supervision are extremely limited or absent altogether throughout the process. In unsupervised learning, the machine is left to identify patterns and draw its own conclusions from the datasets it is given, without a specific outcome identified. The computer does not know if the patterns it identifies will be useful—humans must design experiments to test efficacy or effectiveness of the patterns the machine identifies.
- In supervised deep learning, the data is labeled by a human. In unsupervised deep learning, the data is unlabeled, and the computer assesses patterns, labels data accordingly, and develops algorithms on its own.

■ Predictive analytics

- Often classified as a type of machine learning, predictive analytics uses advanced analytic techniques on large datasets to identify patterns and make predictions about future events.

■ Interrogation of an AI-driven model

- The concept of “interrogating” a computer about how it arrived at its conclusions based on its machine learning-derived model. This is particularly relevant in many areas of deep learning, where the neural networks upon which the computer learned are so complicated that it would be difficult to successfully interrogate the model. The incredibly complex models being developed for things such as autonomously driven cars, for example, may not provide transparency or the ability to interrogate. This would be acceptable if the models are always “correct,” but less so if the model makes a mistake—one of the cars suddenly drives into a river, for example—and we do not have the ability to understand why.

lymph node in the upper right jugular area.”

HR: “Yes, and I performed a biopsy in the clinic of the tonsil mass and sent it to pathology.”

Peggy: “In reviewing Burdell’s history, I noticed that he smokes a pack of cigarettes a day. He also has alcohol challenges.”

HR: “Those are certainly major risk factors.”

Peggy: “Over the past 10 years, you have seen 12 patients with very similar symptoms and risk factors.”

HR: “Anything else common to these patients?”

“They all faced a tough road to recovery.”

HR: “How so?”

Peggy: “Many more follow-ups than expected.”

HR: “And the results?”

Peggy: “70% recovered, with no recurrence of cancer.”

HR: “Good. Let’s sustain this batting average.”

As Ramsey is driving home, Peggy calls him in his car.

Peggy: “The path report of the biopsy showed a diagnosis of squamous cell carcinoma, moderately differentiated.”

HR: “Hmm, interesting. Did they test for p16 as I requested?”

Peggy: “Yes, he was p16+, and as you know the American Joint Commission changed the staging system in late 2017 for oropharyngeal carcinoma that is HPV p16+.

HR: “Yes I am aware of that, and this makes him T2N1Mx Stage I using the new system.

HR: “OK, so please call Mr. Burdell and have him come in for an appointment this week. At the same time please order a PET/CT with contrast to evaluate the extent of disease and if there is any distant metastasis. I would expect there to be none.”

Peggy: “Should I provide the usual explanation.”

HR: “Yes, of course, please let him know that I will talk with him in the clinic. I want to tell him about the cancer in person.”

Peggy: “It looks like his PET/CT scan will be completed in one week. Would you like him added to the tumor board for that following week?”

HR: “Yes, and also please set up appointments with speech pathology and social work along with nutrition. Also please review our clinical trial portfolio, and let me know what is potentially available to Mr. Burdell based on his p16 status.”

Peggy: “I assume you want them to see the patient after you have informed him of the cancer.”

HR: “Exactly.”

Peggy: “By the way, I have found that, over the last two years, more than 100 studies have been published on alternative treatment protocols of squamous cell carcinoma in the oropharynx related to HPV. Do you want a summary of these studies?”

HR: “Is there anything unusual or surprising in this literature?”

Peggy: “No, not really.”

HR: “OK, but let’s have the summary available for the meeting of the tumor board.”

Peggy: “Will do.”

HR: “Also, make your usual scan of the SPOHNC website and see if there are national trials that we don’t participate in.”

Peggy: “Of course. Very familiar with the site, as you know.”

HR: “OK. Enough for today.”

To continue reading this vignette, please see the appendix on page 23.

The scenario described in the vignette may seem a bit futuristic, but many of the AI applications described therein are already in development or in use today. BRAHG recognizes that AI has significant potential to fundamentally change how patients receive and clinicians provide care, but we do not anticipate a dystopian world some futurists predict wherein robots replace the human workforce. Instead, we expect that AI will augment and enhance a clinician’s ability to diagnose and treat patients and will help further personalize the patient experience. Furthermore, we do not pretend to know the speed at which change will occur.

However, it would be myopic to overlook the transformational power these new technologies are likely to have across each component of the AHS tripartite mission. The following framework helps organize the complicated landscape of AI applications across these components:

Clinical domain—including applications for clinical decision support (e.g., diagnosis, treatment options and selection, identification of high-risk patients), apps that help patients and their caregivers prepare for and manage health needs, and operational applications of AI to reduce clerical work and take over and improve back-office functions (e.g., claims submissions and follow-up).

Educational domain—including applications to aid in teaching (e.g., virtual patients, virtual surgery, both of which are not AI themselves but leverage AI and are based on machine learning), as well as changing the type of content to prepare the next generation of health care professionals to incorpo-

rate AI-driven tools into their daily practice.

Research and discovery domain—including applications for basic and translational research, clinical trials, voice-based data entry, and cognitive assistant to comb journals/curate and synthesize study findings, as well as find inconsistencies in the evidence.

AI systems and applications that leverage AI and machine learning are expected to improve efficiency, reduce errors, and help proactively manage population and individual health, while also impacting both patient and clinician experience and patient engagement, with the potential to alleviate and reduce clinician burnout, which has reached epidemic proportions.¹

A recent Accenture study forecast that health care applications of AI will grow to a \$6.6 billion market by 2021.³ In that same vein, a recent Healthcare Information and Management Systems Society survey identified population health and clinical decision support as top areas of potential for AI.⁴ Given this aggressive growth and high potential to fundamentally change how clinicians practice and how patients interact with the health care ecosystem, it is imperative that health care professionals, academic health system leaders, and policymakers understand this topic.

Despite this explosive projected growth, the health care delivery system lags in AI adoption relative to other industries, such as finance and manufacturing.⁵ Health care’s lower adoption rates reflect its greater complexity, including unique moral, ethical, legal, and regulatory issues. These challenges will be explored later in this report.

This year’s report explores how AI will transform various dimensions of health care, though perhaps not at the pace at which some pundits have predicted. Furthermore, we share the American Medical Association’s position that AI solutions that extend and enhance the work of clinicians are more likely to succeed than solutions that replace them. Academic health center leaders have a responsibility to help their own organizations adapt and to actively shape how the use of AI unfolds across the tripartite mission in an ethical, value-driven manner. Herein, we explore emerging applications of AI across the tripartite mission and the implications for academic

medicine and health care more broadly. Finally, we offer a pragmatic call to action for academic medicine leaders to ensure that we are purposeful in our approach to AI.

II. A Principled Approach to Artificial Intelligence

AI presents both great promise and great challenges to AHSs and to health care more broadly. As such, AHSs should arm themselves with a decision-making framework for selection and deployment of AI to prepare for and manage the transformation. As a starting point, it is helpful to recognize where humans excel compared with machines.⁶

Humans excel at:

- Consciousness
- Planning and executive functions
- Creativity and imagination
- Emotion and empathy
- Complex problem solving and dilemmas
- Morals
- Abstract thinking

Computers excel at:

- Input and output
- Information processing, capacity, and memory
- Pattern identification and inconsistency identification

Humans and computers each excel at:

- Vision
- Language
- Complex movement
- Structured problem solving

Understanding these differences is useful in helping identify situations where AI applications may be more effective in creating value. If we don’t control for the aforementioned human factors, AI applications may fall short of our needs, or worse yet, cause harm. AHSs are best positioned to advance AI because of our large numbers of clinician scientists and our access to expertise in other needed fields in most of our universities. However, for many AHSs this is fairly or entirely new and uncharted territory. As such, we recommend that AHSs adopt a set of four core principles to drive decisions:

PRINCIPLES TO ADVANCE AI IN ACADEMIC MEDICINE

1. Adopt an ethical framework to ensure societal benefit in areas of scientific exploration and application of new knowledge. Note that various organizations have explored these issues, providing AHSs a starting point if desired (e.g., the Institute of Electrical and Electronic Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems).⁷ A legal approach—and potentially a new law or set of laws—will likely need to be formed to handle the instances when a clinician does not agree with an AI-driven diagnosis or treatment, or when a mistake or bad outcome occurs in part because of the recommendations of an AI-powered tool. These will take place outside of the domain of AHSs, but AHSs should stay abreast of such legal advancements and make adjustments to their own policies and procedures as necessary.
2. Focus on solving real health care problems (i.e., use cases) rather than on the technology, and then start with the problems where improvements would be of highest value.
3. Promote human-centered design as you re-imagine workflows that will incorporate AI.
4. Promote data literacy—develop mechanisms to collaborate across different domains of knowledge, which often come from other schools within the university or other organizations.

Incorporating these principles into future institutional planning activities should help ensure that AHSs help participate in shaping how AI affects health care, as opposed to passively experiencing the change and being ill-prepared for its impact.

III. The Promise of AI/Machine Learning: Applications Across the Tripartite Mission

There are many promising health care AI applications that could positively impact all three elements of an AHSs tripartite mission:

- Improving health care quality, health outcomes, efficiency, and patients’ family caregivers’, and

- clinicians’ experience—and reducing the cost of care;
 - Augmenting research capabilities and shortening clinical trial timelines; and
 - Enhancing educational tools and resources, while also expanding educational curricula and skill requirements for health care professionals.
- Many AI-based applications are being explored, developed, piloted, and in some cases rolled out by AHSs, health profession schools, and other institutions. We believe AI applications described herein warrant discussion and consideration for investment, with recognition of the very real hurdles in their continued development and adoption.

CLINICAL DOMAIN EXAMPLES

While the use of AI in the clinical setting is at the leading edge or experimental—few tools are yet deployed broadly for everyday practice—there are numerous opportunities to leverage AI to transform health care delivery and improve value, defined as improving quality, safety, and outcomes efficiency. Illustrative opportunities include the following:

Clinical decision support and predictive analytics tools based on AI algorithms

Augmented clinical support tools driven by predictive analytics will enable health professionals to do less “hunting and searching” for information, helping them find clinical analogs to an individual patient’s condition and health history, best practices, and clinical trials to aid in making a precise, accurate diagnosis and decide on the best course of treatment. Clinicians will have more time to evaluate and interpret the available relevant data, synthesize results, and design a personalized course forward for a patient. Patient outcomes will improve as their conditions can be monitored, flagged, and addressed earlier and most appropriately. Examples of these types of tools and applications include the following:

- Diagnosis support—*
- Medical imaging is arguably the most promising clinical area for AI-powered applications to be put into practice and to make an impact. Such applications can help radiologists and other clinicians make diagnoses and prioritize

patients by reading images and comparing them with hundreds of thousands of medical images while also taking into consideration other medical data related to the patient. The American College of Radiology (ACR) is in the process of developing ACRassist, a tool that combines raw clinical content and a communication framework that facilitates content delivery. It uses natural language processing developed through deep learning and brings evidence-based guidelines for recommendations and reporting into a radiologist’s workflow, providing guidance during interpretation and incorporating structured data elements into free-form reporting.^{8,9} In April 2018, the FDA approved the first autonomous diagnostic tool to diagnose diabetic retinopathy. For more detail on this example, see the IDx-DR sidebar.

- Pathology is another area where AI can help clinicians reach a diagnosis more efficiently and accurately. In a 2017 study, researchers at

Case Western Reserve University developed a deep learning algorithm based on 400 biopsy images from multiple hospitals and 200 images from the Cancer Genome Atlas and University Hospitals Cleveland Medical Center. In a study with 600 participants, the algorithm was able to accurately identify the presence of invasive types of breast cancer based on pathology images 100% of the time.¹⁰

- Cardiology is another clinical area that may benefit from AI-driven diagnosis tools. Completing the typical testing required to inform a diagnosis of coronary artery disease can often take a patient weeks or months because of scheduling restraints and common “re-tests” requested by physicians, particularly if the patient is accessing a health system that isn’t coordinated across departments or is relying on third party testing entities. Emerging AI-powered testing devices, such as a scanning tool created by Analytics 4 Life Inc., can take a minutes-

Clinical AI Application Case Study: IDx-DR

- **Problem**
Diabetic retinopathy is one of the leading causes of blindness in adult, working-age men and women in the United States, and the CDC estimates that between 2010 and 2015 the number of adults with diabetic retinopathy is expected to double—from 7.7 million people to 14.6 million.¹¹ If diabetics regularly receive eye exams, early detection is more likely and vision loss/blindness rates can be reduced.¹² However, like the general population in the U.S., less than half of diabetic patients receive annual eye exams.^{13, 14}
- **Solution**
A tool was developed by the IDx company to autonomously detect the presence of diabetic retinopathy. The tool detects lesion characteristics indicative of diabetic retinopathy from images of a patient’s eye and brings its findings together into a diagnosis using an algorithm developed through machine learning.¹⁵ After a clinical trial showed that the tool outperformed pre-set sensitivity and specificity thresholds, the FDA approved the use of this tool in a clinical setting in April 2018.¹⁶ This is one of the first AI-powered autonomous medical diagnostic systems to be approved by the agency.
- **Impact**
Because of the “autonomous” nature of the IDx-DR tool, diagnosis can be reached without a human clinician trained in reading retinal images. The system can therefore be used in the primary care setting, or potentially a retail setting, with simple usage training provided by IDx to staff. While it is too early to know how fast the use of this autonomous diagnostic tool will spread, if it is brought into a wide variety of settings it has enormous potential to reach more diabetic patients, diagnose more diabetic retinopathy cases earlier, and ultimately reduce vision loss and blindness in diabetic patients.
- **Caveat**
While this example is an interesting one with substantial potential to positively impact many patients’ health, we do not anticipate that most diagnostic imaging tools supported by AI will be autonomous. Rather, they will provide information and clinical support to augment the information available to clinicians, improving their diagnostic capabilities. At least for the foreseeable future, these tools will still require human interpretation or human “approval” of a diagnosis recommendation.

long recording of a patient's heart function and collect more than 10 million data points.¹⁷ The data can be analyzed through special software, and a three-dimensional image of the patient's heart along with a detailed report can then be sent to the clinician. The patient can receive a diagnosis after just one test, saving time, reducing costs, and potentially introducing an intervention or treatment sooner and avoiding an adverse event. Analytics 4 Life is in the process of conducting a two-stage study with 12 clinical partners, including Ochsner Health System, Rochester Regional Health, and Sentara Heart Hospital. The first stage consists of a prospective, non-randomized trial to develop a machine-learning algorithm to detect and assess significant coronary artery disease. The second stage consists of a prospective, blinded, non-randomized, paired comparison trial to test the machine learning algorithm.¹⁸

- It is important to note that AI-driven diagnostic recommendations are not always better or faster. Contrary to the previous example, some algorithms will recommend or require many tests and datapoints before a diagnosis recommendation can be made by the machine, whereas the human doctor would arrive at the diagnosis without needing some of those tests. This gap may be reduced as algorithms improve, but for now it is important to note that not all emerging AI-driven diagnosis tools are “better” or more efficient than humans alone.

Tailoring treatment and precision medicine—

- Precision medicine is defined by the National Institutes of Health as “an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.”¹⁹ Clinicians can use this information with the help of AI-powered tools to predict more accurately which medications, treatments, health management, and prevention approaches will work best for a particular patient. In one example, clinicians at Holston Medical Group in northeast Tennessee and southwest Virginia are utilizing a clinical-genomic application called 2bPrecise, embedded in their EHR. The application analyzes drug-to-gene interaction and

applies pharmacogenomics patterns—identified through machine learning during the development of the application—to determine the most effective drug regimen for each patient. The group is using this tool to help improve behavioral health care and tailor treatment plans for opioid patients. The tool was initially tested and adopted by experts at the NIH, including those at the National Human Genome Research Institute, the National Cancer Institute, and the NIH Clinical Center.²⁰

- Another growing area of precision medicine is precision oncology, where molecularly targeted drug therapies are prescribed based on the genomic status of the drug target in order to maximize efficacy. However, the more commonly prescribed early treatment chemotherapy drugs—those that are “nonspecific”—do not have biomarkers and therefore the genomic-based precision oncology approach does not apply. Recognizing this limitation, a team of researchers recently applied deep learning techniques to identify features within genome-scale omics data to develop a model that predicts the effectiveness of drugs in the treatment of various cancer cell lines, regardless of whether the drugs are molecularly targeted or nonspecific. While the results were promising—the new model significantly outperformed the precision oncology approaches that use the genomic status of drug targets as therapeutic indicators—the researchers acknowledged that further refinements will be needed before they bring their model to the clinical environment. Still, the team emphasized that some version of their approach could “significantly broaden the scope of precision oncology beyond targeted therapies, benefiting an expanded proportion of cancer patients.”²¹

Identifying high-risk individuals and managing population health—

- Research has shown that sepsis contributes to up to half of hospital deaths in the U.S.²² and that mortality rates increase by 8% every hour a sepsis patient goes untreated.²³ Therefore, identifying sepsis as early as possible could lead to a reduction in mortality. At the University of Pennsylvania (Penn), clinician-scientists

developed an AI-powered tool to predict sepsis using machine learning and EHR data from more than 162,000 patients. The algorithm is based on dozens of factors that human clinicians would not be able to research, track, process, and synthesize themselves in a reasonable amount of time. The algorithm was piloted on more than 10,000 patients, and the Penn researchers found that it was able to continuously sample real-time EHR data to prospectively identify patients at risk for developing sepsis so that monitoring and treatment could begin. The next step for this pilot is to examine outcomes to determine whether the tool actually leads to a reduction in mortality and morbidity, or simply identifies those on the path toward high mortality and morbidity.²⁴ At Emory University, clinician-scientists and bioinformatics colleagues developed an algorithm to predict sepsis onset four, eight, and 12 hours before human clinicians would detect it. While there were some false-positives, “no predictor is going to be perfect,” said the lead researcher, Shamim Nemati, and any lead-time in identifying potential sepsis patients is valuable.²⁵ The next step for these researchers is to design a prospective trial to assess the clinical utility of deploying this sepsis prediction model.

- Nearly 300,000 cases of *Clostridium difficile*, a common hospital-acquired infection, occur annually in the U.S. To date, the industry has lacked an effective clinical tool to accurately measure patient risk. As part of a collaborative effort among University of Michigan (UM), Massachusetts Institute of Technology, Harvard Medical School, and affiliated hospitals, a machine learning model has been developed to address this issue. The team evaluated EHR data from 191,014 adult admissions to UM Hospitals and 65,718 admissions at Massachusetts General Hospital (MGH), extracting thousands of features, including patient demographics, admission details, patient history, vital signs, and daily hospitalization information. The model developed shows promise, with at least half of cases predicted correctly at least 5 days in advance of sample collection within both study populations.²⁶ One important outcome of

this partnership study: the algorithm developed at UM did not work on patients at MGH, and vice-versa. There are an increasing number of examples of this phenomenon, underscoring that the data underlying algorithms and the environments in which those algorithms learn and are given input significantly impact how they operate and on which populations they will be effective.²⁷

- Increasingly, social determinants of health (SDOH) are being recognized as significant contributors to health and health outcomes. The CDC acknowledges that “by applying what we know about SDOH, we can not only improve individual and population health but also advance health equity.”²⁸ A team from the University of Tennessee Health Science Center recently studied whether they could use a combination of biological and social conditions to predict the likelihood of a hospital re-visit for pediatric asthma patients after an initial visit. They developed the concept of “sociomarkers,” measurable indicators of social surroundings and conditions for a given person. The team applied machine learning to an integrated dataset containing individual-level biomarker patient information and zip code-level sociomarkers. Their resulting algorithm could predict an asthma re-visit for pediatric patients accurately 66% of the time, a percentage the team said could potentially be improved with more and better data—they only used a 12-month time period and did not have data for pediatric patients who went to hospitals outside of their network. Still, their findings demonstrate the promise of advanced risk prediction tools that could be used to assist in population health management efforts.²⁹
- Suicide rates in the U.S. have risen more than 25% since 1999, with some states like North Dakota seeing increases of more than 50%.³⁰ While suicide is often linked to mental illness, only half of those who die by suicide have a known diagnosed mental health condition,³¹ and suicide is often attributed to a multitude of factors, making it difficult to predict. Researchers at Vanderbilt University Medical Center (VUMC) developed an algorithm based on EHR

data to predict the likelihood that a particular patient would attempt suicide in the next week and in the next two years. As lead researcher Colin Walsh stated, “If an AI [tool] with an 80% prediction success rate is used to assess risk of suicide every time a patient comes into contact with medical care, then, in theory, we would be able to predict, treat, and hopefully prevent far more suicide attempts.”³² In a trial using data from 5,000 patients who were admitted to VUMC for a suicide attempt or self-harm, the algorithm was 85% accurate in predicting the likelihood of suicide attempt within a week and 80% accurate in the two-year prediction.

- End-of-life conversations are best had before death is imminent and when introduced by an experienced clinician such as one specializing in palliative care. But predicting the timing of death is difficult when it isn’t looming in the next few days. At Stanford Medicine, Nigam Shah, an associate professor of biomedical informatics, developed an algorithm that predicts the likelihood of patients dying within the next three to 12 months. Palliative care clinicians at Stanford University Medical Center now receive an email every morning with patient record numbers for those whom the algorithm has identified as having a 90% or higher probability of dying within the next three to 12 months, and the clinicians decide which patients on the list will receive a visit from their team. The palliative care staff also respond to clinician referrals over the course of the day but can now identify patients with longer-range mortality likelihood so that appropriate end-of-life conversations can be planned and conducted. The patients identified by the algorithm and selected by the palliative care team are flagged and monitored by Shah’s team so that the algorithm’s accuracy can be tracked over time, and so that tweaks may be made to further improve the model.³³

Population health, health management, and wellness—

- Patients and their families are also crucial in achieving positive outcomes and preventing disease. Machine learning has been incorporated into the development of many Internet and mobile-

based apps that help patients and their caregivers with pre-procedure preparation, recovery from an acute episode or surgery, medication adherence, and general health management. One example is an app which Arkansas Surgical Hospital recently began offering as an option to help patients before and after joint replacement surgery. The system, called PeerWell, includes hundreds of patient programs that are personalized using machine learning and are proven to improve surgery results and accelerate the recovery process.³⁴

- Another example in the health management and wellness space is *higi*, the largest network of consumer-centric health-assessment and biometric stations in the U.S., with more than 11,000 FDA-cleared stations in drugstores and other retail and corporate settings, as well as more than 80 integrated health devices and applications. Consumers can measure key health metrics at physical kiosks and track their progress via kiosk, a secure website, or smartphone app over time, promoting increased patient engagement. Connected health care clinicians can also monitor a patient’s biometric data and progress. In 2017, *higi* announced a partnership with *Interpreta*. Biometric and patient-reported data from *higi* will be combined with claims, clinical, and genomics data from *Interpreta*. Machine learning techniques will be applied to identify patterns in patient data and outcomes. While it is still in development, the goal is to use the identified patterns to develop a personalized “roadmap” or care plan, to be created and delivered in real time to consumers and their connected clinician(s).³⁵

AI-powered support tools to reduce repetitive clerical work

Health professionals spend hours every day entering or verifying data in EHR systems and other technology platforms. Some of these clerical tasks occur at night and on the weekend, resulting in them being not-so-fondly referred to as “EHR pajama time”; the burden this after-hours work places on clinicians has been linked to burnout.³⁶ New tools are being embedded in EHRs and other hospital tech platforms that reduce or eliminate the time required for clerical tasks. These tools

allow clinicians to spend more time with patients, pursue other professional endeavors, and pursue other personal interests, contributing to greater fulfillment and satisfaction. These tools also improve the accuracy of data entry and reduce clerical work for office staff so they can provide better “customer service” with each patient interaction. Examples of these types of tools and applications include the following:

Improved EHR data and dictation entry—

- Platforms are now available that are powered by automated speech recognition (ASR) capabilities, largely developed with use of machine learning techniques, enabling faster and more accurate data entry and dictation in EHRs, with the potential to also reduce errors and omissions. Examples being used in hospitals today include Nuance Dragon Medical, Dolbey, and Entrada, among others. For a detailed example, see the Nebraska Medicine case study in the sidebar.

Scheduling—

- Optimizing clinical scheduling and ensuring that patients show up for their appointments has long been a major challenge

for most health care organizations. Some are using machine learning-based tools to optimize scheduling. Boston’s Beth Israel Deaconess Medical Center (BIDMC) used machine learning to analyze and optimize the hospital’s operating room schedule, using the results of the analysis to adjust the schedules of 15 surgeons and thereby freeing 30% of the total operating room capacity. A team at BIDMC also developed a tool using machine learning that would identify patients at highest risk for appointment cancellation or no-show and then targeted those patients with reminders, transportation services, or other support to increase the likelihood they would arrive for their appointment.³⁷

Automation tools for back-office functions

Business functions that include repetitive and error-prone tasks are ripe for improvement using AI-powered automation tools. The impact has already been demonstrated in health care institutions and in other industries, with proven

Nebraska Medicine

■ Problem

Nebraska Medicine implemented the Epic EHR in 2009. Physicians became increasingly dissatisfied with the amount of time required to complete the physician notes section in Epic.

■ Solution

The health system investigated voice-recognition software, powered by AI, as a potential solution. Their search identified available technologies that also allowed physicians to use ASR apps on mobile devices for dictation and offered features that would analyze physician notes in real time and alert the physician when the notes suggested that more information was needed for compliance, or when the physician should consider ruling out something or taking a particular action that would potentially improve a patient’s outcome. In the end they implemented Nuance Dragon Medical One.

■ Impact

A physician survey conducted after the new tool had been implemented showed that 94% of physicians believed the tool helped them do their job better; 70% of physicians felt that the quality of documentation had improved; and 50% saved more than 30 minutes per day on physician notes.

■ Caveat

While voice recognition tools have made vast improvements in notes entry, the organization of the information and data integrity issues remain to be solved. Another question is whether EHR systems will allow new technologies and software to be embedded in or connected to their platforms, which would affect their potential value and impact.

cost savings and return on investment, error reduction, and improvement in an organization's capabilities and the ability to repurpose humans formerly performing those often-detested repetitive tasks. Examples include the following:

Claims processing—

- **Robotic process automation (RPA)**, also known as intelligent process automation, applies software to complete repetitive tasks quickly, accurately, and tirelessly. While RPA has been around for decades, its capabilities are expanding as machine learning techniques are applied to develop the software—particularly in the banking, retail, telecommunications, and insurance sectors.³⁸ In a growing number of hospitals and health systems, revenue cycle “bots” (in actuality, an algorithm), are being used for front-end EHR assessment to ensure that patient information entered into the EHR and coding for each claim reflects the clinical assessment; this enables submission of accurate claims that are more likely to be approved and paid at the correct reimbursement level. Similar bots are used to respond to back-end claims denials, resubmission, and follow-up. For a detailed example, see the Ascension Health Agilify sidebar.

Data security and compliance—

- Hospitals and health care organizations are ideal targets for cyber criminals and hackers because they house “immutable data”—data

that doesn't change regularly like an email address or phone number might. Immutable data is particularly valuable for identity theft and insurance reimbursement fraud. AI-based data security and compliance tools can automate complex processes to detect data security violations and react to breaches. These algorithms can track and learn user behavior, identify atypical use patterns, and flag a potential breach. Daniel Nigrin, CIO at Boston Children's Hospital, underscores the importance of transitioning from legacy data security systems and incorporating AI-driven systems: “If we continue to use our approach... of being reactive and only addressing attacks once we have seen them, then we're always going to be one step behind the bad guys... so I think [by] using AI, we can do a better job at being more prospective... starting to be able to detect that anomalous behavior or activity as it's happening... and shut down those attacks before they become a problem.”³⁹

CLINICAL DOMAIN IMPLICATIONS

AI has promising applications for clinicians, health care organizations, and patients. However, there are many issues AHSs should consider when exploring the development or rollout of AI-powered tools.

First, AHSs must acknowledge the potential risks of increasing reliance on technology in clinical

decision-making, intervention, and documentation. Examples include the following:

- **Bias**—AI algorithms perpetuate biases in training datasets, e.g., demographics, practice norms, care setting. The algorithms may not replicate in patient populations and settings with differing characteristics, such as race or ethnicity. In addition, there is a risk that as we augment our own work and thinking with AI-powered machines we may bias what we “know.” If we do not recognize how the machines arrive at their conclusions or where biases exist, those may permeate our own thinking and be difficult to remove.
- **Capabilities**—Some overhyped technologies won't deliver their promised capabilities. If too many examples fall short, there is a risk that clinicians and health care professionals will discount or dismiss technologies that do have real potential.
- **Complacency**—There is a risk that reliance on new tools will become dependence, leading to skill erosion, as clinicians no longer repeatedly exercise their full set of clinical abilities. For an analog in the transportation industry, see the Pilot Automation sidebar.

Ascension Health Agilify

- **Problem**
Many necessary business processes involve mundane, repetitive, time-consuming tasks for employees. These processes can be prone to errors because of the potential number of repeated steps involved and the dull nature of the tasks. These tasks include claims processing—a key function that can materially impact the financial health of the hospital if performed incorrectly.
- **Solution**
Ascension Health's Ministry Service Center subsidiary provides shared services to its member hospitals and also

Pilot Automation

- **Problem**
As airplanes have become increasingly advanced, with digital technology and an increasing number of functions that can be controlled by a computer (advanced autopilot), we are running the risk of pilots' skills eroding as computers take the wheel more and more often. In one tragic example in 2009, Air France's Flight 447 crashed after a thunderstorm struck while in flight, ice crystals formed on the wings, an airspeed sensor stopped functioning, and the autopilot disconnected. The human pilots, fatigued and disengaged from the details of the flight and not having actively piloted the plane in the time leading up to the thunderstorm, were unable to quickly identify the set of problems and correct them. While the stall alarm sounded more than 75 times, the least experienced pilot continued to pull up on his “sidestick”—the opposite of what he should have been doing—but couldn't seem to explain this to either of the other captains. Because of a difference

in the design of this particular plan of which they were unaware, the other captains couldn't figure out what was happening. One exclaimed, “What the hell is happening. I don't understand what's happening... We've totally lost control of the plane, we don't understand at all.”⁴¹

- **Result**
The plane tragically crashed, killing all 228 souls aboard. It was concluded that automation made it more likely that pilots would not face a true crisis while in flight and learn how to handle it. The U.S. Federal Aviation Administration tasked researchers to design a new methodology so pilots could experience, study, and practice examples of aerodynamic stall, as well as other in-flight complications.
- **Solution**
New pilot training programs and continuing education programs have been rolled out and new content will continue to be introduced in 2019.⁴²

- **Control**—There is a risk of losing control of critical health care processes, as ownership of some care decisions shifts from clinicians (though we do not foresee clinicians being removed entirely). Some of the tools will produce recommendations without transparency regarding the underlying logic because of deep learning methodology applied (see definitions on page 3).
- **Compliance**—With few exceptions, no clear, consistent framework for legal and ethical compliance and security exists for some of these AI tools.
Second, AHSs must redesign care models and business processes to unlock and optimize the value potential that accompany these AI tools. As precision medicine approaches are further tested and evidence-based protocols are developed, care processes should be redesigned to consistently incorporate the new tools, and educational materials should be developed for clinicians and patients. The new care processes should be human-centered, with workflows designed to incorporate the new tool(s) in a way that will be understandable, executable, and comfortable for a clinician. When an AHS decides to use augmented intel-

ligence to aid in clinical decision-making, there should be clear protocols around when and how to use it, and what to do if a clinician (or patient) disagrees or distrusts the algorithm's output. If automatic process automation tools work well in one business function, the implementation of those tools in other similar areas should be explored.

Third, AHSs should rethink capital deployment in light of a digitally enabled workforce and digitally enabled consumers. **AHSs should consider investing in research and development of new technologies as well as acquisition of existing products and applications.** AHSs should also understand that patients' and family caregivers' expectations for access to cost/quality/service information are changing, making it increasingly important that clinical services and patient-facing business processes (e.g., billing) are as consumer-friendly, personalized, and digitally supported as possible.

Fourth, AHSs (and industry players developing AI-driven tools for health care) must learn from past examples—both positive and negative—in order to follow emerging best practices and avoid demonstrated pitfalls. For example, in the development of IBM's Watson for Oncology, the computers were being trained on a small amount of "synthetic data" with only one or two Memorial Sloan Kettering Cancer Center (MSKCC) physicians providing recommendations for each type of cancer included in the tool. Using synthetic data is not uncommon in machine learning, but it is likely to be more problematic when it comes to patient data and connecting data to recommendations based on anything other than evidence-based, peer-reviewed, clinically accepted findings. Consequently, when other hospitals tried to use the MSKCC-trained Watson, the software was making treatment recommendations guided by the treatment patterns and preferences of the MSKCC physicians instead of a true AI-based interpretation of actual patient data.⁴³

Finally, the BRAHG strongly believes that AI-based tools in the clinical setting will not replace clinicians but rather augment and supplement available information and enable more efficient care delivery. There is much buzz in the popular press about AI causing drastic changes in health care, such as the elimination of radiologists, but

we believe that such a complete "replacement" is unlikely.

EDUCATION DOMAIN EXAMPLES

The digitization of health care and increasing incorporation of AI-based applications in health care settings necessitate new educational curricula and skill requirements for health care professionals. Additionally, AI will introduce improved ways of educating the current and especially the next generation of health care professionals through applications such as AI-driven virtual patient avatars, virtual reality platforms, and AI-driven testing of surgical skills and performance.

Virtual patient avatars

Virtual patients can be programmed through AI-based software to present signs, symptoms, test results, physical emotion and expression, physical exam findings, and diagnostic and therapeutic algorithms, based on thousands of real patient records. Health care professional students can use these virtual patients to practice patient interactions and diagnosis. The overall cost is significantly less than using real patients or actors, and students can access the tool at their convenience.

Similar to advanced flight simulation software that is being implemented in pilot training to avoid potential disasters like the one described in the sidebar on page 13, health professional schools are increasingly incorporating simulated virtual patients for training purposes, presenting health professionals with virtual patient encounters based on extensive real patient data and guiding students through the patient encounter, diagnosis, and treatment recommendation. As AI-based tools become more present in clinical practice, innovative tools such as those described earlier—some of which are still in the experimental phase—could also be incorporated into such training. This will help clinicians stay on top of their "traditional" training that they would apply without any AI-powered tools, as well as train them to incorporate those tools into their practice without losing their human-based decision-making capabilities.⁴⁴

ImmersiveTouch

■ Problem

Traditional medical education, in particular surgery, relies on human patient examples and is hampered by biased sets of examples by design.

■ Solution

Surgical simulation company ImmersiveTouch, founded in 2005, uses head-mounted displays as well as patient-specific anatomy and tactile feedback, analyzed by AI-driven algorithms, to train surgeons. The company currently offers modules for neurosurgery, ophthalmology, orthopedic

surgery, and otolaryngology (ENT) as well as minimally invasive surgeries.

■ Results

In a recent study, ImmersiveTouch training reduced surgical errors by 54%, compared with a control group using current [traditional] training methods. Dozens of academic medical centers are currently using ImmersiveTouch training, including Johns Hopkins and the University of Chicago.⁴⁵

Virtual reality training platforms

Similar to gaming platforms, surgeons-in-training can practice procedures and learn how to work well with a surgical team using a virtual reality platform. Simulation technology has been in practice for years; the newer AI-based training tools are more detailed, more precise, and more human-like. These tools require fewer resources and less time from both educators and learners and allow educators to adjust and normalize the case study examples. The global market for virtual reality in health care is forecast to reach \$3.8 billion by 2020.⁴⁶ For more detail, see the ImmersiveTouch example in the sidebar.

New methods of testing health care professional performance

Much of current health care professional skill evaluation—outside of written tests—relies on human observation and assessment, which always has a subjective element, often resulting in inconsistency. **Newer AI-based testing applications ensure more consistent testing requirements and thresholds, increased ability to track subtle behaviors or mistakes, and reduced subjectivity.**

■ A team at Wayne State University recently used machine learning models to automatically distinguish between "expert" and "novice" performance of surgeons performing robotic-assisted surgery. **They found that the tool "effectively, objectively, and automatically" evaluated surgeons' performance and were able to design the tool to provide more personalized feedback, available for surgeons' to review online.**⁴⁷

EDUCATION DOMAIN IMPLICATIONS

First, AHSs should embrace and leverage the new AI-powered tools such as those described herein to enhance health professional education programs. **Students will have the opportunity to learn at their own pace in an interactive manner using potentially lower-cost, evidence-based tools.** They also will be able to be tested in a more objective way and receive more personalized feedback.

Second, AHSs will need to teach students how to **use emerging AI-based tools that are being introduced into clinical practice,** such as those described in the Clinical Domain section of this report. While some tools may be experimental at this stage, many have the potential to become part of the recommended "normal" practice, and students will be ill prepared if they have not been exposed to them before beginning their professional practice.

Third, health care professionals should also understand how AI-based tools work. AHSs will need to develop health care-literate data scientists and data-literate health care professionals. AHSs should consider adjusting the curricula and/or adding new courses and requirements to teach students the basics of machine learning and how to work with large datasets. AHSs should also help trainees understand the evolving role AI-driven technologies will play in the care of their patients, arming them with information to help them make their own choices and helping them consider the legal, moral, and ethical issues that accompany the incorporation of algorithm-based tools into clinical practice.

Finally, it may be prudent to rethink the selection criteria for health professional students. The new Carle Clinic College of Medicine holds itself out as the “world’s first engineering-based college of medicine,” with every entering student in the inaugural class having an undergraduate engineering background. While we don’t foresee every health professions school mimicking Carle Clinic’s engineering-based curriculum, we do believe that there should be an emphasis on data science literacy within medical education and that the proportion of students with engineering and data science backgrounds should increase. Health professions schools may also want to explore partnerships, cross-discipline courses, or credit requirements with other schools within or external to their own university. Many engineering schools are building out courses on artificial intelligence, machine learning, and advanced data manipulation. **One school—Massachusetts Institute of Technology—is creating an entirely new college for artificial intelligence, with a planned \$1 billion investment,** including a \$350 million commitment from Stephen Schwarzman of the Blackstone Group.⁴⁸

RESEARCH AND DISCOVERY DOMAIN EXAMPLES

Machine learning and augmented intelligence present new opportunities for discovery and innovation, often with enhanced capabilities and reduction of costs. Harnessing these new tools will help AHSs remain research and discovery leaders who are able to accelerate innovations and trials that improve patients’ health.

Basic research

Basic research can be materially augmented with advanced analytic tools and AI-driven algorithms that can help analyze vast datasets and translate patterns into findings to explore further. Examples of these types of tools and applications include the following:

- Data sharing, modeling, and informatics—AI-supported algorithms can connect many currently separate research databases, help identify patterns in the data, and lead to accelerated research findings. An example is My Intelligent

Machines, which helps scientists find, share, and analyze their “omic” data.⁴⁹

- Accelerated drug discovery—Pharma companies are using AI algorithms, such as Atomwise and IBM Watson Health, to analyze enormous datasets to identify new compounds, which could be translated into potential drugs. The algorithms can also predict how well potential drugs will do in testing, uncover combinations of drugs that might work well together, and find new uses for previously tested compounds.⁵⁰
- AI-based research assistant—A virtual research assistant can comb journal articles based on a set of criteria, curate the most relevant articles, and/or synthesize study findings, helping health care professionals stay on top of current medical literature and best practices with limited amounts of available time and focus their attention on the journal articles most relevant to them.⁵¹

Translational research

Algorithms that offer a promising clinical application may be identified in research and then replicated and tested/validated for targeted subpopulations and settings. Then a clinical or operational intervention trial may be designed to further test and demonstrate the feasibility of using the algorithm, with process changes, decision support, and education. Finally, studies can be designed to measure the effectiveness of introducing the algorithm into clinical care or processes and rolled out as a standard of care if shown to be effective and/or more efficient.

Clinical trials

Finding and securing participants is one of the most time-consuming and expensive parts of a clinical trial. Approximately 2 million patients participate in ~3,000 clinical trials in the U.S. every year. But 6 million patients are needed to reach study recruitment requirements, a factor that delays study timelines and hinders ability to draw conclusions and bring new therapies into practice. AI-driven models can help accelerate the clinical trials process. For a more detailed example, see the sidebar.

Accelerating the Clinical Trials Process

■ Problem

The problem: In 2017, more than 1.7 million people in the U.S. will be diagnosed with cancer (first time), and there are an estimated ~10,000 cancer clinical trials in progress. But our health care system has a limited ability to link these patients with the relevant trials in an efficient and comprehensive way. This is detrimental to patients wanting to access the most cutting-edge care and extends clinical trials by extending the time required to recruit appropriate trial participants. And it is costly—it is estimated that more than 90% of clinical trials are delayed or over budget. Researchers have suggested that 25%-50% of cancer patients should be enrolled in trials. But fewer than 5% of cancer patients, on average, enroll in a clinical trial, often because patients and doctors don’t know what trials are available.

■ Solution

AI-powered algorithms that analyze patient data and match with databases of clinical trials can identify eligible patients for clinical trials in minutes.

■ Results

In an early comparison participant search, a principal investigator used a traditional recruitment method to identify and validate 23 eligible patients for a biomarker/non-small cell lung cancer trial over the course of six months. An AI-powered software tool, Deep 6, found and validated 58 eligible matches in less than 10 minutes. The key to making this application work is data sharing among health care institutions.⁵²

RESEARCH AND DISCOVERY DOMAIN IMPLICATIONS

First, AHSs must invest in core resources required to support AI-powered tools that will augment their research and discovery efforts. That includes advanced data storage and manipulation software, tools that enable secure data sharing, clinical trial subject sourcing tools, and personnel who have advanced skills in health care data informatics and a thorough knowledge of machine learning.

Second, AHSs should evolve their approach to collaboration and data sharing. **We believe many of the next big discoveries will be based on very large datasets, which no single institution owns.**

Pharmaceutical companies have and use AI-based tools on enormous datasets—we don’t need to reinvent the wheel; we can form partnerships to access these capabilities. That being said, although partnerships and collaboration will be important, relationships should be structured carefully to protect the data, preserve patient interests, and ensure that proper recognition (if relevant) of any commercialization rights is appropriately assigned.

Finally, the funding for research may shift. Currently, the NIH and the pharmaceutical industry dominate the research funding pipeline at most AHSs. As AI-powered innovations continue

to develop, advocacy efforts toward the government may need to emphasize AI as a worthy area of investment. Furthermore, as patient data is increasingly available through more public (but secure) channels, AHSs and other health care research institutions may be able to pursue research endeavors with data from non-traditional channels.

IV. A Call to Action for Academic Health Systems

Several overarching implications emerged as BRAHG explored examples of AI applications associated with the AHS’s tripartite mission as described earlier. For AHSs, there will be an increasing need to do the following:

- Invest in infrastructure and new competencies;
- Promote data literacy and cross-disciplinary collaboration;
- Pursue and embrace partnerships with other AHSs, health systems, and industry leaders in AI to share data, coordinate data integrity, launch research and development efforts, and disseminate findings and new solutions; and,
- Assure that the interests and welfare of patients and their caregivers are at the forefront.

To that end, we recommend that AHS leaders take the following actions to help deliberately advance the development and deployment of AI at their institutions and beyond. We also suggest that leaders prepare their governing boards for these priorities and investments and present a clear rationale behind them.

- Become knowledgeable about this topic to help bridge the gap between reality and hype. It is incumbent on AHS leaders to do so, as faculty and learners are increasingly knowledgeable about this topic.
- Invest time to understand where AI is headed and what may already be occurring at your own institution, so you can help shape its development and direct resources appropriately. **Develop a thoughtful AI strategy to align resources and optimize investments based on your institution's strategic priorities, competencies, and risk profile. Investments in this area—particularly in development—come at a high cost.** Some investments will likely come to fruition, but others will not be as successful. In this developing area, due diligence and careful examination of forecasted return on investment and potential risks are imperative.
- Invest in the researchers, data scientists, and biomedical informatics scientists as well as the governance infrastructure to support this growing domain of science. Many AHSs have already launched new centers for data science to find/develop scholars needed for this activity, which is different from the mechanics of running IT infrastructure and operations.
- Promote data literacy across disciplines and foster collaboration across domains of knowledge to identify opportunities, evaluate the applicability of those opportunities, and effectively conduct research and development. A faculty member at a leading academic medical center observed that there are not enough “bilingual” innovators who both understand the practice of medicine and are experts in engineering/data science. He feels that without a deep understanding of each language, collaboration efforts sometimes fall short or move too slowly.
- Ensure that educational programs **embed data literacy** to prepare current and future health care

professionals for digitally enabled care models.

- Develop your own institutional framework for pursuing partnerships that focus on value creation. Given the scale of data required to develop AI technologies, AHSs will most certainly require partnerships to succeed—both across AHSs and with industry. When working with pharmaceutical companies and device manufacturers, understand that many AI/ML tools contain intellectual property and/or have a for-profit entity involved. Careful diligence is required to pursue relationships that will make strategic sense and to identify those that are most likely to succeed. In collaboration with profit-oriented businesses, educating those involved in the partnership about potential conflicts in decision-making around patient care is also crucial.
- Come together with other AHSs to create mechanisms and guidelines for data sharing, data aggregation, testing, and best practices. Approach data sharing and “crowd-sourcing” of data carefully—many large datasets, including the data contained in most of our EHRs, is not curated. Developing algorithms using poor data will only result in poor algorithms and ineffective or potentially harmful tools. Coming together in some way to promote data curation will be paramount in our progress toward the development of effective AI-driven clinical tools.
- Work collaboratively with other AHSs to develop industry partnerships in a principled manner: AHSs should lead the charge in educating the industry on what we need and on standardizing approaches to ensure fair participation in value creation.
- Uphold the standards AHSs have applied to all new treatments and clinical process: those of the human clinical trial. Amidst the hype around the potential that AI/ML brings to health care, many AI/ML solutions are being rushed to implementation through in silico testing (testing by computer). These tools merit prospective, randomized clinical trials to ensure efficacy and identify risks. It is important to remember that insurers may be inclined to push the deployment of these new tools, as they are familiar with the efficiency benefits and savings

that come from AI-driven back-office tools. In clinical settings, we should require reasonable evidence that an AI/ML tool brings clinical and not just financial advantages.

- Adopt and advance an ethical framework to guide this transformation to the benefit of all. Known ethical challenges already exist. As AI technology advances, we will likely uncover new issues. As leaders in the field, we will have to be alert to thoughtfully address these issues as they arise.

AHSs have always played an important role in leading the transformation of health care. Our strength is biomedical research, and the associated resources required do not necessarily

change the care delivery paradigm. Artificial and augmented intelligence requires completely new skill sets and technologies, which can and will impact the way that we conduct research, partner on research, source funding, teach, and deliver care. It therefore will be difficult to embrace this new area, and it isn't guaranteed that we will be leaders, though the opportunity certainly exists. Given our shared tripartite mission—discovery and innovation, learning labs for future clinicians and health care leaders, and provision of care to communities across the country and world—it is incumbent upon AHS leaders to be thoughtful in helping shape this next transformation in health care.

References

1. Blue Ridge Academic Health Group. *The Hidden Epidemic: The Moral Imperative for Academic Health Centers to Address Health Professionals' Well-Being*. Atlanta: Emory University; 2018. Available at <http://whsc.emory.edu/blueridge/publications/archive/blue-ridge-winter2017-2018.pdf>. Accessed January 2, 2019.
2. Ahmed M, Toor A, O'Neil K, Friedland D. Cognitive computing and the future of health care. *IEEE Pulse*. May/June 2017. Available at <https://pulse.embs.org/may-2017/cognitive-computing-and-the-future-of-health-care/>. Accessed January 2, 2018.
3. Collier M, Fu R, Yin L. Artificial intelligence: Healthcare's new nervous system. Accenture. December 15, 2017. Available at https://www.accenture.com/t20171215T032059Z_w_/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf. Accessed January 17, 2019.
4. Sullivan T. Half of hospitals to adopt artificial intelligence within 5 years. *Healthcare IT News*. April 11, 2017. Available at <https://www.healthcareitnews.com/news/half-hospitals-adopt-artificial-intelligence-within-5-years>. Accessed January 2, 2018.
5. Chui M, Manyika J, Miremadi M. What AI can and can't do (yet) for your business. *McKinsey Quarterly*. January 2018. Available at <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business>. Accessed January 4, 2019.
6. RodriguezRamos J. Brains vs. computers. Medium. March 12, 2018. Available at <https://becominghuman.ai/brains-vs-computers-f769548010f1>. Accessed December 4, 2018.
7. Institute of Electrical and Electronic Engineers. Available at <https://standards.ieee.org/industry-connections/ec/autonomous-systems.html>. Accessed January 9, 2019.
8. ACR Assist. ACR Informatics. Available at <http://www.acrinformatics.org/acr-assist>. Accessed December 4, 2018.
9. ACR Assist empowers radiologists to efficiently provide value-based imaging. *Imaging Technology News*. November 25, 2015. Available at <https://www.itnonline.com/content/acr-assist-empowers-radiologists-efficiently-provide-value-based-imaging>. Accessed December 4, 2018.
10. Cruz-Roa A, Gilmore H, Basavanahally A, Feldman M, Ganesan S, Shih N, Tomaszewski J, Ganzalez F, Madabhushi A. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Report (Nature)*. 2017;7:46450. Available at <https://www.semanticscholar.org/insights/what-ai-can-and-cant-do-yet-for-your-business>.

- org/paper/Accurate-and-reproducible-invasive-breast-cancer-in-Cruz-Roa-Gilmore/d5b91f292c611dea61f6e95b007ae53c2766a5f9. Accessed January 4, 2019.
11. Watch out for diabetic retinopathy. *CDC Features*. Centers for Disease Control and Prevention. Available at <https://www.cdc.gov/features/diabetic-retinopathy/index.html>. Accessed December 4, 2018.
 12. People with diabetes can prevent vision loss. National Eye Health Education Program, National Eye Institute. 2016. Available at https://nei.nih.gov/sites/default/files/neh-pdfs/2016_NDM_Article_FINAL_508.pdf. Accessed December 4, 2018.
 13. Murchison A, Hark L, Pizzi L, Dai Y, Mayro E, Storey P, Leiby B, Haller J. Non-adherence to eye care in people with diabetes. *BMJ Open Diabetes Research & Care*. 2017;5(1):3000333. Available at <https://www.ncbi.nlm.nih.gov/pubmed/28878930>. Accessed January 4, 2019.
 14. Sixty percent of Americans with diabetes skip annual sight-saving exams. American Academy of Ophthalmology. October 20, 2016. <https://www.aao.org/newsroom/news-releases/detail/sixty-percent-americans-with-diabetes-skip-exams>. Accessed January 16, 2019.
 15. Quellec G, Abramoff M. Estimating maximal measurable performance for automated decision systems from the characteristics of the reference standard application to diabetic retinopathy screening. *Conference Proceedings, IEEE Engineering in Medicine and Biology Society*. August 2014. Available at https://www.researchgate.net/publication/270659182_Estimating_maximal_measurable_performance_for_automated_decision_systems_from_the_characteristics_of_the_reference_standard_application_to_diabetic_retinopathy_screening. Accessed January 4, 2019.
 16. Abramoff M, Lavin P, Birch M, Shah N, Folk J. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *npj Digital Medicine (Nature)*. 2018;1:39. Available at https://www.researchgate.net/publication/327271235_Pivotal_trial_of_an_autonomous_AI-based_diagnostic_system_for_detection_of_diabetic_retinopathy_in_primary_care_offices. Accessed January 4, 2019.
 17. John M. Artificial intelligence launches a new era in cardiac care. *Peter Munk Cardiac Center Magazine (The Globe and Mail)*. November 2017. Available at https://www.uhn.ca/PMCC/media/Globe_Mail/Pages/artificial-intelligence-launches-new-era-in-cardiac-care-2017.aspx. Accessed January 4, 2019.
 18. Coronary artery disease learning and algorithm development (CADLAD) clinical study. Analytics 4 Life. Available at <https://www.analytics4life.com/clinical-study/>. Accessed December 4, 2018.
 19. What is precision medicine? National Institutes of Health, U.S. National Library of Medicine. Available at <https://ghr.nlm.nih.gov/primer/precisionmedicine/definition>. Accessed December 4, 2018.
 20. 2bPrecise, an Allscripts company, launches a genomics and precision medicine Initiative at National Institutes of Health. Allscripts. August 8, 2016. Available at https://2bprecisehealth.com/wp-content/uploads/2017/01/News-Release_Allscripts-NIH-Genomics-Initiative_FINAL.pdf. Accessed December 4, 2018.
 21. Ding M, Chen L, Cooper G, Young J, Lu X. Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*. 2018;16(2):269-278. Available at <https://www.ncbi.nlm.nih.gov/pubmed/29133589>. Accessed January 4, 2019.
 22. Liu V, Escobar G, Greene J, Soule J, Whippy A, Angus D, Iwashyna T. Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*. 2014;312(1):90-92. Available at https://pdfs.semanticscholar.org/2e7c/f2c7ae32c88dc14838991e1570d3fa4be0da.pdf?_ga=2.231972058.41750408.1546632281-1989338097.1546632281. Accessed January 4, 2019.
 23. Kumar A, Roberts D, Wood K, Light B, Parrillo J, Sharma S, Suppes R, Feinstein D, Zanotti S, Tailberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med*. 2006;34(6):1589-96. Available at <http://www.ccmpitt.com/ebm/sepsis/Kumar%20A,%20et%20al.%20%20Duration%20of%20hypotension%20before%20initiation%20o.pdf>. Accessed January 4, 2019.
 24. Giannini H, Chivers C, Draugelis M, Hanish A, Fuchs B, Donnelly P, Lynch M, Meadows L, Parker SJ, Schweikert WD, Mikkelsen, ME, Fishman N, Hansen CW, Umscheid C. Development and implementation of a machine-learning algorithm for early identification of sepsis in a multi-hospital academic healthcare system. Abstracts from the American Thoracic Society International Conference. May 2017. Available at https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2017.195.1_MeetingAbstracts.A7015. Accessed January 4, 2019.
 25. Nemati S, Holder A, Razmi R, Stanley M, Clifford G, Buchman T. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547-553. Available at <https://www.ncbi.nlm.nih.gov/pubmed/29286945>. Accessed January 10, 2019.
 26. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, Washer L, West LR, Young VB, Guttig J, Hooper DC, Shenoy ES, Wiens JA. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infection Control & Hospital Epidemiology*. 2018;39(4):425-433. Available at <https://www.ncbi.nlm.nih.gov/pubmed/29576042>. Accessed January 9, 2019.
 27. Weins J, Shenoy E. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*. 2018;66(1):149-153. Available at <https://www.ncbi.nlm.nih.gov/pubmed/29020316>. Accessed January 10, 2019.
 28. Social determinants of health: Know what affects health. Social Determinants of Health, Centers for Disease Control and Prevention. Available at <https://www.cdc.gov/socialdeterminants/>. Accessed December 4, 2018.
 29. Shin EK, Mahajan R, Akbilgic O, Shaban-Nejad A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *npj Digital Medicine (Nature)*. 2018;1:50 (2018). Available at https://www.researchgate.net/publication/328028243_Sociomarkers_and_biomarkers_predictive_modeling_in_identifying_pediatric_asthma_patients_at_risk_of_hospital_revisits. Accessed January 10, 2019.
 30. Suicide rising across the US. *Vital Signs*. Centers for Disease Control and Prevention. Available at <https://www.cdc.gov/vitalsigns/suicide/infographic.html>. Accessed December 4, 2018.
 31. Suicide rates rising across the U.S. CDC Newsroom. Centers for Disease Control and Prevention. Available at <https://www.cdc.gov/media/releases/2018/p0607-suicide-prevention.html>. Accessed December 4, 2018.
 32. Goldhill O. Prescription AI: Machines know when someone's about to attempt suicide. How should we use that information? Quartz. September 5, 2018. Available at <https://qz.com/1367197/machines-know-when-someones-about-to-attempt-suicide-how-should-we-use-that-information/>. Accessed December 4, 2018.
 33. Newby K. Compassionate intelligence: Can machine learning bring more humanity to healthcare? *Stanford Medicine*. Summer 2018. Available at <http://stanmed.stanford.edu/2018summer/artificial-intelligence-puts-humanity-health-care.html>. Accessed December 4, 2018.
 34. Top Arkansas joint replacement center invests in technology to engage patients. OrthoFeed. April 20, 2017. Available at <https://orthofeed.com/2017/04/20/top-arkansas-joint-replacement-center-invests-in-technology-to-engage-patients/>. Accessed December 4, 2018.
 35. Philippidis A. Interpreta and high eye healthcare 'roadmaps' using patient data collected at grocers, pharmacies. *Clinical OMICS*. September 28, 2017. Available at <https://www.clinicalomics.com/articles/interpreta-and-high-eye-healthcare-roadmaps-using-patient-data-collected-at-grocers-pharmacies/1262>. Accessed December 4, 2018.
 36. Arndt B, Beasley J, Watkinson M, Temte J, Tuan WJ, Sinsky C, Gilchrist V. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations. *Annals of Family Medicine*. 2017;15(5):419-426. Available at <http://www.annfammed.org/content/15/5/419.full>. Accessed January 10, 2019.

37. Rath D. One CIO's view of the path to precision medicine. *Healthcare Informatics* blog. May 10, 2018. Available at <https://www.healthcare-informatics.com/blogs/david-raths/innovation/one-cio-s-view-path-precision-medicine>. Accessed December 4, 2018.

38. Kumar S. Robotic process automation across industries. *Digitalist*. March 8, 2018. Available at <https://www.digitalistmag.com/digital-economy/2018/03/08/robotic-process-automation-across-industries-05955888>. Accessed December 4, 2018.

39. Faggella D. AI in healthcare IT security—Why hospitals are targets. *emerj*. October 8, 2017. Available at <https://emerj.com/ai-podcast-interviews/ai-healthcare-security-hospitals-targets/>. Accessed December 4, 2018.

40. Ascension debuts subsidiary to help other companies succeed through process automation. *Business Wire*. June 18, 2018. Available at <https://www.businesswire.com/news/home/20180618006054/en/Ascension-Debuts-Subsidiary-Companies-Succeed-Process-Automation>. Accessed December 4, 2018.

41. Air France Flight 447's lessons—four years later. CBS News Online. June 1, 2013. Available at <https://www.cbsnews.com/news/air-france-flight-447s-lessons-four-years-later/>. Accessed December 4, 2018.

42. Preventing a plane crash—Research helps pilots train for aerodynamic stalls. *Phys Org*. June 8, 2018. Available at <https://phys.org/news/2018-06-plane-crashresearch-aerodynamic-stalls.html>. Accessed December 4, 2018.

43. Ross C, Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *STAT*. July 25, 2018. Available at https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/?utm_source=STAT+Newsletters&utm_campaign=beb06f048d-MR_COPY_08&utm_medium=email&utm_term=0_8cab1d7961-beb06f048d-150085821. Accessed December 4, 2018.

44. Wu N. Simulation improves medical education—one case at a time. *Healthcare Innovation News*. 2015;2(1):1-2. Available at <https://www.i-human.com/wp-content/uploads/2015/01/Healthcare-Innovation-News-Jan-2015-Wu.pdf>. Accessed January 10, 2019.

45. Fink C. How VR saves lives in the OR. Immersive Touch. April 23, 2018. Available at <https://www.immersivetouch.com/press-archive/2018/4/23/how-vr-saves-lives-in-the-or>. Accessed December 4, 2018.

46. The Top 7 virtual reality digital health companies. Medtech Boston. February 26, 2018. Available at <https://medtechboston.medstro.com/blog/2018/02/26/the-top-6-virtual-reality-digital-health-companies/>. Accessed December 4, 2018.

47. Fard M, Ameri S, Chinnam R, Pandya A, Klein M, Ellis RD. Machine learning approach for skill evaluation in robotic-assisted surgery. *Proceedings of the World Congress on Engineering and Computer Science*. October 2016. Available at <https://arxiv.org/pdf/1611.05136.pdf>. Accessed January 10, 2019.

48. Lohr S. M.I.T. plans college for artificial intelligence, backed by \$1 billion. *The New York Times*, October 15, 2018. Available at <https://www.nytimes.com/2018/10/15/technology/mit-college-artificial-intelligence.html>. Accessed December 4, 2018.

49. Jenna S. How artificial intelligence will make basic research self-sustainable. TEDx Concordia. December 19, 2017. Available at https://www.youtube.com/watch?reload=9&v=tPb_IqoKNDY. Accessed December 4, 2018.

50. Accelerating research with AI & Watson for drug discovery. IBM. Available at <https://www.ibm.com/blogs/insights-on-business/healthcare/accelerating-research-ai-watson-drug-discovery/>. Accessed December 4, 2018.

51. Milmo C. AI system will trawl through millions of academic journals to find links missed by scientists. *Independent*. November 3, 2015. Available at <https://www.independent.co.uk/news/science/ai-system-will-trawl-through-millions-of-academic-journals-to-find-links-missed-by-scientists-a6719876.html>. Accessed December 4, 2018.

52. Bookbinder M. The Intelligent trial: AI comes to clinical trials. *Clinical Informatics News*. September 29, 2017. Available at <http://www.clinicalinformaticsnews.com/2017/09/29/the-intelligent-trial-ai-comes-to-clinical-trials.aspx>. Accessed December 4, 2018.

Appendix

Opening vignette, *continued from page 4*

The tumor board is convened, including Dr. Henry Ramsey in otolaryngology, Dr. Maria Moreau in medical oncology, and Dr. Steven Wilson in radiation oncology, as well as several other clinicians.

HR: "This is a stage 1 T2N1M0 p16 positive squamous cell carcinoma of the tonsil."

The case is reviewed by all present, and a therapeutic approach is recommended based on the clinical picture, as well as on Peggy's summary of the relevant scientific literature and her review of available clinical trials that might be relevant.

HR: "The treatment options include transoral robotic resection of the tumor with neck dissection followed by post-operative radiation therapy or concurrent chemotherapy/radiation therapy to standard dosing. Mr. Burdell has expressed interest in proceeding with surgery and if possible would like to avoid chemotherapy."

Various comments and gestures of agreement come from members of the tumor board.

Most of those present have their laptops open during this discussion. It is easy to spot their cognitive assistant icons on the upper left of their screens. Etiquette for interactions of clinicians and cognitive assistants has emerged over the few years that this technology has been available.

First, during meetings clinicians can query their assistants by typing but not by talking. Second, assistants cannot communicate with each other during meetings. Before adopting these rules, meetings had become rather chaotic, with everybody talking at once but not to each other. Communications among assistants often resulted in a second meeting being conducted below the surface.

The meeting concludes by all agreeing on a summary of the tumor board recommendations, which are sent to primary care physician, Dr. Edward Cummings.

Drs. Ramsey, Moreau, and Wilson meet with the patient, Mr. George Burdell and his older sister, Agnes Hunsinger. They walk through the treatment options, i.e., transoral robotic surgery versus concurrent chemoradiation.

GB: "What is the surgical procedure?"

HR: "We use a "robot" to remove the tumor in the tonsil. We have to remove the enlarged cancer lymph node from your neck, including the surrounding lymph nodes as well."

GB: "Is there a name for the procedure?"

HR: "It's is called TORS with selective neck dissection."

GB: "Dissection! Sounds terrible. Like high school biology."

HR: "I know, but, medically speaking, it just means the surgical removal of tissues which are diseased."

AH: "I am going to be helping George with the practical issues associated with all this. What will that include?"

HR: "Well, for the surgery we would perform the neck dissection and then one week later return to the OR and remove the tumor using the robot."

He hands Ms. Hunsinger a summary sheet.

SW: "We would then wait at least four to six weeks and then start your post-operative radiation."

GB: "Well, what about the other options?"

MM: "Well, instead of surgery we would give you a higher dose of radiation and also a dose of chemotherapy to make the tumor respond better."

GB: "Why would I choose one over the other?"

AH: "This seems complicated, and George has never been that good at being organized."

GB: "You always say that Agnes. I am not that bad."

HR: "Do you have a cell phone, Mr. Burdell?"

GB: "Sure, an iPhone actually."

HR: "Great, we are going to provide you an app called HealthHelp that will assist you with all this."

GB: "Like the app that keeps track of how much I walk?"

HR: "Something like that. It will know about all your appointments and prescriptions. It will send you text messages to remind you. It can review what we discuss today to help you make a decision or understand the treatment options better. You can ask it questions. You can even give it a name."

GB: "I think Agnes would be a good name."

HR: "You can provide your sister Agnes with access to HealthHelp so she can keep track of things with you and provide help when you need it."

GB: "Then I think I will change the name to Ruth, who was our mother. I wouldn't want to be confused by two Agneses."

The conversation then switches back to the benefits and risks of the various treatment options. Mr. Burdell ultimately decides to proceed with TORS with neck dissection followed by radiation.

GB: "Do you interact with HealthHelp as well?"

SW: "Yes, we all do. You will receive communications from your whole cancer team, including Drs. Ramsey, Moreau, and me, as well as several other people you will meet along the way, including your speech pathologist, nutritionist, and social worker."

GB: "And, I can ask questions of anybody?"

SW: "Yes, of course. You might find it interesting that we keep track of all questions that patients ask, as well as all the answers. This has enabled us to continually improve patients' experiences."

GB: "I guess that I better be careful of what I ask."

SW: "We don't keep track of who asked what questions. Your privacy is assured."

GB: "That's nice to know."

Following the completion of all planned therapy, Drs. Ramsey, Moreau, and Wilson follow Mr. Burdell on a

three-month basis for signs and symptoms of recurrence. Dr. Cummings sees Mr. Burdell for routine checkups for any other chronic disease.

Millie: "Mr. Burdell has become a model patient. He schedules checkups and takes his prescriptions, quite unlike his previous behaviors."

EC: "Seems like HealthHelp changed his life. His sister, Agnes, even sent me an email about that."

Millie: "I wonder if it isn't the combination of a serious health scare and the availability of HealthHelp?"

EC: "Well, Millie, why don't you research that question and let me know what you find?"

Millie: "I already have. That's why I brought it up."

About the Blue Ridge Academic Health Group

The Blue Ridge Academic Health Group studies and reports on issues of fundamental importance to improving the health of the nation and its health care system and enhancing the ability of the academic health center (AHC) to sustain progress in health and health care through research—both basic and applied—and health professional education. In 22 previous reports, the Blue Ridge Group has sought to provide guidance to AHCs on a range of critical issues. (See titles, page 26.)

For more information and to download free copies of these reports, please visit www.whsc.emory.edu/blueridge.

Previous Blue Ridge Reports

See <http://whsc.emory.edu/blueridge/publications/reports.html>.

Report 22: *The Hidden Epidemic: The Moral Imperative for Academic Health Centers to Address Health Professionals' Well-Being*. 2018.

Report 21: *The Academic Health Center: Delivery System Design in the Changing Health Care Ecosystem—Sizing the Clinical Enterprise to Support the Academic Mission*. 2017.

Report 20: *Synchronizing the Academic Health Center Clinical Enterprise and Education Mission in Changing Environments*. 2016.

Report 19: *Refocusing the Research Enterprise in a Changing Health Ecosystem*. 2015

Report 18: *A Call to Lead: The Case for Accelerating Academic Health Center Transformation*. 2014

Report 17: *Health Professions Education: Accelerating Innovation Through Technology*. 2013.

Report 16: *Academic Health Center Change and Innovation Management in the Era of Accountable Care*. 2012.

Report 15: *The Affordable Care Act of 2010: The Challenge for Academic Health Centers in Driving and Implementing Health Care Reform*. 2012.

Report 14: *The Role of Academic Health Centers in Addressing the Social Determinants of Health*. 2010.

Report 13: *Policy Proposal: A United States Health Board*. 2008.

Report 12: *Advancing Value in Health Care: The Emerging Transformational Role of Informatics*. 2008.

Report 11: *Health Care Quality and Safety in the Academic Health Center*. 2007.

Report 10: *Managing Conflict of Interest in AHCs to Assure Healthy Industrial and Societal Relationships*. 2006.

Report 9: *Getting the Physician Right: Exceptional Health Professionalism for a New Era*. 2005.

Report 8: *Converging on Consensus? Planning the Future of Health and Health Care*. 2004.

Report 7: *Reforming Medical Education: Urgent Priority for the Academic Health Center in the New Century*. 2003.

Report 6: *Creating a Value-driven Culture and Organization in the Academic Health Center*. 2001.

Report 5: *e-Health and the Academic Health Center in a Value-Driven Health Care System*. 2001.

Report 4: *In Pursuit of Greater Value: Stronger Leadership in and by Academic Health Centers*. 2000.

Report 3: *Into the 21st Century: Academic Health Centers as Knowledge Leaders*. 2000.

Report 2: *Academic Health Centers: Good Health Is Good Business*. 1998.

Report 1: *Academic Health Centers: Getting Down to Business*. 1998.



EMORY
UNIVERSITY

Woodruff Health
Sciences Center

D

AI in Financial Services

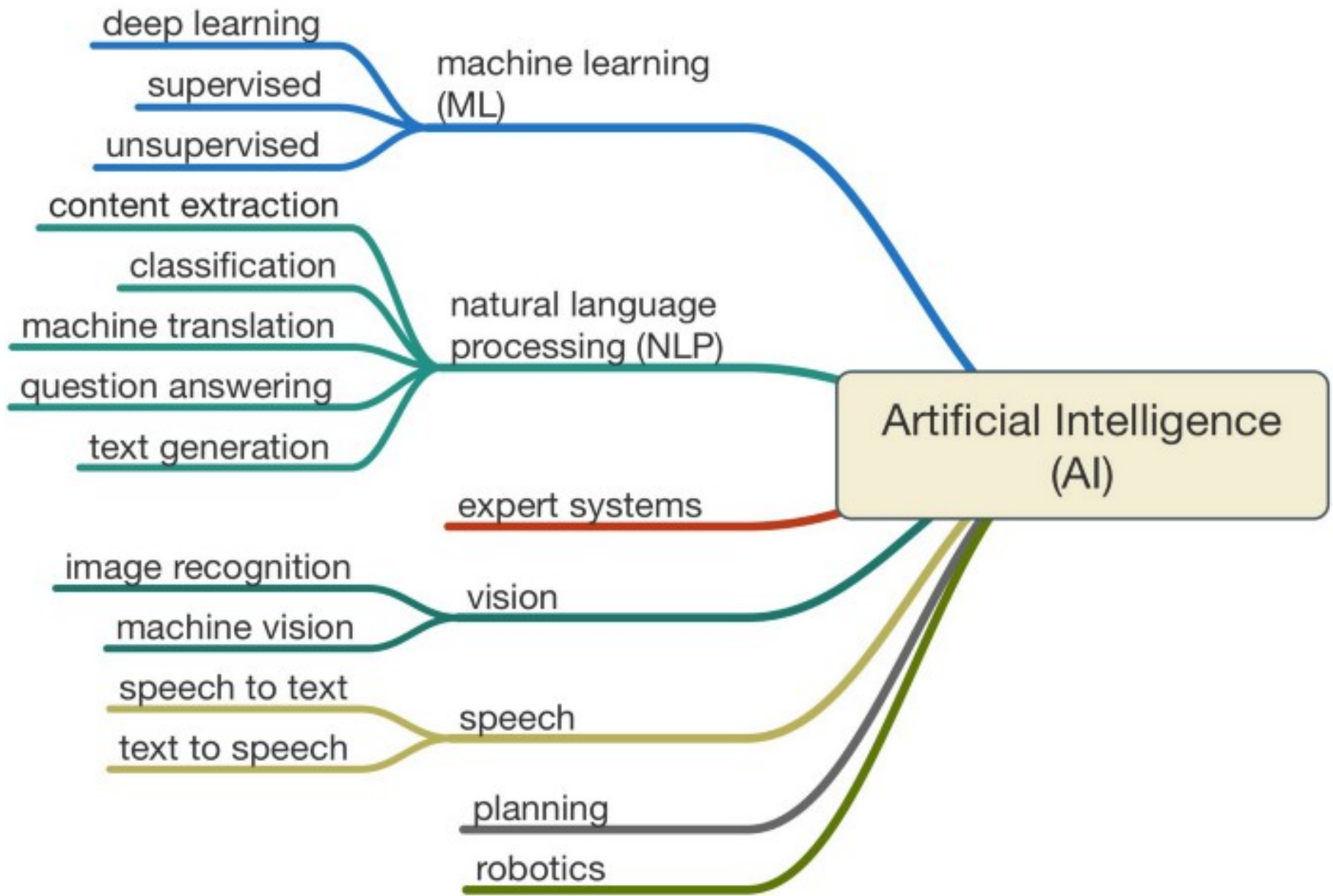
Artificial Intelligence in Financial Services

Ted Claypoole
Partner



What is AI?





AI Terminology

Artificial Intelligence:

- The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.
- “The study . . . on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”
- “The field of computer science dedicated to solving cognitive problems commonly associated with human intelligence, such as learning, problem solving, and pattern recognition.”
- The study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.



AI Terminology

- **Narrow AI**: Technology outperforming humans in a narrow field of endeavor, i.e. chess or weather prediction.
- **General AI**: Technology able to perform any intellectual task that humans can perform.
- **Machine Learning**: a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
- **Training Dataset**: The data used to train an algorithm, distinguished from the testing dataset or validation dataset used to test that the algorithm has been properly trained. The training dataset is used at the beginning of the AI creation process to help a program understand how to apply technologies like neural networks to learn and produce sophisticated results.



AI Terminology

- **Unsupervised Learning**: is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. Most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. This method helps find previously unknown patterns in data sets without pre-existing labels. Also known as self-organization and allows modeling probability densities of inputs.
- **Supervised Learning**: Task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. When the model is getting trained on a labelled dataset. Labelled dataset is one which have both input and output parameters.
- **Machine Perception**: is the ability to use input from sensors (such as cameras (visible spectrum or infrared), microphones, wireless signals, and active lidar, sonar, radar, and tactile sensors) to deduce aspects of the world. Applications include speech recognition, facial recognition, and object recognition.



Anomaly Detection

- AI models can learn what is “normal” for a person, group, system, network, etc., and can adjust for dynamic baselines
- By learning what is normal, these models can very accurately predict when something is “not quite right”
- New systems give the user a way to prioritize by detecting the level of rarity and the potential importance of an anomaly
- Behavioral data models can use attributes like log-in time, email traffic and calendar items to predict unusual behavior
- Heat mapping can also detect deviations from the norm that may point to broader trends (e.g., takeup rates in commission sales)
- KYC solutions can use attributes like mouse movement and click speed along with contextual data to authenticate consumers



How is AI Used in Financial Services?



Federal Reserve Gov. Brainard's Speech on AI

1. Superior ability for pattern recognition
2. Cost efficiencies
3. Greater accuracy
4. Better predictive power
5. Better at accommodating large volumes of unstructured data

Four Areas Where AI Will Impact Banking

1. Customer facing uses (chatbots, robo-advising, credit evaluation)
2. Back office operations (capital optimization, stress testing)
3. Trading and investment strategies
4. Compliance and risk mitigation (fraud screening)

Deloitte Insights for AI in Financial Services

- **Embed AI in strategic plans:** Integrating artificial intelligence (AI) into an organization's strategic objectives has helped many frontrunners develop an enterprisewide strategy for AI that various business segments can follow. The greater strategic importance accorded to AI is also leading to a higher level of investment by these leaders.
- **Apply AI to revenue and customer engagement opportunities:** Most frontrunners have started exploring the use of AI for various revenue enhancements and client experience initiatives and have applied metrics to track their progress.
- **Utilize multiple options for acquiring AI:** Frontrunners seem open to employing multiple approaches for acquiring and developing AI applications. This strategy helps them accelerate the adoption of AI initiatives via access to a wider pool of talent and tech solutions.



How AI is Currently Used in Financial Services

- **Credit Decisions** (Auto lending, Digital banks)
- **Risk Management** (Market Analysis, Much faster analytics)
- **Fraud Prevention**
- **Trading** (Structured and unstructured data, Bloomberg)
- **Personalized Banking** (Apps, Tools, Interfaces)
- **Process Automation** (verifying data, generating reports)

SEC's Use of AI

The SEC recently announced that it is using unsupervised ML on unstructured data to detect red flags that may indicate fraud: “Machine learning algorithms may help our examiners by pointing them in the right direction in their identification of possible fraud or misconduct.”

Sample use case: Investment advisor reporting

- They review unstructured reporting data to predict the presence of idiosyncratic risk
- They have found that this method is 5 times more accurate at predicting risk
- However, they warn that they use staff to carefully analyze red flags for false positives

What Are Legal Issues and How Do They Apply to Financial Services?



Machines Making Decisions



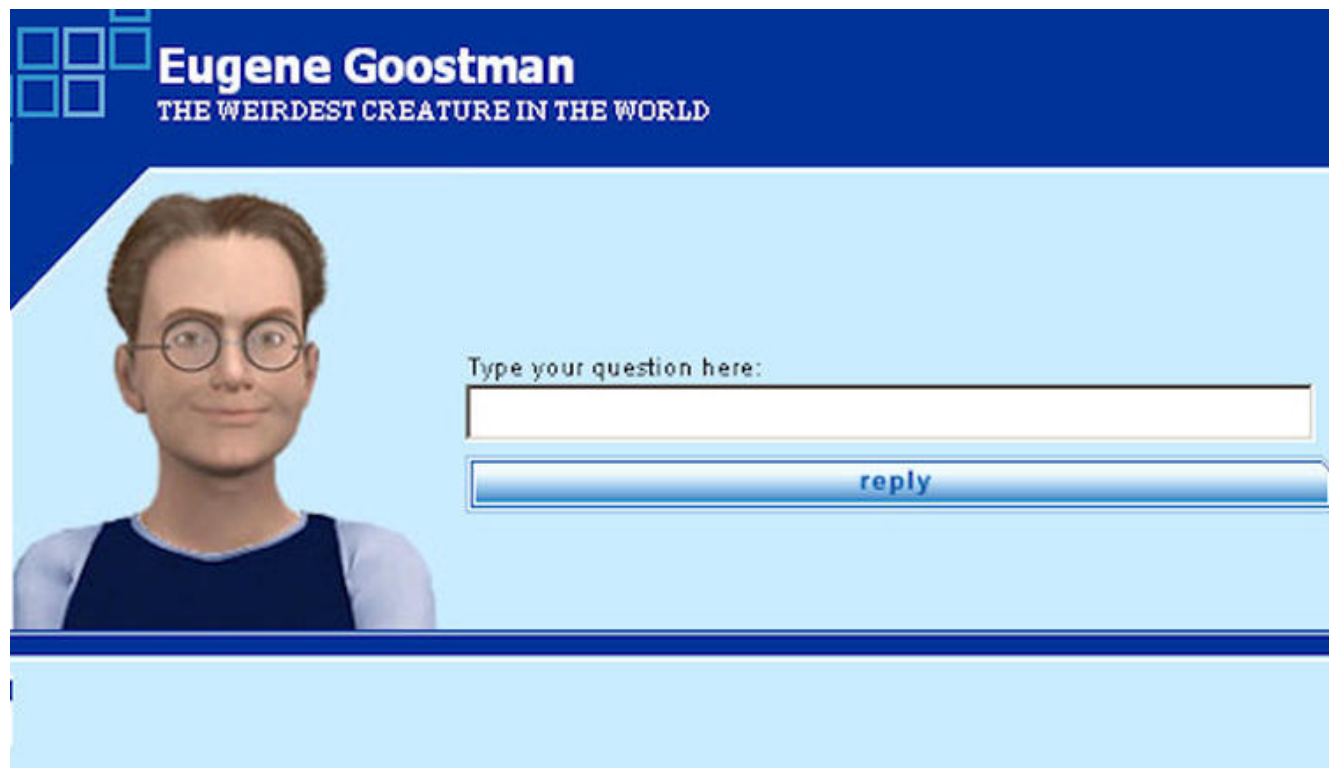
Can They Contract For Me? With Each Other?
Can They Contract With Me?
Who is Responsible?



Machines Chatting to People

Did You Mean That?

**Insulting
Offensive
Defamatory**



Machines Create Their Own Art

Writing Copy

Painting

Photo and Video

Software

Industrial Design



Are Financial Institutions Vulnerable to IP Trolls?



- IP as collateral
- Trolling Process (Katz)
- Software Ownership

Machines As Patent Inventors



UKIPO/EPO – Machine called Dabus invented two products under review for patent protection:

- ❖ Shape-shifting Food Container
- ❖ Emergency Flashlight System

USPTO has same filings: 2019

Sex and the Singularity: AI's Intrusion on Privacy

Deriving Secondary Meaning from Collections

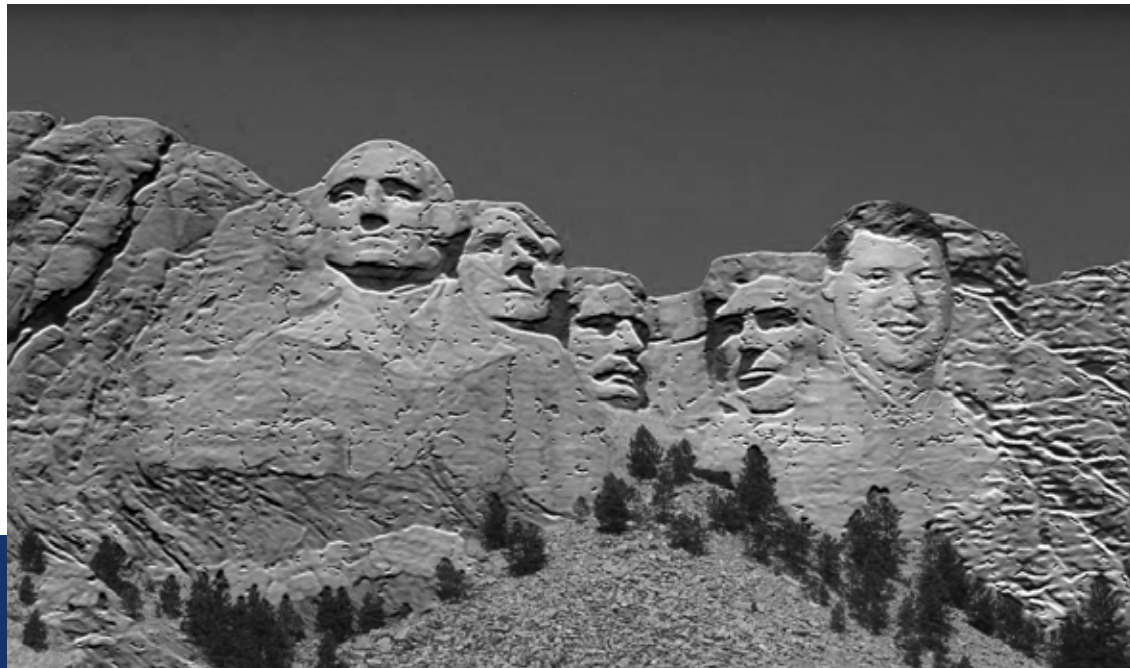


What is Meaningful Consent?
De-Anonymization
Data Aggregation: IoT Collection



Machines Falsify Pictures/Video

Deep Fakes
Manufactured Evidence
Virtual Reality



Financial Services

- Deep Fakes in Insurance Claims
- Forged Documents in Financial Deals
- Faked Documents for Securitizations
- Artificial Trading Partners (speed fakes)

AI's Impact on Security

Malware Using AI Kidnapping AI



Autonomous Vehicles

How to Make Impossible Decisions?



Self-Driving Cars

The Ethical Dilemma

Who is Liable for Loss of Life and Property?



Insuring AI Decisions

Who is left holding the Bag?

How do You Calculate Damage?

AI Programmer?

Machine Owner?



How Are Regulators Addressing AI Concerns?



Regulating AI

Who has Right and the Expertise to Regulate it?
How do you Test a Black Box Algorithm?
Should Regulators use AI to Audit Companies?
Can an AI Decision be Unfair and Deceptive?



Art. 22 GDPR: Automated Individual Decision-Making

- The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
- Controversy: GDPR mandates a “right to explanation” from machine learning models (?)
 - In plain English, the text suggests that a data subject is entitled to enough information about the automated system that she or he could make an informed decision to opt out.



Best Practices for Implementing AI Solutions

- Adopt a phase-in approach to test the fit and a governance plan with separation of duties
- Understand the training data, model design, assumptions and testing - no “black boxes”
- Determine additional compliance needs
- Don’t sign up for more red flags than you can handle
- Build in a human “gut check”
- Consider confidentiality



Ethics Guidelines for Trustworthy Artificial Intelligence

EU document, released April 2019, prepared by the High-Level Expert Group on Artificial Intelligence.

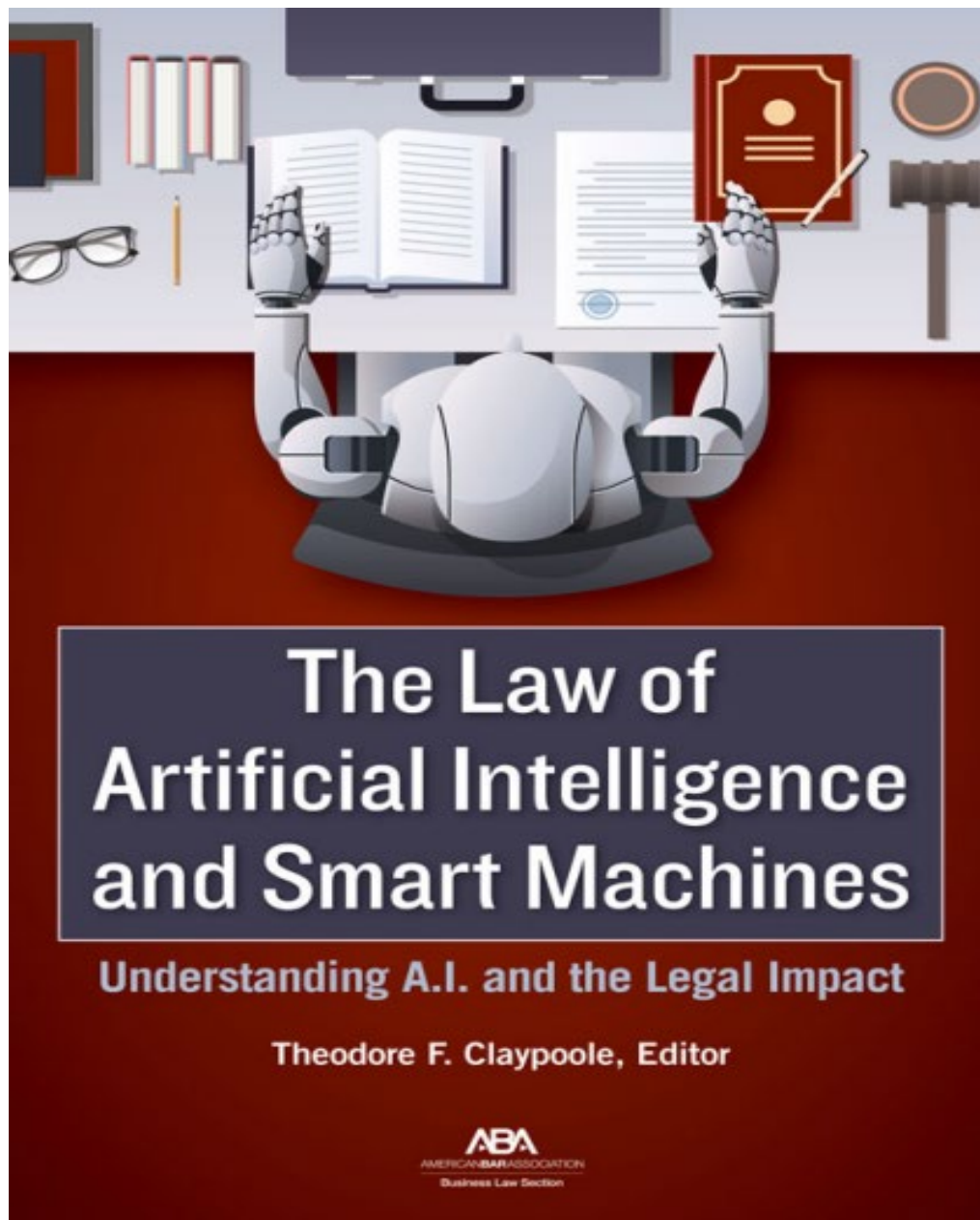
Organized by the European Commission in June 2018, as part of the AI strategy announced earlier that year.

Developing AI under ethical principles.

Ethics Guidelines for Trustworthy Artificial Intelligence

- (1) it should be lawful, complying with all applicable laws and regulations
- (2) it should be ethical, ensuring adherence to ethical principles and values and
- (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm





NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

Artificial Intelligence, FTC Authority and International Best Practices in Emerging Technologies and Financial Services



Deon Woods Bell

Federal Trade Commission

January 9, 2020



THE PREMIER SOURCE FOR CLE



FTC Authority, Business Guidance
and International Best Practices



Related Enforcement Actions



FTC Hearings &
Compliance Questions

The FTC has been in the “AI” business for decades



OECD AI PRINCIPLES



- OECD AI Definition - machine-based system that can make predictions, recommendations or decisions
- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organizations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

OECD Council Recommendation on Artificial Intelligence (2019) *available at*
[HTTPS://LEGALINSTRUMENTS.OECD.ORG/EN/INSTRUMENTS/OECD-LEGAL-0449](https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449).

What Does FTC Research Tell us About Big Data?

- The ability to collect consumer data from a variety of sources and use algorithms to:
 - extract hidden information
 - identify correlations
 - make predictions
 - draw inferences
 - glean new insights
- Three Vs:
 - volume,
 - velocity, and
 - variety



FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? (2016), available at <https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf>

Benefits

- Increase educational attainment for individual students.
- Provide access to credit using non-traditional methods.
- Provide healthcare tailored to individual patients' characteristics.
- Provide specialized healthcare to underserved communities.
- Increase equal access to employment.

Risks

- Result in more individuals mistakenly being denied opportunities based on the actions of others.
- Create or reinforce existing disparities.
- Expose sensitive information.
- Assist in the targeting of vulnerable consumers for fraud.
- Create new justifications for exclusion.
- Result in higher-priced goods and services for lower income communities.
- Weaken the effectiveness of consumer choice.

FTC Legal Landscape

- Federal Trade Commission Act (FTC Act)
- Fair Credit Reporting Act (FCRA)
- Equal Credit Opportunity Act (ECOA)
 - And potentially other legislation such as:
- Gramm-Leach-Bliley Act (GLBA)
- Children's Online Privacy Protection Act (COPPA)



FTC Jurisdiction

The FTC enforces the FTC Act, which prohibits
UNFAIR and DECEPTIVE practices

■ **Section 5 prohibits:**

- “Unfair methods of competition”

- “Unfair or deceptive acts or practices”



■ **Deceptive practices**

- Material representation or omission that is...
- likely to mislead consumers...
- who are acting reasonably under the circumstances

■ **Unfair practices**

- substantial injury that is...
- not reasonably avoidable and...
- not outweighed by benefits

FTC PRINCIPLES THAT INFORM THE AI DISCUSSION



Companies should clearly explain their practices and consumers' options – and then honor them.

FTC PRINCIPLES THAT INFORM THE AI DISCUSSION



Companies
must live up
to the privacy
and data
security
promises they
make

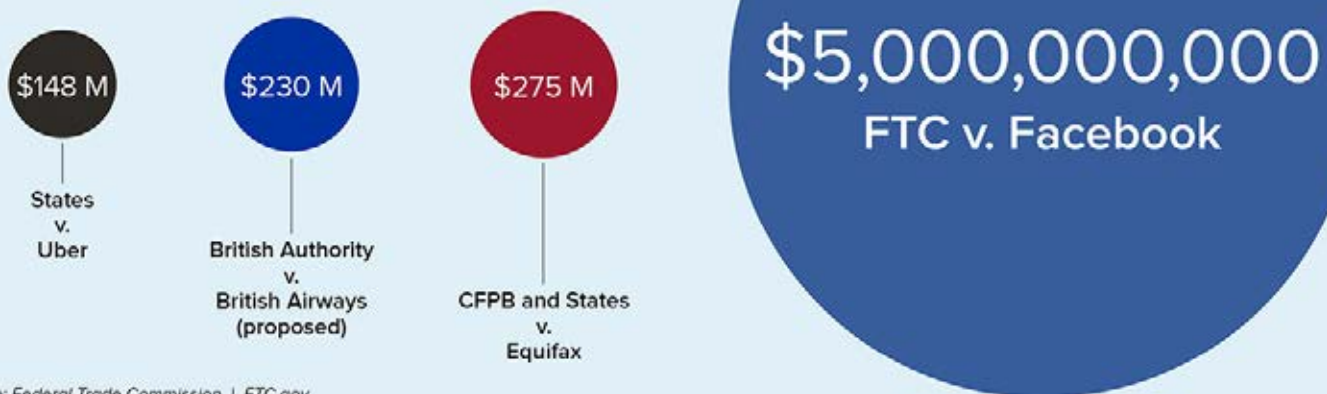




FACEBOOK ALLEGATIONS

- Violated the 2012 order by ***deceiving its users about sharing Facebook friend data*** with third-party app developers
- ***Did not screen the developers or their apps*** before granting them access to user data
- ***Misrepresented users' ability to control the facial recognition*** technology
- Misrepresentation to state ***collecting phone numbers for security, without disclosing also used numbers for advertising***

Highest Penalties in Privacy Enforcement Actions

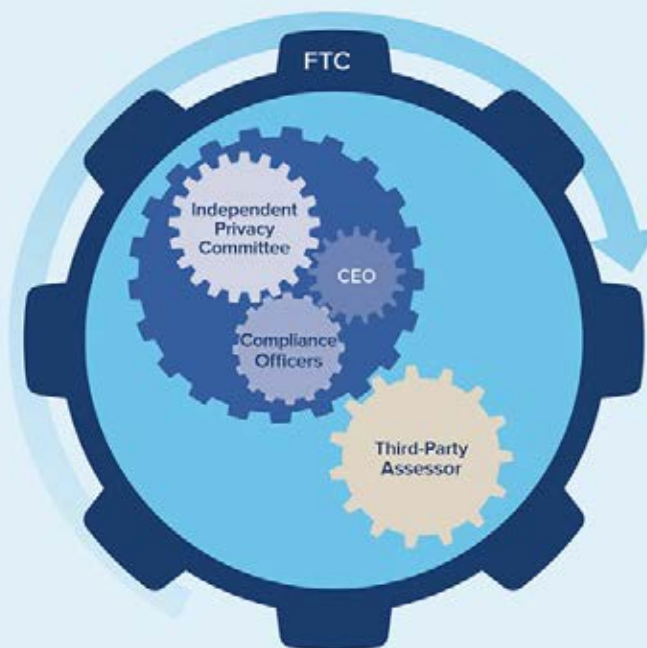


Source: Federal Trade Commission | [FTC.gov](https://www.ftc.gov)

New Facebook Privacy Compliance System

A multilayered incentive structure of accountability, transparency, and oversight

Source: Federal Trade Commission | [FTC.gov](https://www.ftc.gov)





CAMBRIDGE ANALYTICA

- Alleged to have ***used app to harvest Facebook user profile by obtaining App Users' consent to collect their Facebook profile data through false and deceptive means***
- Order as to officer and app developer
 - ***Prohibits false or deceptive statements regarding collection, use, or sale of data***
 - ***Requires destruction of data and any related work product, including any algorithms, derived from data***

Aleksandr Kogan and Alexander Nix, filed July 24, 2019

In the Matter of Cambridge Analytica, LLC, No. 9383, filed July 24, 2019

FTC PRINCIPLES THAT INFORM THE AI DISCUSSION



Companies have a legal obligation to maintain consumers' sensitive information securely.



D-LINK

- Alleged as to routers and IP cameras
 - Misrepresented ***secure in promotional materials and GUIs***
 - Misrepresented ***reasonable measures had been taken in response to security event***
 - ***Failed to take reasonable steps to secure software*** for routers and IP cameras
- Order
 - Comprehensive ***software security program***
 - ***Monitor, update***, accept vulnerability reports
 - ***Independent assessments*** of software security program from an FTC approved assessor

THE FAIR CREDIT REPORTING ACT

- The Fair Credit Reporting Act (“FCRA”)
 - Eligibility determinations
- CRAs must
 - implement reasonable procedures to ensure maximum possible accuracy of consumer reports
 - provide consumers with access to their own information, along with the ability to correct any errors.

EQUAL CREDIT OPPORTUNITY ACT

- The Equal Credit Opportunity Act (“ECOA”)
 - Disparate treatment
 - Disparate impact
- Analysis turns on whether the decisions have a disparate impact on a protected class and are not justified by a legitimate business necessity. Even if evidence shows the decisions are justified by a business necessity, if there is a less discriminatory alternative, the decisions may still violate ECOA.



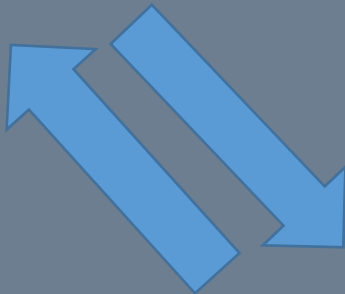
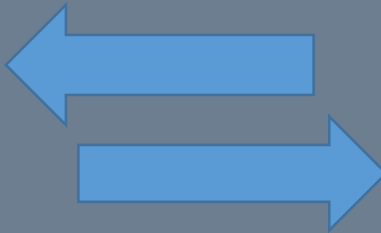
REALPAGE

- Alleged **violated FCRA** requirement that CRAs follow reasonable procedures to assure maximum possible accuracy:
 - Caused major problems for consumers with names or birthdates similar to people with criminal records
- Required to ***maintain reasonable procedures to assure the maximum possible accuracy*** of the information it includes about individuals in its consumer reports
- \$3 million civil penalty

FTC PROGRAM DEVELOPMENT DIALOGUE

investigations and
enforcement

policy initiatives,
workshops,
conferences,
studies, reports



education and
advocacy



Hearings on
**Competition and
Consumer Protection**
in the 21st Century



November 13-14, 2018:

Algorithms, Artificial Intelligence & Predictive Analytics

FTC Hearing #7: The Competition and Consumer Protection Issues of Algorithms, Artificial Intelligence, and Predictive Analytics, FTC (2018), <https://www.ftc.gov/news-events/events-calendar/ftc-hearing-7-competition-consumer-protection-21st-century>.



HEARING SEVEN TOPICS

“The consumer welfare implications associated with the use of algorithmic decision tools, artificial intelligence, and predictive analytics.

Of particular interest to the Commission: (a) the welfare effects and privacy implications associated with the application of these technologies to consumer advertising and marketing campaigns;

(b) the welfare implications associated with use of these technologies in the determination of a firm’s pricing and output decisions; and

(c) whether restrictions on the use of computer and machine learning and data analytics affect innovation or consumer rights and opportunities in existing or future markets, or in the development of new business models.”



CONSUMER PROTECTION IMPLICATIONS DISCUSSION

- Security
- Privacy
- Transparency
- Fairness
- Advantage taking
- Transformation of information
- AI by design
- Old law versus new technology
- Cost benefit balance



EXAMPLE: AI IN ONLINE ADVERTISING

- Online advertising markets match advertisers with consumers
- Machine learning models:
 - Divide customers into fine-grained and automatically-generated segments
 - Set reserve prices in auctions based on user modeling and bidder behavior
 - Automatically generate creatives fit to customers' predicted wants
- Reinforcement-learning-based tools help advertisers bid better on fine-grained segments



EXAMPLE: AI IN CREDIT SCORING

- Credit scoring context
 - scores used to assess eligibility for
 - credit where adverse action may be taken
- Benefits Lenders and Consumers
 - Better lending decisions: greater insights and more accurate scores
 - Financial inclusion: ability to include more data to broaden access to credit



AI AND ETHICS

- . . . the use of big data analytics also poses ethical risks involving fairness, accountability, and transparency. Assessments of their eligibility for credit, housing, employment, parole, school admission are enormously consequential for people.
 - Ethical Principals for AI and Data Analytics, Software & Information Industry Association



FTC's ROLE in a CHANGING WORLD

March 25-26, 2019:

Artificial Intelligence Case Study

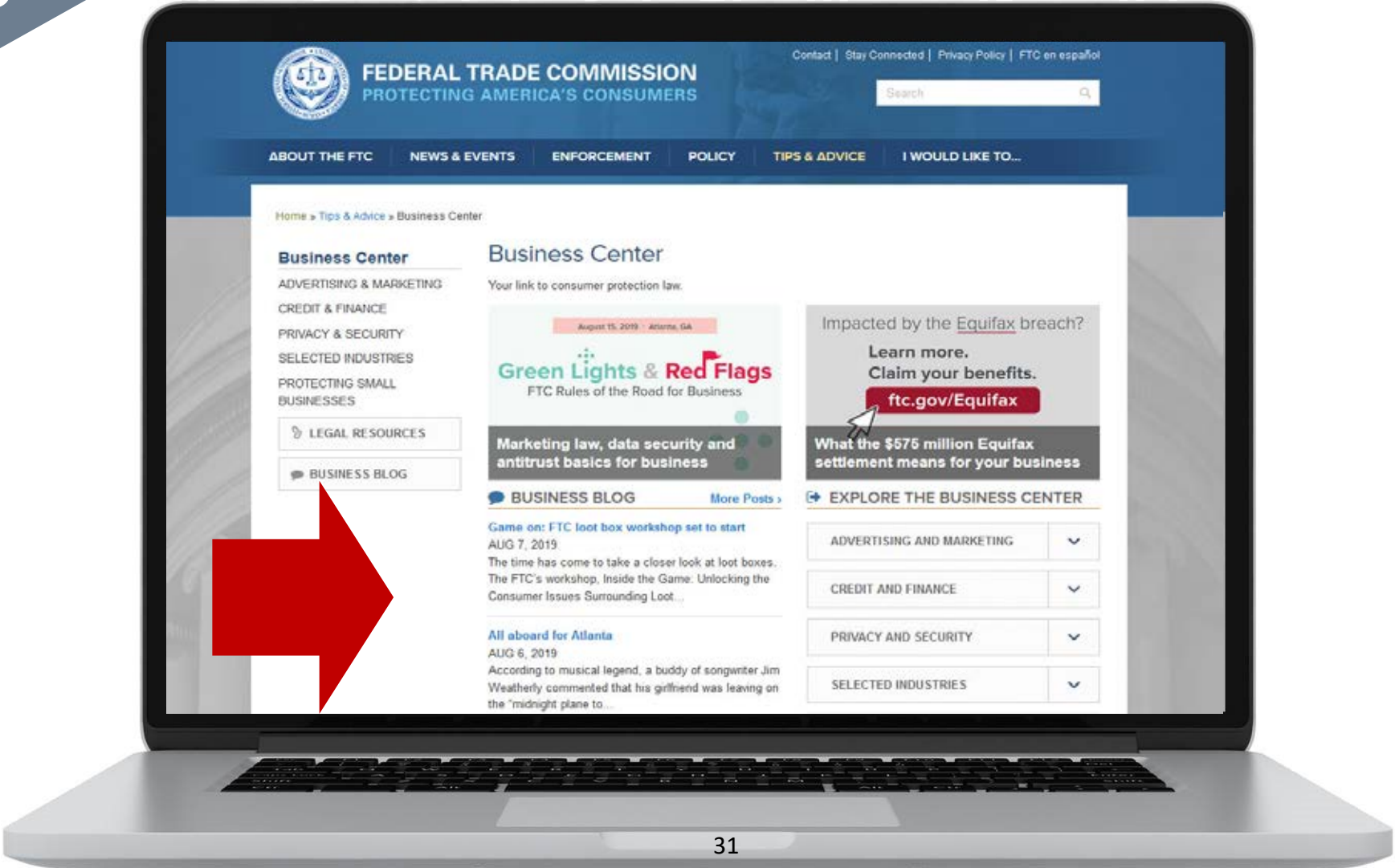
The AI Case Study session focused on two questions:

- How can the FTC best work with foreign agencies to develop effective policies on competition, consumer protection, and privacy concerning emerging technologies, such as AI? What are the challenges?
- From a practical perspective, what are the consequences of having differing approaches internationally to competition, consumer protection, and privacy enforcement regarding AI and other emerging technologies?

FTC Hearing #11: The FTC's Role in a Changing World (2019), <https://www.ftc.gov/news-events/events-calendar/ftc-hearing-11-competition-consumer-protection-21st-century>.

Where to find FTC legal resources – and to catch up on developments

business.ftc.gov

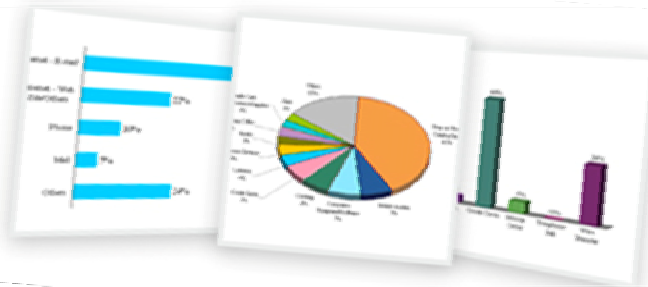


US SAFE WEB Act of 2006

- Section 3 of the Act expressly confirms:
 - 1) FTC's authority to redress harm in the US caused by foreign wrongdoers; and
 - 2) the availability in cross-border cases of all remedies available to the FTC, including restitution.
- Crucial for all jurisdictions to be able to effectively collaborate across borders with other authorities



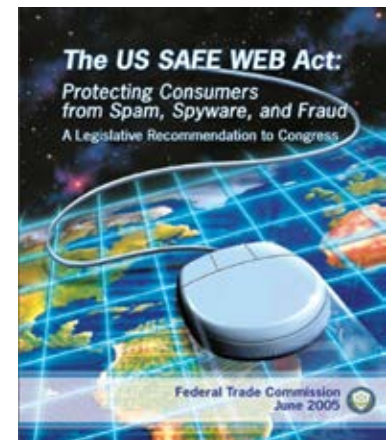
FTC Mechanisms for Cross-Border Enforcement Cooperation



Informal
information
sharing and
assistance

Complaint
sharing

Formal
investigative
assistance
and
information
sharing



Questions for Legal Compliance

- If you use AI and/or compile big data for others who will use it for eligibility decisions (such as credit, employment, insurance, housing, government benefits, and the like), are you complying with the accuracy and privacy provisions of the FCRA?
- If you receive AI or big data products from another entity that you will use for eligibility decisions, are you complying with the provisions applicable to users of consumer reports?

Questions for Legal Compliance

- If you are a creditor using AI or big data analytics in a credit transaction, are you complying with the requirement to provide statements of specific reasons for adverse action under ECOA?
- Are you complying with ECOA requirements related to requests for information and record retention?
- If you use AI or big data analytics in a way that might adversely affect people in their ability to obtain credit, housing, or employment:
 - Are you treating people differently based on a prohibited basis, such as race or national origin?
 - Do your policies, practices, or decisions have an adverse effect or impact on a member of a protected class, and if they do, are they justified by a legitimate business need that cannot reasonably be achieved by means that are less disparate in their impact?

Questions for Legal Compliance

- Are you honoring promises you make to consumers and providing consumers material information about your data practices?
- Are you maintaining reasonable security over consumer data?
- Are you undertaking reasonable measures to know the purposes for which your customers are using your data?
- If you know that your customer will use your big data products to commit fraud or for discriminatory purposes, do not sell your products to that customer. If you have reason to believe that your data will be used for these purposes, ask more specific questions about how your data will be used.

These comments are
mine and don't reflect
the official position of
the FTC.

THANKS!



Deon Woods Bell

Office of International Affairs

Federal Trade Commission

dwoodsbell@ftc.gov

+1-202-326-3307

Ethics of the Use of AI in the Practice of Law

E



-Ethics of the Use of AI in the Practice of Law-

Professor of Law Drew Simshaw – Gonzaga Univ.

Adam Nguyen, eBrevia – NY NY

Rafa Baca, Adelante IP Law – Palo Alto, CA

- <https://www.youtube.com/watch?v=VkizYljxcD8>

Overview of ABA Model Rules 1.6, 2.1 and 1.1

- What is AI – (developer technical concepts distilled for practicing lawyers)
- AI use in the Legal Field (Drew):
- “Money Balling” Litigation to elite clients and firms
- Closing the justice gap – access to legal help to the masses for greater individual empowerment within a common law democracy
- MODEL RULE 1.1: Practicing lawyers must keep abreast of Technology – last 7 years, 36 states adopted a duty of Technology Competence the above info provided to satisfy that requirement.
- Duty to supervise developers
- Duty to communicate use of AI systems to clients, other lawyers, and non-developers
- MODEL RULE 1.6 & SOFT AI – AI-driven software tools to improve efficiencies of a law practice that were previously tasked to young law firm associates – document review, eDiscovery, legal research, outcome prediction
- Data Lakes & Protection of Client Confidentialities. Lawyers insist on Anonymization algorithms for their big data driven models, secure data stores
- Underlying bias in data sets
- MODEL RULE 2.1 & HARD AI - Lawyer shall exercise independent professional judgment and render candid advice despite becoming increasingly reliant to intelligent systems
 - Draws question to extent of a Lawyer’s professional judgment is “independent” based on received AI output

What is AI (Artificial Intelligence)

- In computer science – artificial = machine; natural = all living things. Hence – artificial intelligence and natural language processing (NLP).
- Artificial or Machine Intelligence is a field of computer science and mathematics that has been around since the 1950s that promotes machines mimicking the cognitive functions of living things, i.e. human thought among others. (Alan) Turing Test – ability to execute AI that is indistinguishable from human thought.
- Then 1990s-2000s brought exponentially more powerful computer (chip) processing hardware; and the 2000s-2010s enabled historically unprecedented amounts of data to be collected and stored on dispersed computers networks (server farms) at commercially practical speeds i.e. “cloud computing” at low cost.

What is AI (Artificial Intelligence)

- AI relies on applying mathematical (statistical) probability models (or “algorithms”) to collected data to predict or “artificially reason” what a future outcome would be.
- mid-2010s saw commercially practical ability to make accurate artificial reasoning as a tool applied to our daily lives – namely super powerful, MOBLIE and fast computing power (ex: GPUs) coupled with mindboggling amounts of data (cheaply) collected to apply (statistical) algorithms to.
- The field of computer science most relevant to lawyers is Natural Language Processing AI - using the mathematical tendencies of some words to statistically clump together more than others “It is so ordered”.
- Model Rule 1.6 AI tools for lawyer efficiencies, ie doc review; Model Rule 2.1 AI tools that render professional judgment, automatically writes contracts or court motions or rules on cases

AI use in Legal Field

- “Money Balling” Litigation to elite clients and firms
- Closing the Justice Gap – AI to gain greater access to legal help

Soft AI in Law Practice

- Advanced legal research
- Automation of tasks



- Contract review
- E-discovery
- Machine learning

Hard AI in Law Practice



- Auto generated Legal Opinions
- Auto generated contracts

The World's First Robot Lawyer

The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

 [Download Now](#)



U.S. EDITION

SIGN IN

SUBSCRIBE

Newsweek



ROBOT LAWYER OVERTURNS \$4 MILLION IN PARKING TICKETS

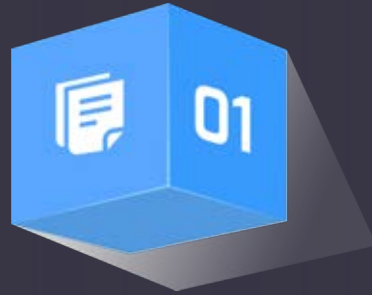
BY **ANTHONY CUTHBERTSON** ON 6/29/16 AT 6:42 AM



Legal writing assistance

- Even though “robots” are not (yet) drafting briefs, emerging tools are assisting with different stages of the legal writing process
- For example, some services provide data-driven review of drafts of legal writing by analyzing cited authority, checking the accuracy of quotations and citations, and providing suggestions based on a review of similar cases

Use cases for AI-assisted contract review



Contract
Management &
Digitization



M&A & Other
Transactional
Diligence



Audit &
Compliance



Vendor &
Customer
Management



Real Estate &
Other Leases



IP Procurement &
Management



Human
Resources



Custom Uses

AI & Model Rule 1.1:

Lawyers must be *Tech Savvy*

(Duty of Technical Competence)

- Lawyer duty to communicate use of AI systems to CLIENTS, other legal professionals?
- Lawyer Duty to Supervise Developers?

MRPC 1.1—“Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation.”

ABA Model Rule 1.1 Comm. [8] (formerly [6])—“To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology”

Competence

- At least 37 states have adopted a technological competence rule
- Commentary stresses understanding both the risks and benefits of new technologies
- Does not require lawyers to become technology experts, but might need to consult with experts when designing, adopting, and using a new tool
- Challenges presented by “black box” problem

AI & Model Rule 1.6 & Soft AI: AI Tools to improve efficiencies of Law Practice

- Document review, legal research
- eDiscovery and legal [data] analytics aka “outcome prediction”

AI & Model Rule 1.6 & Soft AI:

- Data Lakes & Protection of Client Confidentialities.
- Lawyers SHOULD insist on anonymization algorithms for their big data driven models, secure data stores

AI & Model Rule 1.6 & Soft AI:

- Understand that humans design and curate datasets so that there is an *underlying bias* inherent to any data set used in AI

MRPC 1.6—(with limited exceptions) “A lawyer **shall not reveal** information relating to the representation of a client unless the client gives informed consent”

MRPC 1.6(c)—“A lawyer shall make reasonable efforts to prevent the inadvertent or unauthorized disclosure of, or unauthorized access to, information relating to the representation of a client.”

Confidentiality

- There is now a positive obligation in the black letter of the Model Rules to take affirmative steps to protect confidentiality
- AI will create unique challenges beyond those presented by cloud computing
- Client data is being collected, managed, used, and stored (indefinitely) in new ways with AI
- The legal AI ecosystem is complex; vendor sophistication and transparency is critical

MRPC 5.1 & 5.3—Partners or managers “shall make reasonable efforts to ensure that the firm has in effect measures giving reasonable assurance that”:

5.1—“all **lawyers** in the firm conform to the Rules of Professional Conduct”

5.3— that the conduct of a **non-lawyer** employed by, retained by, or associated with the lawyer, “is compatible with the professional obligations of the lawyer”

ABA Model Rule 5.3—
“Responsibilities
Regarding Nonlawyer
~~Assistants~~ Assistance”

Supervision

- The rule title change from non-lawyer “Assistants” to “Assistance” recognizes the increased use of third party technology
- Commentary lists cloud computing as one such form of assistance, which includes many AI-driven tools

AI & Model Rule 2.1 & Hard AI:

To what extent of a Lawyer's professional judgment is

“independent” based on received AI output

- Lawyer shall exercise independent professional judgment and render candid advice despite becoming increasingly reliant to intelligent systems
- Ex: Automated document generation, judicial opinion and brief writing

ABA Model Rule 2.1—“In representing a client, a lawyer shall exercise independent professional judgment and render candid advice,” which might involve referring “not only to law but to other considerations such as moral, economic, social and political factors, that may be relevant to the client’s situation.”

ABA Model Rule 1.4—requires appropriate communication with clients “about the means by which the client’s objectives are to be accomplished.”



-Ethics of the Use of AI in the Practice of Law-

Contact Us we look forward to hearing from you:

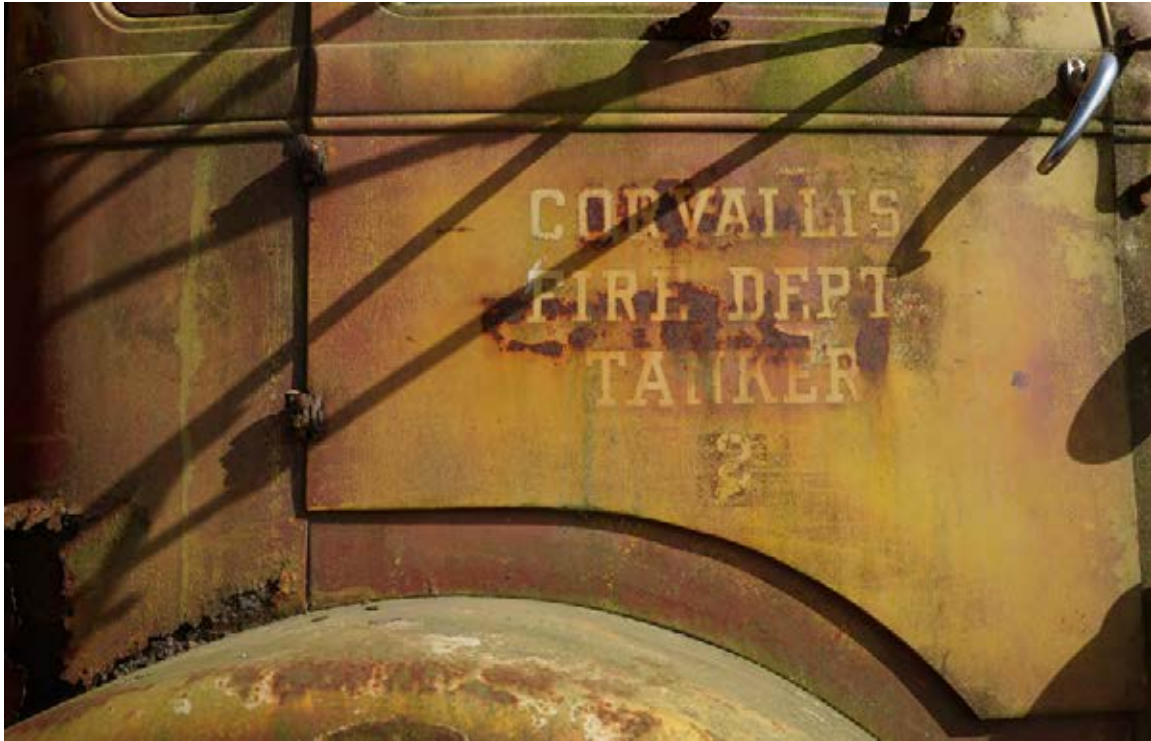
Link to Drew's Paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3308168

Adam's Link: <https://www.ebrevia.com>

Rafa's Link: <http://www.adelanteiplaw.com>

Digital Ethics, Morality and the Law

by Rafael “RAFA” Baca, Esquire & MS Data Science
San Francisco, California



FORWARD: Why Digital Ethics, Morality & the Law?

Data and computer scientists often encounter great difficulties when their professional work is applied within national legal systems, other than common and civil law systems, that do not guarantee human rights protections to its citizens. Further concerns might arise as current employers may require some professional activity that is legal but may be objectionable as being personally immoral. As section to my Lit'l Legal Handbook -- https://github.com/Rafa-Baca/LEGAL_HANDBOOK, the following strives to unravel the often-intertwined concepts in ethics, law, and morality for developers, data, and computer scientists.

FOO GOOD:

As a profession, we programmers strive to do good. “Real World” motivations are often murky and, comparably, not as clear as our programming which reliably provides well-structured output in bits and bytes. In the end, we should always strive to hack the real world based on our personal moral compass to do good to the many who will use our data and software. *Coding foo should never bar our personal and professional obligations to foo good.*

So how do we know our own personal brand of foo is good? Let's first gather a distinguishing understanding of law, ethics and morality . . .

Digital Ethics, Morality, and the Law:

Introduction: The Digital Realm -

As technology within the digital domain continues to rapidly expand and affect our everyday lives, building resonate and inspirational communities on Internet frontier is arguably one of the most noteworthy efforts in recent human history. By its very nature, creation is a messy process where a sense of wonder, lawlessness, amusement, love, and anger can all be seen all at once during this Internet “Wild West” period.

Presently, many communities across the Internet are struggling for a shared moral, ethical, and legal ethos for a fair, consensus of agreeable conduct that can be reliably enforced - a search for an innate digital sense of right and wrong. All online communities tend to agree that there should be baseline understanding of respect and accountability, although there are many approaches for how this understanding should enforced.

Illustratively, groups of entities, namely individuals, corporations, and nations, have recently charted quite different paths toward this unified goal for cyberspace. In many instances, online communities of individuals develop guidelines for conduct within their corresponding fields of interest. For example, in making a small, solitary pledge at Softwareethics.org,¹ individuals working within their professional occupations are in the process of actively developing codes of ethics for software and social engineering on the Internet. Other individuals are compelled toward online activism and even vigilantism to address many digital and physical world objectives, through the actions of such groups as Anonymous; Never Again’s “Tech Pledge” (<http://neveragain.tech/>), and with social engineering certification training programs for hackers.²

With the absence conduct in the digital realm that consistently advocates respect for basic human liberties as well as the peaceful, well-being of citizens in other nations, companies across the globe are taking a collective initiative toward creating online communities for promoting fundamental online norms while addressing a variety of issues. In response to the San Bernardino shooting, Apple and a community of companies, collectively defended their position that defends writing software code, even for screen-locks, as protected free speech and for the right to resist the orders of the government to assist in potentially incriminating a U.S. citizen in the name of counter-terrorism.

(<https://www.apple.com/newsroom/2016/03/03Amicus-Briefs-in-Support-of-Apple/>). In light of the Russians interfering with US presidential elections, Facebook, Twitter, and a community of companies have actively and favorably responded in support of “electoral integrity”, the democratic right of free and fair elections.³ As a consequence of Volkswagen software engineers programming a hack around mandatory government auto carbon emissions testing, the S&P Dow Jones Indices, the London Stock Exchange, and RobecoSAM permanently removed, for social and ethical reasons, Volkswagen AG stock from the Sustainability Indices for globally responsible companies.⁴

From a sense of right and wrong embedded deep down within our shared human experience, all these participants no matter how big, small, or sophisticated that they may be ultimately know that many of the above acts just “feel wrong”. The issue is how can that individual feeling for determining “wrongness” be applied to something as abstract and wild, at times, as the Internet? What behaviors would be acceptable to you, and how can these behaviors be fairly applied to activity on the Internet in terms of morality, ethics or laws? First, let us address the basic differences between law, ethics, and morality. Second, we see how various communities are struggling to apply various laws, ethics, and morality to the digital world at this time.

I. Morality, Laws & Ethics 101: The Basic Differences between Law, Ethics, & Morality

Ultimately, I believe that a reputation of trust is the currency that each individual actively gains while on the Internet and information or “data” is the fuel that drives one’s perceived reputation. Formulaically, as expressed in terms of statistical machine learning, one’s reputation is based on how likely others will have a “good feeling” that an individual will act favorably within accepted norms based on information of validated past events. As fake information is often purposely juxtaposed next to the truth, the Internet is a murky space to make a “good feeling” judgment from one’s perceived prior behaviors. So how do we go about navigating or perhaps socially designing some basic level of accepted norm for trust as humans increasingly connect digitally and interact with one another on the Internet?

As an overview to the following discussion, consider that each of us have personal set of values of what innately feels right or wrong. In my opinion, this innateness tends to come from a user manual for all humanity. Therefrom, morals tend to become a validated set of values when recognized by at least one other individual as fundamentally agreeable. Ethics tends to add some civic structure to some select moral values, when becoming a majority consensus of what particular behaviors within a community are agreeable and tolerable. Notably, ethics and laws are applied differently in practice – although laws can directly arise from ethical standards that a government decides to apply to preserve peace and order. Ethics are an active means for dynamic engagement within a community whereas laws tend to be static rules by which to govern effectively from. However, great social tensions happen within a nation when some ethical principles are incorrectly made law.

Let’s first gain a deeper understanding about the concept of morality, stemming from the Latin word *mores* referring to “folkways”⁵ (or values) of agreeable significance shared within a community.⁶ Our personal liberty interests are a good example of morals and are also referred to as our fundamental personal rights⁷ that, in part, comprise the Bill of Rights of the U.S. Constitution -- such as freedom of religion, right to peacefully assemble as well as the right to self determination and movement. Morals are self-understood from ones’ personal opinions of right and wrong.⁸ By extension, a moral community thus develops a core set of guiding principles that two or more individuals can tolerably agree.⁹

Thus, moral values form the foundational layer of right and wrong that corresponding moral communities ultimately ascribe to.¹⁰

“Having a moral duty” is a popular phrase and suggests that morals occupy the same social space within our greater community just as our laws do.¹¹ In reality, as morals are a personal experience shared by individuals, morals tend not to be written or clearly defined as our laws, which are systematically written and codified by the government. However, individuals often mistakenly (and dangerously) confuse morals with the law when justice is called for. Many legal actions are taken to what is actually considered a moral wrong, such as many disputes between one’s religious values and many other personal ideologies that are not collectively shared by the tolerant majority.¹² Further, to avoid the tragedies of genocide, most modern nations strive to be respectful of diverse moral interests to protect the interests of all citizens.¹³

“The wellspring of ethical change is personal morality”¹⁴ as individuals expand their shared values toward a consensus that forms a broader community’s sense of right and wrong. As illustrated by many groups of players within a single online multiplayer gamer community, the variety online sub-communities and their specific customs may be incomprehensively vast such that ethics dynamically strives to underscore those acceptable customary behaviors shared by the majority within the broader community for that particular game, such as the perennial Warcraft series.

Ethics derives from the Greek word *ethos* meaning “customary”.¹⁵ Ethical behaviors derive from a dominant moral community that is respectfully tolerant of other moral views in the minority. Ethical principles remain an open-ended narrative within a specific community as well as in a cluster of communities alike. In short, ethics is principally conveyed as shared moral values in action.¹⁶ Unlike the law, ethics cannot easily be generalized into a defined set of written rules.

In the United States, our contemporary understanding of written guiding principles or codes of ethics were first modeled from the Belmont Report of 1978 advocating ethical treatment of human subjects in scientific research.¹⁷ Influenced by the Belmont Report, many communities now tirelessly work to provide updated written codes as ethical behaviors change with time, especially within professional (workplace) communities such as software engineers’ codes of ethics under the Institute of Electrical and Electronics Engineers - Computer Science (IEEE-CS) and the Association for Computing Machinery (ACM). Although a noble idea, many professional ethics codes *per se* are not enforced just as regional governments effectively enforce legal codes. Again, ethical principles represent the continuously evolving norms of action and written ethics codes often represent aspirational exemplifications for behavior within a community - as opposed to the laws governed by nation states.¹⁸

Justice or fairness is a subset of the concept of ethics and not the law. Specifically, a “just” decision is a fair decision to the extent that each individual within a community is treated equally in terms of what they need or deserve.¹⁹ In other words, justice is a principle by which “we render to each what is due and treat like cases alike.”²⁰ Modern democratic legislatures, the courts although to a lesser extent, and similar governing bodies decide what emergent ethical principles,

including principles of justice, should be transformed to written law.²¹

Deriving from the Latin word *lex* meaning law or rule, the law creates and maintains an ordered society and ensures safety and welfare of such governed citizens.²² As discussed above, most constitutionally democratic governments further ensure that their citizens are guaranteed fundamental personal rights.

In contrast to ethics and ethical principles of justice, the law fashions well-defined civic boundaries that are easily generalized, written, and reliably enforced.²³ To this end, the law is often exalted within its own formulaic written terminology that very much differs from our everyday use of language. Further, the law requires well-defined commitment of all those who choose to be within the community of citizens governed by that law as a principal manifestation of a sense of nationalism.²⁴

As humanity is never perfect, law and justice very often collides as is manifested in the world's many civil wars and cultural wars, respectively. Laws may not be strictly enforced and give way to the unpredictably turbulent ethical battles, in the name of justice, between various moral communities underneath the canopy of a governed citizenry. Specifically, in the United States, the ethical sensibilities of justice to treat all equally and fairly have episodically intersected with the contemporary written law within the broader discourse of the nation. For example, consider the following equations: the U.S. Slave Codes were inherited from British slavery laws and enforced early on by many of the southern colonial states in United States history = legal but immoral thus unethical. After the written addition of the 13th, 14th, and 15th Amendments to the U.S. Constitution – the supreme law of the land, former slave states enact state and local “Jim Crow” laws to prevent African American and other populations from exercising their legally superior constitutional rights as citizens to vote for well over 100 years after the U.S. Civil war = illegal and immoral thus unethical – but yet wedges one community against another to the present day. There are many other unspeakable examples within this blurry war of collisions, such as the mandatory sterilization of the mentally disabled and vulnerable ethnic minority populations in the 20th Century within the United States and its protectorates²⁵ to illegalization of marijuana by Federal Bureau of Narcotics Chief Harry Anslinger²⁶ and its subsequent legalization in the 21st century (<http://fortune.com/2017/09/11/california-marijuana-drone-delivery-ban/>). Society can remain conflicted in its legal and ethical codes but each individual's sense of moral fairness toward human rights tends to ultimately bubble-up and prevail as the eventual societal norm (- with much hopeful optimism!).

II. Got Lovn Feeln? A Survey of Getting Along in the Digital Wild West

With a better understanding of morality, ethics, and the law, let us now turn to what might work best in curtailing that growing feeling of “wrongness” on the Internet today as shared by a consensus of many communities of individuals, corporations, and governments.

A. Imparting Morals to Machines: 21st Century Morality Transplant Surgery -

In the field of artificial intelligence, there is a concept of supervised machine learning where you train a computer to statistically draw inferences based on past information when given a description or “label”. In particular, to establish an initial point of reference to the learning process of a machine, humans typically provide labeled descriptions. Thereafter, for unseen instances, the machine draws on those labeled past experiences to guess the answer of greatest probability.

As such, Massachusetts Institute of Technology (MIT)’s Moral Machine project (<http://moralmachine.mit.edu/>) currently crowd sources human labeling in the context of asking human’s to provide information to assist computers to innately decide on “what is the moral thing to do?” The project prompts thousands of humans for their moral value intuition so that each dilemma shown on the website is statistically synthesized as a moral value. Those labeled moral values will be “transplanted” to machines through algorithms as if given an innate sense of what an artificial intelligent being will personally value. I suppose a community of robots that share the same values would eventually form the first artificial moral community rooted in from the many human contributors to this project.

Similarly, the U.S. Government’s Intelligence Advance Research Projects Activity (IARPA)’s CREATE project²⁷ applies traditional philosophy to artificial intelligence that assists humans to make enhanced, “human augmented” judgments to best mitigate the harms associated with human conflict. Collectively, a sourced crowd of thousands of people is submitting speech and debate arguments that assist machines to identify the good and bad parts of such arguments, even those arguments based on unreliable pretenses. Ultimately, an artificial intelligence can be developed for application in analytical toolkits to assist humans from making poorly biased judgments as well as promote better communication that yields enhanced human reasoning and conclusions.

B. Imparting Ethics to Coders -

Communities of like-valued coders are currently doing amazing things on the Internet to standardize their respective codes of ethics. Ethical Hacking certifications and professional codes for developers are the most predominant.

Many communities of hackers, computer scientists, software engineers, and digital security experts collectively agree that a coder’s social (“ethical”) responsibility is to make the public aware of software vulnerabilities; this is called “Responsible Disclosure”. In support of the “responsible disclosure” movement, some tech companies will pay individuals who find such software vulnerabilities monetary rewards, aka “bug bounties”.²⁸ Moreover, there are widely available ethical hacker certifications²⁹ as well as ethical hacking and social engineering initiatives. Such certifications are quite popular in the digital security industry.

Professional codes of ethics have been introduced to software programming in the 1940s by MIT professor Norbert Wiener.³⁰ Since then, the predominantly recited ethics codes in the digital field are the above mentioned Software Engineering Code of Ethics and Professional Practice, IEEE (Principle 1: Public 1.00 *et seq.*):³¹

- ~~• Approve software only if they have a well-founded belief it is safe and meets specifications.~~
- ~~• Accept full responsibility for their own work.~~

- **Not knowingly use software that is obtained or retained either illegally or unethically.** – IP (copyrights)
 - **Identify, define, and address ethical, economic, cultural, legal and environmental issues related to work projects.** – business & employment law, environmental law
 - **Ensure that specifications for software on which they work satisfy the users' requirements and they have the appropriate approvals.** – contract law
 - **Ensure adequate testing, debugging, and review of software.** – contract law
 - **Not engage in deceptive financial practices such as bribery, double billing, or other improper financial practices.** – criminal law ”,
- and the ACM Code of Ethics & Professional Conduct, § 1 (General Moral Imperatives):³²

- “ • ~~Contribute to society and human well-being.~~
- **Avoid harm to others.** – criminal law (& civil penalties in California)
- ~~Be honest and trustworthy.~~
- **Give proper credit for intellectual property.** IP law
- **Respect the privacy of others.** – due process, US Constitution
- **Honor confidentiality.** – IP law (trade secrets) ”.

Notably, I have crossed out the above provisions of ethical code with no foundation in U.S. Law to find that the majority of items in these two ethical codes for professional conduct are generalizations that are duplicated from existing law. In short, the majority of the IEEE and ACM ethical codes also may be separately enforced in both the criminal and civil courts by at least the underlying U.S. laws.

C. Imparting laws (and order?) to the Digital Wild West - Hacker Laws:

Currently, the Computer Fraud and Abuse Act “CFAA” (18 U.S.C. § 1030 *et seq.*) is the most infamous federal law applied to Internet hacking. The CFAA prohibits accessing or conspiring to access a computer without authorization and leads the way to individuals being prosecuted in both criminal and civil court systems. In practice, the CFAA is applicable to computers greater than ten (> 10) that are related to the federal government, across interstate boundaries, and for losses exceeding USD\$5,000.

Many other U.S. states apply much of the same provisions of the federal CFAA law but in an equivalent state law context. California, for example, subsequently enacted the Computer Data Access And Fraud Act, Cal. Pen. Code, §502 (“CDAFA”) that does not require unauthorized breaking into a computer for access as with the federal CFAA but merely requires logging into a database with a valid password - but without permission - to subsequently take data. Accordingly, California’s CDAFA has expanded the prosecutorial reach of the federal CFAA where common hacking techniques, such as webscraping, can arguably be punished under the state’s CDAFA statute after receiving notice, often the click-thru-notice from a website’s Terms & Conditions clause. At this present time, the U.S. Supreme Court is (Oct. 10, 2017) denied review of this very matter in the case *Facebook, Inc. v. Power Ventures, Inc.*, No. C-08-05780 (N.D. Cal. July 20, 2010) where California might have overstepped its prosecutorial authority.

Whistleblower Statutes:

Federal whistleblower statutes (5 U.S.C. §1201 *et seq.*) theoretically shield and potentially, financially reward employees that witness some illegal activity during

the course of their employment. Whistleblower statutes generally refer to federal protections afforded to government workers and some contractors in areas typically designated by the statute that include digital intelligence (arguably: such as contractor Edward Snowden), military (arguably: U.S. Army veteran Chelsea Manning through the whistleblower repository, WikiLeaks), securities and banking regulation, and labor and federal employment matters among others. Employees of private companies and non-profits may also be eligible for federal whistleblower protection under the law if their work is related to Environmental, Occupational, and Health reporting. There are some states that also enact their own state whistleblower statutes based on the federal law as a guideline.

Whistleblowing is another example of an unsettled cultural war where opposing ethical communities struggle to define what a whistleblower actually is in the digital realm. The overarching whistleblower laws that give protection to just one community strictly falling under the static written law but controversially denies other potential whistleblowers from those legal benefits further conflate this murky confusion. Again, the static whistleblower statutes attempt to cover a dynamically unsettled issue between ethical communities at this time. Moreover, from my personal observations, employees actually attempting to evoke whistleblower statutes have found the process disappointingly ineffective, as the U.S. Department of Justice will only accept quick, easy “open-and-shut” cases with little evidentiary holes. However, if you are professionally active in the digital security and intelligence communities, these whistleblower statutes are helpful to keep in mind as they were initially enacted by our legislatures to help those who come forward with the truth, which along with justice must be at the heart of an online democratic society.

***FOO ALARM FIRE* – SOUNDING THE ALARM:**

The law may be regarded as a formal governmental confirmation of what is already generally acceptable by the community with the largest ethical consensus. Given new social situations such as the Internet, the path from ethical principles to laws is fluid, often indirect, and unsettled. Some codes of conduct may be arguably unethical but legally valid and vice versa. In my opinion, justice is often obtained at the junction of being both clearly legal and ethical and with minimal murkiness among most individuals. It should be said that the human narrative continues to include those who struggle for justice.

As professional, individual, and business communities take positive steps to apply codes of digital ethical conduct or become signatories to online petitions, such ethical actions may not be entirely enforceable under the vast, expansive skies and natural landscapes of the lawless Digital Wild West. In my humble opinion, just as many flags were set on the Antarctic and Lunar territories under respective Treaties, I believe to provide effective enforcement a multinational legal agreement, such as an Accord (enacted by the U.S. executive branch) or, better, a treaty (enacted by U.S. Congress) would need to be in place and coordinated by some international body like the United Nations. As a smaller scale exemplary roadmap, to combat recurring digital data breeches by corporations, similar global efforts have been recently made to ensure commercial regulatory compliance with the European

Union's General Data Protection Regulation (GDPR – Regulations (EU) 2016/672) for privacy and protection of EU citizens' personal data.

In short, with the concerted approval from our U.S. representative government, a body of international law advocating a *Digital Bill of Rights* for all users remains a valid option. Our journey should begin with each of us as individual citizens and tech companies commanding our U.S. Congress to take legal action based on growing shared ethical concerns for the Internet, and directly contacting our representative through the help of the Electronic Frontier Foundation: (<https://democracy.io/#!/>).

REFERENCES:

1. <http://www.softwareethics.org/>.
2. *See eg.*
https://web.archive.org/web/20070307020902/http://www.mitnicksecurity.com/media/msc_course_outline.pdf.
3. https://www.washingtonpost.com/business/technology/facebook-to-release-russia-ads-to-congress-amid-pressure/2017/09/21/643c867c-9f10-11e7-b2a7-bc70b6f98089_story.html?utm_term=.c6d42d56fdd6.
4. <http://www.detroitnews.com/story/business/autos/foreign/2015/09/29/vw-sustainability-indexes/73019318/>.
5. "Law Morals & Ethics", by Geoffrey C. Hazard, Jr., Yale Law School, 19 S. Ill. L. U. L. J. 447-458 (1994-1995); p. 451.
6. Morality, Ethics, and Law: Introductory Concepts, Jennifer Horner, Ph.D., J.D., Seminars in Speech and Language, Vol. 24, No. 4, pp. 263 -274 (2003); pp. 263-64.
7. Id. Horner at p. 270.
8. Id. Hazard at p. 453.
9. Id. at Horner p. 264.
10. Id. Horner at p. 273.
11. Id. Hazard at p. 453.
12. Id. Hazard at p. 452.
13. Id. at Horner p. 272.
14. Id. Hazard at p. 457.
15. Id. Hazard at p. 453.
16. Id. Hazard at pp. 454-55.
17. *Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.* (1978).
(https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf).
18. Id. Hazard at p. 457.
19. Id. Horner at p.267.
20. Id. Horner at p. 268.
21. Id. Hazard at p. 456.
22. Id. Horner at p. 272.
23. Id. Hazard at p. 455.
24. Id. Hazard at p. 455.
25. https://en.wikipedia.org/wiki/Compulsory_sterilization.
26. <http://www.wnyc.org/story/man-who-declared-war-drugs/>.
27. <https://www.iarpa.gov/index.php/research-programs/create>.
28. https://hackerone.com/directory?query=type%3Ahackerone&sort=resolved_reports_closed%3Adescending&page=1.
29. <https://iclass.eccouncil.org/our-courses/certified-ethical-hacker-ceh/>.
30. https://en.wikipedia.org/wiki/Programming_ethics(hereinafter "Wiki Programming").
31. Id. at wiki programming, *see also* <http://www.acm.org/about/se-code>.
32. Id. at wiki programming, *see also*

Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law

DREW SIMSHAW[†]

As in many other industries, artificial intelligence (“AI”) is poised to drastically transform the legal services landscape. “Bots,” automated expert systems, and predictive analytics are already changing the way consumers seek, and lawyers provide, legal services. Among other impacts, AI has the potential to increase access to justice in the self-help, individual, and corporate law firm markets by lowering costs and expanding services to untapped markets. A prominent question in early literature on AI in law is whether these services constitute the unauthorized practice of law. Threshold questions of whether and by whom such services should be regulated are important, but will likely not be answered (or even answerable) until AI’s impact on the profession is more cognizable. In the meantime, there is currently no comprehensive guidance for attorneys on how AI should be developed, adopted, and used in ways that conform to a lawyer’s ethical obligations. Without such guidance, law firms and third-party services risk designing and adopting AI-driven tools that fail to provide effective client-centered services, inhibit wide-spread access to justice, and undermine lawyers’ ethical obligations to current and former clients, including the obligations to practice competently, maintain confidentiality, effectively supervise third parties, communicate with clients, and exercise independent judgment and render candid advice. This Article initiates this critical dialogue by exploring the types of AI being implemented in the profession, and identifying characteristics of these emerging services that will present ethical tensions and challenges. It rigorously examines existing guidance from the ABA and state bar authorities concerning new technology in practice, and identifies areas where this guidance is not sufficient to confront the unique ethical issues presented by AI. This article does not attempt to provide detailed or prescriptive guidance on these issues, but rather identifies the imminent challenges not currently being addressed in the literature on AI and legal ethics, or by bar authorities. The concluding recommendations will set the stage for and inform future scholarship and discussions concerning legal ethics, access to justice, and unauthorized practice of law in the age of AI.

[†] Drew Simshaw is a Visiting Associate Professor of Law, Legal Practice, at Georgetown University Law Center. He thanks the participants of Yale Law School’s 2017 We Robot Conference and the Mid-Atlantic Clinicians Workshop hosted by Georgetown Law, as well as his colleagues at Georgetown Law and Elon Law for their helpful comments and feedback. He also thanks Michael Yoder, Joshua Jordan, and Nicholas Schloss for their research assistance.

TABLE OF CONTENTS

INTRODUCTION	174
I. THE ROLE OF ARTIFICIAL INTELLIGENCE IN IMPROVING ACCESS TO JUSTICE.....	179
A. THE IMPORTANCE OF FOSTERING THE DEVELOPMENT OF A ROBUST LEGAL SELF-HELP MARKET	180
B. THE LIMITS OF ARTIFICIAL INTELLIGENCE IN IMPROVING ACCESS TO JUSTICE	183
II. THE RISE OF ARTIFICIAL INTELLIGENCE IN LAW AND THE IMMEDIATE AND INEVITABLE CHALLENGES.....	187
A. SOFT AI, STRONG AI, AND “DATAFICATION”	187
B. AI’S IMPACT ON THE DEMAND FOR LEGAL SERVICES AND NEED FOR HUMAN LAWYERS	189
C. SIGNIFICANT CHARACTERISTICS OF DEPLOYED AND DEVELOPING FORMS OF AI IN LAW PRACTICE	192
III. CONFRONTING AI’S CHALLENGES THROUGH A RENEWED COMMITMENT TO ETHICAL OBLIGATIONS	195
A. CURRENT ETHICAL GUIDANCE IS INSUFFICIENT TO ADDRESS THE UNIQUE CHALLENGES POSED BY AI IN LAW PRACTICE	195
1. <i>Competence</i>	196
2. <i>Confidentiality</i>	198
3. <i>Supervising Third Parties</i>	201
4. <i>Communicating with Clients</i>	202
5. <i>Independent Judgment and Candid Advice</i>	204
6. <i>Obligations to Former Clients</i>	205
B. NEEDED GUIDANCE CONCERNING THE DESIGN, ADOPTION, AND USE OF AI IN LAW PRACTICE.....	206
1. <i>The Need for Urgency</i>	207
2. <i>Guidance Is Needed During Critical Design Stages of Early AI</i>	207
3. <i>The Need for Proactive, but Not Prescriptive, Guidance</i>	208
CONCLUSION.....	210
APPENDIX	212

INTRODUCTION

In late 2014, an overturned parking ticket signaled a pivotal moment in a revolutionary era of legal services. In some ways, the process of overturning the ticket was familiar. The driver believed the ticket was unjust, the argument was documented and placed before a decision maker, and the decision maker

determined that a fine was not warranted. But in many other ways, this was not a typical transaction of legal services. The driver did not interact with any humans, look up any laws, or fill out any forms. Instead, the driver answered a few questions asked online by a “bot” that then automatically filled out the necessary forms and filed them with the appropriate local government office, free of charge.

Over the next twenty-one months, recipients of a quarter million other parking tickets in New York and London followed suit by seeking the services of the same bot. What these drivers might not have realized is that the mastermind and coder behind the bot did not pass the bar exam, earn a J.D., or even attend a single law school class. In fact, he had only just recently graduated from high school. But most of these drivers probably did not care, because the bot worked well—really well. By mid-2016, it had successfully overturned 160,000 tickets (a 64% success rate), helping those who used the service avoid over \$4 million in fines.¹ By July 2017, the service had saved users \$9.3 million by disputing 375,000 parking tickets.²

Joshua Browder, the bot’s now 21-year-old creator, has called this service something that likely resonates with users—“the world’s first robot lawyer.”³ But in many ways, the service is unlike a robot and unlike a lawyer. DoNotPay.co.uk, the website where users interact with the bot, does not have arms, a voice, or any of the anthropomorphic features typically associated with “robots.” Moreover, people do not typically turn to lawyers to address legal needs as minor as parking tickets. However, as DoNotPay expanded to cities like Seattle, the service began to take on more lawyer-like tasks, including using a driver’s answers to questions to draft a letter to a city’s parking enforcement office to challenge a ticket.⁴ In July 2017, DoNotPay expanded its service to all fifty states.⁵ If the early demand and success of DoNotPay is any indication, parking tickets are just the beginning of artificial intelligence’s (“AI’s”) transformation of legal “self-help” services, and, indeed, the legal services industry as a whole.

Individual consumers of legal services are not the only ones engaging with AI. Rather, lawyers and law firms are too, and in big ways. ROSS Intelligence⁶ (“ROSS”) has marketed itself as “the world’s first artificially intelligent

1. Samuel Gibbs, *Chatbot Lawyer Overturns 160,000 Parking Tickets in London and New York*, GUARDIAN (June 28, 2016), <https://www.theguardian.com/technology/2016/jun/28/chatbot-ai-lawyer-donotpay-parking-tickets-london-new-york>.

2. John Mannes, *DoNotPay Launches 1,000 New Bots to Help You with Your Legal Problems*, TECHCRUNCH (July 12, 2017), <https://techcrunch.com/2017/07/12/donotpay-launches-1000-new-bots-to-help-you-with-your-legal-problems/>.

3. Gibbs, *supra* note 1.

4. Arezou Rezvani, *‘Robot Lawyer’ Makes the Case Against Parking Tickets*, NPR (Jan. 16, 2017, 3:24 PM), <http://www.npr.org/2017/01/16/510096767/robot-lawyer-makes-the-case-against-parking-tickets>.

5. Shannon Liao, *‘World’s First Robot Lawyer’ Now Available in All 50 States*, VERGE (July 12, 2017, 2:44 PM), <https://www.theverge.com/2017/7/12/15960080/chatbot-ai-legal-donotpay-us-uk>.

6. ROSS INTELLIGENCE, <http://www.rossintelligence.com/> (last visited Nov. 21, 2018).

attorney,” and in May 2016, BakerHostetler “hired” the service.⁷ ROSS answers natural language questions asked by subscribing attorneys, “reads” over one million pages per second that it accesses from its partnering legal publisher,⁸ and provides answers along with specific text from laws, cases, and secondary sources.⁹ Unlike existing legal “data providers,” ROSS’s co-creator describes its service as providing “insight” into the law that is “jurisdictionally aware,” and able to provide updates as the law and its interpretation change.¹⁰ ROSS uses IBM’s Watson technology—the same technology that defeated humans on *Jeopardy!*¹¹—in a way that uses semantics that match not only keywords, but similar concepts.¹² Other large firms are jumping on board quickly.¹³ In addition to changing the way lawyers perform legal research, AI is poised in the near term to drastically transform the nature and efficiency of document review, e-discovery, and the way lawyers predict the outcomes of their decisions and cases.¹⁴

But overturning parking tickets, improving lawyer efficiency, and reducing costs for law firm clients is just the beginning of AI’s potential in the legal profession. If implemented responsibly, AI could expand access to legal services to parts of society that have historically been shut out. For example, in 2016, DoNotPay expanded its service from contesting parking tickets to combating homelessness by helping recently evicted people apply for emergency housing.¹⁵ In addition, in March 2017, it began exploring the possibility of helping refugees seek asylum and file immigration applications in the United States and Canada.¹⁶ Similarly, in an effort to “democratize the law,” ROSS strives to make its technology “easily accessible to all legal service providers and educators,”¹⁷ as

7. Karen Turner, *Meet ‘Ross,’ the Newly Hired Legal Robot*, WASH. POST (May 16, 2016), <https://www.washingtonpost.com/news/innovations/wp/2016/05/16/meet-ross-the-newly-hired-legal-robot/>.

8. TED Institute, *The World’s First AI Legal Assistant* | Andrew Arruda | TED Institute, YOUTUBE (Dec. 21, 2016), <https://www.youtube.com/watch?v=wwbr0fombFs>.

9. Vanderbilt University, *Andrew Arruda: Artificial Intelligence and the Law Conference at Vanderbilt Law School*, YOUTUBE (May 6, 2016), https://www.youtube.com/watch?v=LF08X5_T3Oc.

10. *Id.*

11. John Markoff, *Computer Wins on ‘Jeopardy!’: Trivial, It’s Not*, N.Y. TIMES (Feb. 16, 2011), <http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?pagewanted=all>.

12. John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041, 3049 (2014).

13. See ROSS INTELLIGENCE, *supra* note 6 (listing K&L Gates, Latham & Watkins, Salazar Jackson, von Briesen & Roper, Bryan Cave, Womble Carlyle, and Dickinson Wright as law firms using ROSS).

14. See *infra* Subpart II.C.

15. Elena Cresci, *Creator of Chatbot That Beat 160,000 Parking Fines Now Tackling Homelessness*, GUARDIAN (Aug. 11, 2016), <https://www.theguardian.com/technology/2016/aug/11/chatbot-lawyer-beat-parking-fines-helping-homeless-do-not-pay>.

16. Elena Cresci, *Chatbot That Overturned 160,000 Parking Fines Now Helping Refugees Claim Asylum*, GUARDIAN (Mar. 6, 2017), <https://www.theguardian.com/technology/2017/mar/06/chatbot-donotpay-refugees-claim-asylum-legal-aid>.

17. ROSS INTELLIGENCE, *How to Leverage Legal Technology and Bridge the Justice Gap: ROSS Intelligence’s Mission to Democratize the Law*, <https://rossintelligence.com/leverage-legal-technology-bridge-justice-gap/> (last visited Nov. 21, 2018) (“Arruda expressed our pledge to give ROSS away for free to all lawyers on the front lines to best help them do their jobs. . . . [W]e are committed to partnering with . . . bar

well as nonprofits like Upsolve,¹⁸ which provides free bankruptcy assistance to low-income New Yorkers.

As these examples demonstrate, there is great demand for AI in the law. AI is likely to be integrated into the profession at unprecedented rates, if not out of a sense of duty to close the justice gap, then out of a sense of competitive and economic necessity.¹⁹ Even so, there is a growing consensus that the future of legal services is not likely one in which AI fully replaces human lawyers.²⁰ To date, however, most attention has been paid to the corporate law firm setting, and there is less certainty what effect AI will have on the market for individual legal services.²¹

In any event, as illustrated by DoNotPay, lawyers, AI services, and third parties will likely all be involved at some point during a large majority of cases. Leading up to an expansion of service in 2017, the creator of DoNotPay relied heavily on actual lawyers.²² He also consulted extensively with lawyers in the United States, United Kingdom, and Canada in his effort to try to make the service effective in helping refugees initiate immigration and asylum applications.²³ At least one human lawyer has suggested that, once applications are submitted, immigration attorneys will embrace the opportunity to step in and pick up where the bots left off,²⁴ perhaps themselves utilizing a ROSS-like AI service in the near future. Because the future of legal services is one involving a complex ecosystem of lawyers, artificially intelligent systems, third-party service providers, and other non-lawyers, the legal profession must take a

associations . . . justice commissions, the courts, pro bono and public interest groups, legal service organizations, and law schools . . .”).

18. See UPSOLVE, <http://upsolve.org/> (last visited Nov. 21, 2017) (“Our mission is to help low-income Americans in financial distress get a fresh start through Chapter 7 bankruptcy at no cost.”); see also Joe Borstein, *Can Technology Automate Your Rights? Upsolve Thinks So*, ABOVE THE L. (Sept. 28, 2016, 3:58 PM), <https://abovethelaw.com/2016/09/can-technology-automate-your-rights-upsolve-thinks-so/> (discussing Upsolve and ROSS’s partnership).

19. Julie Sobowale, *Beyond Imagination*, 102 A.B.A. J. 47, 52 (2016) (“Law firms are feeling the pressure from clients, particularly in-house counsel, to lower costs. And artificial intelligence is born out of necessity.”); see also *infra* note 99 and accompanying text.

20. See, e.g., Dana Remus & Frank Levy, *Can Robots Be Lawyers?: Computers, Lawyers, and the Practice of Law*, 30 GEO. J. LEGAL ETHICS 501, 501 (2017) (“[A]utomation has a measurable impact on the demand for lawyers’ time, but one that is less significant than popular accounts suggest.”).

21. Tanina Rostain, *Robots Versus Lawyers: A User-Centered Approach*, 30 GEO. J. LEGAL ETHICS 559, 574 (2017) (“In the individual legal services sphere, [unlike the corporate market] [], legal technologies are a response to unmet legal needs and regulatory barriers to developing other forms of access to the legal system. What effect these technologies will have on the individual market for legal services is unknown.”).

22. See Liao, *supra* note 5 (describing how Browder recruited “volunteer and part-time lawyers to help him with the legal aspect of the tool”).

23. See Cresci, *supra* note 16 (explaining that Browder worked with lawyers in each country, and quoting him as saying “I wanted to make sure I got it right because it’s such a complicated issue. I kept showing it to lawyers throughout the process and I’d go back and tweak it”).

24. *Id.* (quoting immigration lawyer Sophie Alcorn as saying, “It will be easier for applicants to submit their applications and it will empower legal aid organisations [sic] to assist a larger numbers of clients”).

comprehensive approach to ensuring that AI is integrated responsibly, morally, and ethically into all forms of legal services.

Moreover, AI's transformation of the legal profession will not be without practical and ethical challenges. On the legal self-help front, courts, state legislatures, and bar associations in the near term will have to decide whether increasingly sophisticated services such as DoNotPay constitute the unauthorized practice of law. Some have argued that such prohibitions would not only represent ill-advised and short-sighted policy, but would also be immoral and unethical in light of the current access-to-justice crisis and likely concurrent uninhibited proliferation of AI in large law firms, which serve mostly corporate clients. In addition, a robust market for artificially intelligent legal self-help services will increasingly involve more human lawyers who have their own ethical obligations, making legal ethics oversight a critical forum for confronting AI's challenges. The extent of such oversight, and how it is structured, are important questions, but ones that might not be settled before AI is implemented in the profession to an even greater degree. Accordingly, urgent guidance is needed regarding emerging forms of "soft AI" in law practice, and possible forms of "strong AI" in the future.²⁵ This paper offers a starting point for this guidance by identifying the topics on which guidance is needed.

Part I examines the role of technology broadly, and AI specifically, in improving access to justice, including the importance of facilitating the development of a robust legal self-help market, while also recognizing AI's limits in these efforts. Part II identifies in more detail the various kinds of AI that are affecting the practice of law. In part, this section examines AI's impact on the demand for legal services and need for human lawyers, and the specific characteristics of deployed and developing forms of AI in law practice, including those associated with document review, e-discovery, legal research, and outcome prediction.

Part III of the Article examines how the legal profession should confront these challenges, recognizing past ethical guidance concerning other less transformative technologies, and focusing on the specific implications with regard to lawyers' obligations concerning competence; confidentiality; supervising third parties; communicating with clients; exercising independent judgment and rendering candid advice; and obligations regarding former clients. Subpart III.B. stresses the urgent need for guidance concerning the design, adoption, and use of AI, especially during critical design stages. It also examines the need for, and stakeholders' willingness to issue, proactive, humanistic guidance. The Article concludes by summarizing the areas that should be the subject of initial guidance from within the profession. Without such guidance, law firms and third-party services risk designing and adopting AI-driven tools that fail to provide effective client-centered services, inhibit wide-spread access

25. See *infra* Subpart II.A.

to justice, and undermine lawyers' ethical obligations to current and former clients. The concluding recommendations will set the stage for future discussions concerning legal ethics, access to justice, and unauthorized practice of law in the age of AI.

I. THE ROLE OF ARTIFICIAL INTELLIGENCE IN IMPROVING ACCESS TO JUSTICE

The United States is in the midst of an access to justice crisis. Too many people lack access to the legal services they need, usually because they cannot afford them. The Brennan Center for Justice reports that “[e]ighty percent of low income people have trouble obtaining legal representation or otherwise accessing the civil court system to protect their property, family, and livelihood.”²⁶ The Legal Services Corporation (“LSC”) defines the “justice gap” as “the difference between the unmet need for civil legal services and the resources available to meet that need,” and has determined that technology can be a powerful tool in narrowing it.²⁷ While AI alone cannot close the gap, previous transformative technologies have been credited with making some significant strides.²⁸

The most transformative technology to date in the legal services industry, as in most industries, has been the Internet, which has, among other things, helped link low-income clients to free legal services.²⁹ In addition, many legal services and resources are now available online. For example, the advent of “online courts” has improved access to court systems,³⁰ and “collaborative technology” has proven especially helpful in alternative dispute resolution forums.³¹ Various forms of automation in online dispute resolution processes have also demonstrated an ability to improve access to justice.³² Indeed,

26. BRENNAN CTR. FOR JUST. AT N.Y.U. SCH. OF LAW, CLOSING THE JUSTICE GAP <https://www.brennancenter.org/issues/closing-justice-gap> (last visited Nov. 21, 2017).

27. LEGAL SERVS. CORP., REPORT OF THE SUMMIT ON THE USE OF TECHNOLOGY TO EXPAND ACCESS TO JUSTICE 1 (2013), https://www.lsc.gov/sites/default/files/LSC_Tech%20Summit%20Report_2013.pdf.

28. Melissa A. Moss, *Can Technology Bridge the Justice Gap?*, 90 FLA. B.J. 83, 86 (2016) (“While it is apparent technology alone cannot bridge the justice gap, it is also apparent the justice gap cannot be bridged without embracing technology.”).

29. See Raymond H. Brescia et al., *Embracing Disruption: How Technological Change in the Delivery of Legal Services Can Improve Access to Justice*, 78 ALB. L. REV. 553, 597 (2015) (exploring how Pro Bono Net developed web-based tools to increase access to pro bono services for the poor and unrepresented).

30. See Mark A. Cohen, *Online Courts: Using Technology to Promote Access to Justice*, LEGAL MOSAIC (Aug. 15, 2016), <http://legalmosaic.com/2016/08/15/online-courts-using-technology-to-promote-access-to-justice/> (online courts “provide the population inexpensive, fast, and easy access to justice for a range of civil disputes”).

31. See Michael J. Wolf, *Collaborative Technology Improves Access to Justice*, 15 N.Y.U. J. LEGIS. & PUB. POL’Y 759, 762 (2012) (“[T]hese powerful yet accessible tools can *dramatically* improve access to civil justice in America both in traditional court cases and alternative dispute resolution (ADR) forums.”).

32. See Anjanette H. Raymond & Scott J. Shackelford, *Technology, Ethics, and Access to Justice: Should an Algorithm Be Deciding Your Case?*, 35 MICH. J. INT’L L. 485, 491 (2014) (noting that online dispute resolution systems have increased access to justice).

DoNotPay's automated intelligence is yet another example of a service accessed via the Internet.

Online solutions to closing the justice gap have their limits, including existing barriers to Internet access for large portions of the population.³³ However, to the extent that technology has been successfully leveraged in the past to improve access in the legal services industry, AI will be an even more impactful force than previous tools, and has the potential to magnify and transform benefits of existing technologies. Some of these benefits are being brought to life through innovation in programs and clinics at law schools throughout the country.³⁴ One important component of this progress will be fostering the development of AI in the legal self-help market, while still confronting the ethical challenges AI presents.

A. THE IMPORTANCE OF FOSTERING THE DEVELOPMENT OF A ROBUST LEGAL SELF-HELP MARKET

One major barrier to individuals accessing the legal services they need is prohibitively high costs.³⁵ Legal self-help, including the various iterations of DoNotPay, is one way that people have historically avoided these high costs, by simply not hiring a lawyer and instead opting to "do-it-yourself." Of course, these services have been around far longer than DoNotPay, dating back to before the Internet and even before widely available consumer software. The evolution of legal publisher Nolo is representative of the way that some within the industry have adjusted their business model to recognize and meet this massive demand.³⁶ Founded in 1971, Nolo began by publishing do-it-yourself law books, before eventually offering affordable software that helped users fill out common legal forms without the assistance of an attorney.³⁷ Other services have since emerged as online start-ups. The popular and controversial service LegalZoom can, among other things, generate a draft will based on input regarding assets and intentions for estate disposal.³⁸ As DoNotPay has demonstrated, AI is poised to

33. See *infra* Subpart I.B.

34. See, e.g., *Georgetown's Iron Tech Lawyer Competition 2018*, GEO. L. INST. FOR TECH. L. & POL'Y, <http://www.georgetowntech.org/irontechlawyer/> (last visited Nov. 21, 2018); see also Ronald W. Staudt & Andrew P. Medeiros, *Access to Justice and Technology Clinics: A 4% Solution*, 88 CHI.-KENT L. REV. 695, 698 (2013) (proposing that law schools offer an Access to Justice Technology Clinic and discussing the program's success at Chicago-Kent College of Law).

35. See Michael Zuckerman, *Is There Such a Thing as an Affordable Lawyer?*, ATLANTIC (May 30, 2014), <https://www.theatlantic.com/business/archive/2014/05/is-there-such-a-thing-as-an-affordable-lawyer/371746/> (discussing the failure of the legal market to provide affordable services).

36. See Kelly Phillips Erb, *Are We Ready for Robot Lawyers?*, 38 PA. LAW. 54, 55 (May/June 2016) (explaining that Nolo.com "has proven that some clients are looking for solutions to legal problems without the need to hire a lawyer and pay fees").

37. See Nolo, www.nolo.com (last visited Nov. 21, 2018) (listing do-it-yourself books and products).

38. McGinnis & Pearce, *supra* note 12, at 3050. Indeed, "[t]rust and estate planning is already ripe for this kind of mechanization because this area of law has relatively few kinds of forms and unique factual situations that arise for the large majority of people." *Id.*

make major inroads in the legal self-help industry as services become increasingly advanced, while requiring less of users.

There is no shortage of important legal issues that will need to be confronted as AI expands the availability and effectiveness of legal self-help services, including what the appropriate liability standard is for these services when something goes wrong.³⁹ Legal self-help, especially when offered online, also blurs state lines when trying to determine what jurisdiction's practice rules should apply.⁴⁰ Perhaps the most widely publicized issue, though, is whether these services constitute the unauthorized practice of law ("UPL").⁴¹ This Article does not attempt to answer line-drawing UPL questions concerning AI. Rather, this Article argues in part that there are fundamental ethical issues that must be articulated in order to not only more fully inform the UPL debate, but also to guide the development, adoption, and use of AI by consumers and firms that are not waiting for answers to the UPL questions. Even so, it is important to acknowledge the UPL debate when framing ethical issues.

Approaches to dealing with the challenges presented by emerging technology-fueled self-help services currently vary widely. Some state bars have aggressively tried to prohibit certain legal self-help services, including LegalZoom,⁴² a move that some commentators have discouraged. Caroline E. Brown, for example, has responded to prohibitions by arguing that lawyers should support legal self-help "because providing access to affordable legal services works to close the justice gap without significantly threatening the legal profession," and believes accordingly that "unauthorized practice of law regulations should be amended to include an exception to the definition of

39. See generally Benjamin H. Barton, *Some Early Thoughts on Liability Standards for Online Providers of Legal Services*, 44 HOFSTRA L. REV. 283 (2015).

40. See Thomas E. Spahn, *Artificial Intelligence: Ethics Issues*, TSZJ10 ALI-CLE 1 (2018) (discussing how states have begun to de-emphasize lawyers' "physical presence" and acknowledge that lawyers can practice "virtually" and permanently in a state where they are not licensed); see also Jordan Bigda, Note, *The Legal Profession: From Humans to Robots*, 18 J. HIGH TECH. L. 396, 425 (2018) (arguing that new law regarding the jurisdictional limitations surrounding artificially intelligent lawyers should mimic the rules of paralegals); Julee C. Fischer, Note, *Policing the Self-Help Legal Market: Consumer Protection or Protection of the Legal Cartel?*, 34 IND. L. REV. 121, 127-28 (2000) (noting that technological advancements in lawyering "knocks down the barriers between persons, states and even countries").

41. See, e.g., Spahn, *supra* note 40 ("Artificial Intelligence represents the latest and perhaps the most advanced step in a continuum of non-human processes for providing what could be seen as legal advice."); William J. Connell, *Artificial Intelligence in the Legal Profession—What You Might Want to Know*, 66 R.I. B.J. 5, 43 (2018) ("If computer programs are writing briefs, or at least creating preliminary drafts, is that the practice of law? Will programs that incorporate artificial intelligence need to be licensed by the Bar Association and the Supreme Court?"); Bigda, *supra* note 40, at 423 ("If lawyers begin outsourcing work to robots and artificially intelligent programs, will this lead to ethical issues of the unauthorized practice of law?").

42. See, e.g., Rachel M. Zahorsky, *Alabama Bar Group Files Suit to Ban LegalZoom*, A.B.A. J. (July 15, 2011, 8:48 PM), http://www.abajournal.com/news/article/alabama_lawyer_group_files_suit_to_ban_legalzoom/; *LegalZoom Targeted in Legal Software Ban in Missouri*, SOCAL TECH (Aug. 1, 2011), http://www.socaltech.com/legalzoom_targeted_in_legal_software_ban_in_missouri/s-0037234.html; Bill Draper, *Missouri Lawyers Challenge LegalZoom's Service*, CBS ST. LOUIS (Aug. 1, 2011), <https://www.insurancejournal.com/news/midwest/2011/08/01/208821.htm>.

‘practice of law’” that explicitly permits such services.⁴³ Others commentators have taken a more middle-of-the-road approach by acknowledging that restrictive regulations might be out of date, but still advocating for some form of oversight.⁴⁴ Most recently, in 2017, Dana Remus and Frank Levy argued that, “[t]o make informed regulatory decisions, lawyers generally and bar committees in particular will have to become more informed and more skilled with new legal technologies,” and that “[b]oth groups will . . . need to struggle with the bounds of the ‘practice of law’ and with the increasingly mixed nature of legal expertise and other forms of expertise.”⁴⁵

The questions of whether and by whom AI-driven services should be regulated are important, but will likely not be answered (or even answerable) until AI’s impact on and within the profession is more cognizable. Although some states continue to fight emerging self-help services on UPL grounds, the current prevalence of such services suggests that states will not be able to completely suppress the availability of AI-driven services.⁴⁶ It is hard to overlook, however, that the legal profession’s advocacy for crippling restrictions on legal self-help solutions could potentially stunt the development of the larger AI revolution in law in ways that would ultimately favor large firms over the public interest. Such predictions have led some commenters, such as Cody Blades, to offer defenses of services like LegalZoom which could also apply to emerging AI services:

The legal community has spoken repeatedly throughout history about a duty that each attorney has to provide services to those that cannot otherwise afford them. Although this ideal has not been met by the legal community, LegalZoom provides an alternative that is working. To block access to legal services because of something as amorphous as “practice of law” statutes is to effectively deny access to legal services to those whom the legal community has neglected: a miscarriage of justice and a failure of the profession’s ethical obligations.⁴⁷

The UPL debate is just one example of why AI must be comprehensively addressed within the legal profession. At the very least, the profession should not rush to prohibit self-help services utilizing AI if large law firms simultaneously remain permitted to incorporate AI services into their delivery

43. Caroline E. Brown, Note, *LegalZoom: Closing the Justice Gap or Unauthorized Practice of Law?*, 17 N.C. J.L. & TECH. 219, 222–23 (2016).

44. See, e.g., Mathew Rotenberg, Note, *Stifled Justice: The Unauthorized Practice of Law and Internet Legal Resources*, 97 MINN. L. REV. 709, 712 (2012) (offering “solutions to anachronistic and inconsistent unauthorized practice of law statutes” while also recognizing that some regulation of internet legal providers is needed).

45. Remus & Levy, *supra* note 20, at 555–56.

46. See Spahn, *supra* note 40 (“As th[e] technological evolution has demonstrated, lawyers often fight rearguard actions in attempts to prohibit laymen from using books, software, etc.—contending that such non-human aids constitute the illegal unauthorized practice of law by their creators. But lawyers ultimately lose each fight. It would be safe to presume that the same outcome will occur with artificial intelligence.”).

47. Cody Blades, *Crying over Spilt Milk: Why the Legal Community Is Ethically Obligated to Ensure LegalZoom’s Survival in the Legal Services Marketplace*, 38 HAMLINE L. REV. 31, 55 (2015).

models. Such inequity in access to AI's potential could serve to increase the justice gap, rather than narrow it.

However, as this Article will demonstrate, the questions of whether to adopt certain AI services and how to properly use them are difficult (and perhaps counterintuitive at times). For example, the co-creator of ROSS has suggested that there is an ethical obligation to use AI in practice because it lowers prices for clients.⁴⁸ Others have suggested that, even if such an obligation does not yet exist, future advances and benefits might make it irresponsible to refrain from using AI.⁴⁹ Regardless of whether such an obligation is ever widely adopted, the increased availability of AI-driven legal services will force all lawyers to consider the extent to which they are obligated to exercise, among other things, competence and zealotry in understanding and adopting AI services or tools that improve objective efficiency in practice, despite parallel vulnerabilities associated with the use of these technologies that implicate other obligations, such as the duty to protect client confidentiality.⁵⁰

Moreover, even if legal self-help is permitted to continue to advance in some form, the underlying AI that drives it, and the similar services utilized by lawyers themselves, have their limits when it comes to closing the justice gap.

B. THE LIMITS OF ARTIFICIAL INTELLIGENCE IN IMPROVING ACCESS TO JUSTICE

For the potential benefits of AI to come to fruition in the legal field, both lawyers and those seeking legal services will require access to AI services and associated technologies. For many, such access has been historically elusive. At a fundamental level, if someone lacks even basic Internet access, that person cannot utilize online legal self-help services such as DoNotPay. Similarly, if a small public interest law firm or public defender's office lacks the funds necessary to contract for emerging third-party AI services, the benefits of those

48. Andrew Arruda, *An Ethical Obligation to Use Artificial Intelligence? An Examination of the Use of Artificial Intelligence in Law and the Model Rules of Professional Responsibility*, 40 AM. J. TRIAL ADVOC. 443, 455–57 (2017).

49. See, e.g., Roy D. Simon, *Artificial Intelligence, Real Ethics*, N.Y. ST. B. ASS'N J., http://www.nysba.org/Journal/2018/Apr/Artificial_Intelligence,_Real_Ethics/ (last visited Nov. 21, 2018) (“Do you have a duty to alert your clients to the option of using AI products that may save substantial fees or arrive at quicker or more accurate results? Right now the answer to that question is unclear—but before long, practicing law without using AI will be like practicing law with an Underwood manual typewriter, and you will have to tell your clients that there is a better, cheaper, faster way.”); Turner, *supra* note 7 (quoting Ryan Calo as saying “Eventually, I bet *not* using these systems will come to be viewed as antiquated and even irresponsible, like writing a brief on a typewriter.”); see also Tejas G. Patel, Note, *Document Automation Software: Solving the Dichotomy Between Meeting Attorneys’ Financial Needs and Ethical Obligations*, 19 SUFFOLK J. TRIAL & APP. ADVOC. 361, 393 (2014) (“[T]he current ethical rules continue to allow lawyers to be inefficient and charge what they believe is a reasonable fee, but in reality is unreasonable when considering how much lower their fees can be if they use a new system of billing using automation software.”).

50. See *infra* Subparts III.A.1, III.A.2.

services will remain elusive to those lawyers and their clients. This lack of access could prove to be a serious impediment to improving access to justice with AI.

Clients and their lawyers have not always had sufficient access to other forms of technology that have otherwise had significant impact on the delivery of legal services. Many individuals, especially those who are indigent, lack access to the Internet and other technological resources necessary to make full use of other emerging and potentially transformative technological resources.⁵¹ Some communities, especially in rural areas, still lack basic Internet access.⁵² Even in more urban areas, where Internet access is more widely available, it has been reported that communities of color experience lower connection speeds than those provided to wealthier communities served by the same provider—a term known as “redlining.”⁵³ Many poor Americans rely on their cell phones as their sole means of accessing the Internet,⁵⁴ subjecting them to inferior and limiting interfaces when accessing services only available online. And in many instances, those that do have access to more robust technology, nevertheless lack the experience necessary to make effective use of it.⁵⁵

Allowing technology, including AI, the opportunity to help close the justice gap necessarily requires efforts to mitigate these inequalities.⁵⁶ Although there have been federal initiatives that recognize the need for, and are aimed at improving, Internet access for low income Americans,⁵⁷ these programs have experienced significant opposition and cutbacks from the federal government

51. See Eric J. Magnuson & Nicole S. Frank, *The High Cost of Efficiency: Courthouse Tech and Access to Justice*, 22 PROF. LAW. 16, 17 (2014) (“For all the benefits that the justice system stands to gain from technology, however, there are unanticipated consequences that affect the most vulnerable of society. Indigent people have fewer resources, including access to technology.”).

52. Darrell M. West & Jack Karsten, *Rural and Urban America Divided by Broadband Access*, BROOKINGS (July 18, 2016), <https://www.brookings.edu/blog/techtank/2016/07/18/rural-and-urban-america-divided-by-broadband-access/>.

53. See Jon Brodtkin, *AT&T’s Slow 1.5Mbps Internet in Poor Neighborhoods Sparks Complaint to FCC*, ARS TECHNICA (Aug. 24, 2017, 10:20 AM), <https://arstechnica.com/information-technology/2017/08/atts-slow-1-5mbps-internet-in-poor-neighborhoods-sparks-complaint-to-fcc/> (discussing the formal complaint which alleges that lower-income AT&T subscribers receive slower Internet than their higher-income counterparts); see also Formal Complaint of Joanne Elkins, Hattie Lanfair & Rachelle Lee at 7, Joanne Elkins et al. v. AT&T Corp., No. EB-17-223 (FCC Aug. 24, 2017).

54. See generally Radhika Marya, *Cellphones Are Now Essentials for the Poor*, USA TODAY (Sept. 14, 2013, 9:14 AM), <https://www.usatoday.com/story/money/personalfinance/2013/09/14/cellphones-for-poor-people/2805735/>.

55. See Courtney Gilmore, *The Impact of Technological Illiteracy*, RICH. J.L. & TECH. BLOG (Jan. 21, 2018), <http://jolt.richmond.edu/2018/01/21/the-impact-of-technological-illiteracy/> (noting that “access to the web does not render a person, in this case a school aged students [sic], as having computer literacy”).

56. See Magnuson & Frank, *supra* note 51, at 18 (“[A]pproached with an eye toward mitigating this inequity, technology can help close the justice gap.”).

57. See, e.g., *Lifeline Support for Affordable Communications*, FED. COMM. COMMISSION, <https://www.fcc.gov/consumers/guides/lifeline-support-affordable-communications> (last visited Nov. 21, 2018).

following the 2016 presidential election.⁵⁸ This could harm the development and deployment of AI in the legal field because some consumers will not have the means to access AI services. Accordingly, there must be increased efforts from within the legal community to improve consumers' technology access and literacy as the profession continues to rely on such technology, and especially as it continues to integrate, and rely on, AI. As explained below, effective and responsible use of AI by lawyers will require clients to comprehend AI to some extent,⁵⁹ and they will only be able to understand AI if they have access to, and understand, the associated technology.

Consumers are not the only players who lack access. Access to technology has also been a historical barrier for some lawyers, and especially public interest lawyers with fewer resources than large firms. Some lawyers find themselves at a disadvantage if they are unable to afford emerging services and tools,⁶⁰ or are unable to adjust their service and business models to incorporate a new technology. Even the financial and time costs associated with testing a new service to see if it is useful, are costs too great for small legal offices to bear. On the other hand, large law firms typically have more resources to invest,⁶¹ and more flexibility to experiment with and adjust to the changing technological landscape.⁶²

If large law firms are the only consumers, or the only *paying* consumers, of AI legal services, then these barriers could result in design bias that favors the needs of the types of clients that hire the services of large law firms. Inequalities that marginalize or remove certain lawyers from the AI market could place certain parts of the profession at a competitive disadvantage, to the detriment of

58. See, e.g., Issie Lapowsky, *Millions Need the Broadband Program the FCC Just Put on Hold*, WIRED (Feb. 14, 2017, 9:30 AM), <https://www.wired.com/2017/02/millions-need-broadband-program-fcc-just-put-hold/>.

59. See *infra* Subpart III.A.1.

60. Connell, *supra* note 41, at 41 ("The costs of these programs may be expensive, so this may result in even more pressure being placed upon smaller firms or the solo practitioner who may not have the resources to purchase these programs. Lawyers who do not have access to these services will be competing with those who do.").

61. Sean Semmler & Zeeve Rose, Note, *Artificial Intelligence: Application Today and Implications Tomorrow*, 16 DUKE L. & TECH. REV. 85, 90 (2017) ("[There is a] possibility that big firms, with their resources and profit margins, are well situated to gain access to this disruptive technology at an earlier stage than smaller firms. Subscriptions to legal A.I. applications may be expensive (early on), and if big firms can buy this technology, become familiar with it now, and use it to attract new clients while retaining their old clientele, then by the time smaller firms get access to the same technology, it may be too late.").

62. Kurt M. Saunders & Linda Levine, *Better, Faster, Cheaper—Later: What Happens When Technologies Are Suppressed*, 11 MICH. TELECOMM. & TECH. L. REV. 23, 43–44 (2004) ("[S]mall firms may be adequate for handling minor innovations, but other innovations may be so large that only a large firm can mass the needed funds, equipment, talent, and sustained effort. Also, the risk may be so high that only secure dominant firms can take the chance. . . . [I]nnovation is often speeded when several firms race to invent or innovate first. The resulting gain in competitive speed may offset any economies of scale in innovation that might exist." (alteration in original) (quoting WILLIAM G. SHEPHERD, *THE ECONOMICS OF INDUSTRIAL ORGANIZATION* 145 (3d ed. 1990))).

large parts of society.⁶³ Such inequality prevents technology from fulfilling its potential role as “the great equalizer.”⁶⁴ If technology is able to fill this role, then, in theory, this would ultimately benefit clients that are members of historically disadvantaged groups. However, the transformative role of technology has its limits, many of which stem from systematic inequality.⁶⁵ As this article explains, there is reason to believe that AI will be adopted at much quicker rates than other technology, by those that have the means to do so—namely large firms with significant financial resources—meaning that inequalities could be magnified quickly if the profession does not address access soon.

The challenges resulting from a possible design bias favoring paying clients of AI services is compounded by inevitable underlying and often unconscious biases of the designers of AI,⁶⁶ as well as underlying bias in the data that are fed into AI’s algorithms⁶⁷ and the resulting disparate impact that manifests in legal systems.⁶⁸ All of these challenges warrant urgent and comprehensive attention with an eye toward the potential risks and benefits of AI, as well as guidance concerning lawyers’ ethical obligations.

However, in many significant ways, AI will be different from other technologies that have been the subject of guidance from within the profession to date. To fully understand why AI will be different, and to appreciate the significance and implications of these differences, a closer look at AI and the way it has manifested, and will manifest, itself within the legal profession, is necessary.

63. See Semmler & Rose, *supra* note 61, at 90 (“Legal tech companies that wish to create more universal access to legal technology should be careful to ensure that their technology is not used to entrench larger firms in positions of power (even more than they already are).”).

64. Dimitri Kanevsky, *Technology Change as the Great Equalizer*, WHITE HOUSE (May 7, 2012, 12:55 PM), <https://obamawhitehouse.archives.gov/blog/2012/05/07/technology-change-great-equalizer>.

65. See Adrienne LaFrance, *Technology, the Faux Equalizer*, ATLANTIC (Mar. 31, 2016), <https://www.theatlantic.com/technology/archive/2016/03/half-full-tech/476025/> (“Silicon Valley’s sunny outlook on technology and opportunity ignores systematic inequalities,” like the fact that not everyone has Internet access, and so technology alone cannot be the “equalizing force.”).

66. See, e.g., Kate Crawford, Opinion, *Artificial Intelligence’s White Guy Problem*, N.Y. TIMES (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0 (“Like all technologies before it, artificial intelligence will reflect the values of its creators.”); Peter Rejcek, *The Struggle to Make AI Less Biased than Its Creators*, SINGULARITYHUB (Jan. 31, 2017), <https://singularityhub.com/2017/01/31/the-struggle-to-make-ai-less-biased-than-its-creators/>.

67. Jamie J. Baker, *Beyond the Information Age: The Duty of Technology Competence in the Algorithmic Society*, 69 S.C. L. REV. 557, 558, 569 (2018) (“[T]here are problems with blindly relying on algorithms because they lack transparency in generating results. With this lack of transparency, lawyers must be extra vigilant in ethically relying on these results in the face of machine learning bias or other. . . . [D]ata-drive decision-support systems can perpetuate injustice, because they can be biased either in their design, or by picking up human biases” (quoting Iyad Rahawn, *Society-in-the-Loop: Programming Social Contract*, 20 ETHICS INFO. TECH. 5, 6 (2018))).

68. See generally Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (discussing the harm that can result from relying on algorithmic techniques).

II. THE RISE OF ARTIFICIAL INTELLIGENCE IN LAW AND THE IMMEDIATE AND INEVITABLE CHALLENGES

A. SOFT AI, STRONG AI, AND “DATAFICATION”

AI has been defined as “the ability of machines to execute tasks and solve problems in ways normally attributed to humans.”⁶⁹ At a fundamental level, there are two kinds of AI. The first has been called “soft AI.”⁷⁰ Like early examples of groundbreaking uses of “big data,”⁷¹ soft AI is purely focused on mimicking human intelligence and attempts to produce outcomes that to a high degree match those that would have been produced by humans acting alone.⁷² Soft AI does this without any attempt to replicate the underlying processes by which humans actually reach those outcomes.⁷³ Many of the emerging instances of AI in law are examples of this soft AI, including AI tools that aid with document review, e-discovery, legal research, and outcome prediction.⁷⁴

One major challenge posed by soft AI is its primary, if not exclusive, use of what Daniel Katz describes as “observational data.” Katz explains that, “[u]sing large segments of observational data, today’s soft AI is built upon modeling what people actually do, thereby allowing a machine to probabilistically emulate their behavior under analogous conditions.”⁷⁵ This is problematic when trying to emulate the behavior of lawyers because legal strategy often involves considering factors that are not currently observable by machines because certain associated data are never, or at least less often, “datafied.”

“Datafication,” a term coined by Viktor Mayer-Schönberger and Kenneth Cukier, refers to the act of transforming something into “a quantified format so it can be tabulated and analyzed.”⁷⁶ Many pieces of client information are not currently datafied, and for good reason. For instance, a legal brief will not reference certain pieces of embarrassing or sensitive client information that for any number of reasons a lawyer and the client may have determined should be

69. *What’s Next for Artificial Intelligence*, WALL ST. J., <http://www.wsj.com/articles/whats-next-for-artificial-intelligence-1465827619> (last updated June 14, 2016, 1:14 AM) (quoting Yann LeCun, then director of artificial-intelligence research at Facebook).

70. Irving Wladawsky-Berger, *‘Soft’ Artificial Intelligence is Suddenly Everywhere*, WALL ST. J. (Jan. 16, 2015, 12:49 PM), <https://blogs.wsj.com/cio/2015/01/16/soft-artificial-intelligence-is-suddenly-everywhere/>.

71. See generally VIKTOR MAYER-SCHONBERGER & KENNETH CUKIER, *BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK* (2013) (surveying big data’s growing effect on business, government, science and medicine, privacy, and the way we think).

72. Wladawsky-Berger, *supra* note 70.

73. Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909, 918 (2013) (“Today’s AI is ‘soft AI’ because it attempts to mimic human intelligence in outcomes, but not in its underlying processes.”).

74. See *infra* Subpart II.C.

75. Katz, *supra* note 73, at 918–19 (citing Steven Levy, *The AI Revolution Is On*, WIRED (Dec. 27, 2010, 12:00 PM), http://www.wired.com/magazine/2010/12/ff_ai_essay_airevolution/).

76. MAYER-SCHONBERGER & CUKIER, *supra* note 71, at 76–78.

excluded from formal or informal documentation during the case. For example, a sexual assault survivor might not want a certain electronic communication to be a part of a case because it could open an opportunity for defense counsel to distort that communication through the perpetuation of a rape myth.⁷⁷ The fact that this information never makes its way into an internal or external database does not mean that the underlying facts are not important to the case. In fact, it is quite the opposite; sensitive information often affects, if not drives, the overall legal strategy employed in a case. However, if this information is never formalized, it is not “observable” to soft AI assistance that might be able to otherwise make valuable use of it. This paradox inevitably leads to tensions between a lawyer’s different ethical obligations.

Communication between lawyers and their clients, including discussion of sensitive facts or secrets, is a critical component of effective and ethical lawyering. The duty to discuss with a client the means by which the client’s objectives are to be achieved necessarily involves discussing and dealing with sensitive facts when crafting legal strategy.⁷⁸ As AI development progresses to include tools that can help develop legal strategy (for example one based on past outcomes),⁷⁹ a lawyer who adopts a service that fails to account for, or fails to make appropriate use of, such information, risks unethically marginalizing or even ignoring the client’s objectives during key decision-making phases of the representation.⁸⁰ However, a lawyer who *does* utilize an AI tool that not only incorporates, but also deeply analyzes, such sensitive information faces unique confidentiality concerns beyond those currently associated with more prevalent technology.⁸¹ At the same time, a lawyer who ignores such potentially helpful, efficient services risks failing to competently and zealously represent their client at an affordable price.⁸² This Article argues that these tensions make it imperative that ethical obligations are rigorously scrutinized in light of any given system’s proposed service, and that lawyers, firms, bar associations, and legal

77. See generally Drew Simshaw, *Title IX in the Technological Age—Challenging Rape Culture and Myths Through Fairer Use of Electronic Communications*, 6 TENN. J. RACE, GENDER & SOC. JUST. 275 (2017) (advocating for the use of electronic communications to enable Title IX enforcement and subsequent criminal rape trials).

78. See MODEL RULES OF PROF’L CONDUCT r. 1.4 (AM. BAR ASS’N 2016) (governing attorney-client communications, and requiring lawyers to “promptly” communicate and consult their clients).

79. Daniel Ben-Ari et al., “*Danger, Will Robinson?*” *Artificial Intelligence in the Practice of Law: An Analysis and Proof of Concept Experiment*, 23 RICH. J.L. & TECH. 2, 35 (2017) (“Computers could do the work of a lawyer—examining a case, analyzing the issues it raises, conducting legal research, and even deciding on a strategy.”).

80. See *infra* Subpart III.A.1.

81. See *infra* Subpart III.A.2.

82. See *infra* Part III; see also Turner, *supra* note 7 (quoting Ryan Calo as saying “Eventually, I bet not using these systems will come to be viewed as antiquated and even irresponsible, like writing a brief on a typewriter.”).

ethics oversight bodies immediately initiate a dialogue regarding these services, including developing formal or informal guidance.⁸³

The second kind of AI, “strong AI,” or “hard AI,” looks beyond mere outcomes based on inputs, and actually attempts to mimic real human processes. AI that is this advanced is still a thing of the future. Luke Nosek explains:

[W]e remain stages away from creating an artificial general intelligence with anywhere near the capabilities of the human mind. We don’t yet understand how general, human-level AI (sometimes referred to as AGI, or strong AI) will work or what influence it will have on our lives and economy.⁸⁴

There is no doubt that there will continue to be tremendous demand for AI services that take on increasingly central components of legal research and case development. As a result, there is reason to believe that there will be some amount of pressure from both partners and clients of law firms to adopt such advanced services due to their business efficiencies compared to human labor.⁸⁵ This will raise a host of issues concerning the moral and ethical implications of such advanced services, and inevitably raises the question of whether “robot lawyers” will take human lawyers’ jobs.

B. AI’S IMPACT ON THE DEMAND FOR LEGAL SERVICES AND NEED FOR HUMAN LAWYERS

It is at this point in most legal AI discussions that some lawyers question why the profession is focusing on responsible use of AI when it should be plotting how to prevent inevitable “robot lawyers” from taking their jobs. Indeed, many, including Rickard Susskind, believe that to some degree this is what the legal services industry has in store.⁸⁶ These concerns are not limited to the legal profession. Indeed, automation has reduced the need for many forms of labor. Between 2000 and 2012, roughly a half million auto manufacturing jobs were lost, largely due to automation.⁸⁷

83. See *infra* Parts III, IV.

84. *What’s Next for Artificial Intelligence*, *supra* note 69 (quoting Luke Nosek, co-founder of PayPal and the Founders Fund).

85. See, e.g., Erb, *supra* note 36 (“Integrating robot lawyers or programs that can run repetitive tasks is cheap. Robots don’t ask for promotions, and they don’t want bonuses. . . . In the age of apps and the Internet, consumers increasingly want answers immediately. A firm that relies on computers and not on people can spit out answers almost instantaneously, and it can do so 24 hours a day. Robots don’t need breaks, they work weekends and evenings, and they don’t go on vacation.”).

86. RICHARD SUSSKIND & DANIEL SUSSKIND, *THE FUTURE OF THE PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS* 66–71 (2015) (describing how artificial intelligence will replace attorneys, and other professionals, by providing the same services at low-to-no cost).

87. Erb, *supra* note 36.

However, it is likely that AI will have a more profound impact on the workforce than past technological transformations. Some have predicted that 47% of jobs in the U.S. could be automated and replaced by robots controlled by computers in the next two decades.⁸⁸ Andrew Ng has elaborated:

The age of intelligent machines will see huge numbers of individuals unable to work, unable to earn, unable to pay taxes. Those workers will need to be retrained—or risk being left out in the cold. We could face labor displacement of a magnitude we haven't seen since the 1930s.⁸⁹

Depending on the number of individuals that need to be “retrained” at a given time in a given sector, there will not necessarily be a sufficient number of jobs available by the time those individuals reenter the workforce. Some believe that the legal services industry will not be immune to this trend,⁹⁰ even despite the highly specialized training that sets lawyers apart from what many consider to be more vulnerable professions.⁹¹ Regardless of whether AI's emergence can be characterized as “taking jobs,” its role will certainly remove lawyers from certain components of the current legal services model.⁹²

However, when it comes to machine learning, as opposed to just automation in general, AI cannot yet replicate human capabilities.⁹³ Even if it could, especially in the legal services industry, humans are too essential to completely remove from the lawyering process. As the deputy director of the Florida Bar Foundation, Melissa Moss, has explained, “When the technology is simply too much or the user has an emergency situation that demands immediate attention, alternatives that involve immediate human intervention have to be built into systems.”⁹⁴ In 2017, Professors Remus and Levy argued that artificial intelligence will change, but not replace, the work performed by lawyers, and concluded that the hours worked by lawyers in corporate firms will be reduced by only about 2.5% annually over the next five years.⁹⁵

In the long term, if concerns are addressed, it might be that AI is less likely to “take” lawyers' jobs, and more likely to enable them to make services available to untapped markets. For example, as previously referenced,

88. *Id.*

89. *What's Next for Artificial Intelligence*, *supra* note 69 (quoting Andrew Ng, chief scientist at Chinese Internet giant Baidu).

90. *See, e.g.*, Katz, *supra* note 73, at 963 (“[W]ith respect to the existing market for legal services, the total number of humans needed to service the current demand for legal services is simply going to decline.” (footnote omitted)).

91. *See, e.g.*, Erb, *supra* note 36 (“According to traditional wisdom, the best way to avoid being replaced by a robot was to get an education and land a job that doesn't rely on manual labor, the employment sector viewed most at risk. But as it turns out, robots can do anything. Even lawyering.”).

92. *See infra* Subpart II.B.

93. *What's Next for Artificial Intelligence*, *supra* note 69 (quoting Yann LeCun as saying “Despite these astonishing advances, we are a long way from machines that are as intelligent as humans—or even rats. So far, we've seen only 5% of what AI can do.”).

94. Moss, *supra* note 28, at 84.

95. Remus & Levy, *supra* note 20, at 536.

DoNotPay relied on lawyers when designing its self-help service to initiate immigration applications, and some anticipated that the service would increase clients for human lawyers after refugees were brought into the system.⁹⁶ Similarly, self-help services like LegalZoom, LegalShield, and Rocket Lawyer have begun to contract with lawyers or otherwise enable consumers to connect with human lawyers when needed.⁹⁷ Others have noted that as more clients are brought into the legal system, there could even be increased hiring of certain indispensable human legal service providers, if appropriate funding is part of broader investment in technology.⁹⁸ Moreover, if expanding legal services to untapped markets does not occur as the result of profession-wide efforts to fulfill a professional responsibility to improve access to justice, it will likely occur out of economic necessity.⁹⁹

However, there is a risk that not all lawyers will benefit equally from the rise of AI. As McGinnis and Pearce explain:

Machines may actually aid two kinds of lawyers in particular. First, superstars in the profession will be more identifiable and will use technology to extend their reach. Second, lawyers who can change their practice or organization to take advantage of lower cost inputs made available by machines will be able to serve an expanding market of legal services for middle-class individuals and small businesses, meeting previously unfulfilled legal needs.¹⁰⁰

So, while some access will be increased as a result of (1) “superstars” extending their reach, and (2) versatile practices adjusting their services to the middle class, less high profile and less versatile lawyers—like public defenders—will likely not be able to implement AI as quickly, if at all. At the speed at which AI is developing, this could be detrimental and put significant portions of the profession at a competitive disadvantage.¹⁰¹

96. Cresci, *supra* note 23 and accompanying text.

97. Bigda, *supra* note 40, at 407–08 (“LegalZoom is beginning to offer legal advice for clients by contracting lawyers from different states. . . . Rocket Lawyer provides an ‘On Call’ service for its monthly subscribers, which allows customers to consult with attorneys from around the country. . . . LegalShield is implementing new technology into their platform by allowing clients to work with an attorney through the client’s smartphone.” (footnotes omitted)).

98. Magnuson & Frank, *supra* note 51, at 18 (“Increased funding for the justice system ensures not only increased technological resources available to the poor, but also adequate court staffing and the availability of legal service providers such as legal aid and public defenders, who are indispensable in filling the client-service gaps that evolving court processes and burgeoning technology create.”).

99. See Katz, *supra* note 73, at 963 (“Without tapping previously untapped markets (and there is good reason to believe they can be tapped), law is an otherwise mature industry whose total labor market participation will likely never exceed its prior peak.”).

100. McGinnis & Pearce, *supra* note 12, at 3042.

101. Katherine Medianik, Note, *Artificially Intelligent Lawyers: Updating the Model Rules of Professional Conduct in Accordance with the New Technological Era*, 39 CARDOZO L. REV. 1497, 1506 (2018) (“The[] elements of implementing AI technology generate margins superior to competing firms, thereby creating a competitive advantage.”).

Ultimately, certain parts of the legal profession are likely to see some degree of integration of machine learning into everyday practice.¹⁰² If this integration engages lawyers, clients, and the public without creating massive inequality of access, it could be beneficial to all parties, and could calm fears of robots completely replacing human lawyers while also discouraging extreme reactions such as prohibiting certain forms of AI innovation that could help the public at large.

An examination of current forms of specific soft AI making their way into the profession, with an eye toward the more advanced iterations to come, will help identify some specific challenges that must be comprehensively considered to ensure responsible implementation.

C. SIGNIFICANT CHARACTERISTICS OF DEPLOYED AND DEVELOPING FORMS OF AI IN LAW PRACTICE

In many ways, soft AI is already part of the legal profession, perhaps making the biggest impact in the areas of document review, e-discovery, legal research, and, increasingly, outcome prediction.

Document review is perhaps the task most obviously suitable for basic use of soft AI-based assistance. But AI is not only changing the speed and accuracy of review during critical initial stages of a case, it is also changing the very nature of this process.¹⁰³ Whereas document review was previously tasked to young associates, AI is drastically reducing the need for human hours to be spent on this task.¹⁰⁴ If effectively implemented, AI “can aggregate data and match a finite set of outcomes to the answers to questions” or “rely on data sets to provide answers as well as products from automated letters to document review, all with a few clicks of a mouse.”¹⁰⁵ AI’s ability to transform the task of document review is indicative of its potential to impact other, more complex tasks, like e-discovery.

Electronic discovery (“e-discovery”), is another historically laborious and increasingly expensive task where soft AI is making a tremendous impact.¹⁰⁶ Historically, e-discovery has been defined as “the process by which computers

102. Katz, *supra* note 73, at 963 (“For white-collar professions such as law, medicine, or finance, the medium-term future centers on a mixture of humans and machines working together to more efficiently deliver the services than either could alone.”).

103. *Id.* at 947 (“In short, while the existing methods differ and a significant number of technical questions still remain unanswered, document review . . . as we currently know it is about to be substantially reset.”).

104. *See id.* at 944 (“In the ‘golden days’ of document review, the days prior to the proliferation of electronically stored information, law firms would execute manual review of paper documents using teams of young associates.”).

105. Erb, *supra* note 36.

106. *See* Katz, *supra* note 73, at 942–43 (“The total cost of litigation is driven by a number of factors: lawyers, expert witnesses, investigators, employee time and distraction, and to an ever-increasing extent the costs of discovery.”).

search a database for keywords that lawyers agree are marks of relevance.”¹⁰⁷ AI is altering this definition as e-discovery moves toward predictive coding practices.¹⁰⁸ AI and effective use of algorithms can now predict how relevant a particular document is, much faster and more accurately than a human acting alone.¹⁰⁹ The more accurate and less expensive e-discovery becomes, the more prevalent the practice will eventually be within the profession.¹¹⁰

However, there are two practical aspects of AI in e-discovery that could slow or even inhibit closing the access to justice gap. One is that the benefits of AI-driven e-discovery might, at least at first, only be recognized by large firms because many smaller practices lack designated e-discovery units.¹¹¹ Another is that lawyers have not themselves been involved in the technological innovation in the area of e-discovery, and, despite the fact that discovery is a highly legal process, have outsourced the task to third parties.¹¹² The reliance on third-party innovation, with no lawyer involvement during design, could lead to ethical challenges as AI continues to advance.¹¹³

Legal research, like that performed by ROSS,¹¹⁴ is in tremendously high demand, and becoming highly sophisticated very quickly. Simply put, AI can help predict which past cases will be helpful to a lawyer’s case. This is becoming an increasingly greater departure from current database searches where a tool, such as Lexis or Westlaw, returns search results that a lawyer must then read, analyze, and Shepardize or KeyCite. As McGinnis and Pearce have explained, “in the past forty years, legal computer programs have perfected only keyword searches. However, because of technological acceleration, in less time computers will be able to pick and choose for themselves the best precedent to

107. McGinnis & Pearce, *supra* note 12, at 3047 (citing Steven C. Bennett, *E-Discovery by Keyword Search*, 15 PRAC. LITIGATOR 7, 9 (2004)).

108. Katz, *supra* note 73, at 945 (“We now stand on the cusp of the next generation of e-discovery centered around ‘predictive coding’ technology” (footnote omitted)).

109. See McGinnis & Pearce, *supra* note 12, at 3047 (noting that technicians can construct algorithms that predict whether a document is relevant from a large set of documents, thereby increasing the range of documents reviewed and the speed at which they are reviewed).

110. See *id.* at 3047–48 (in fact, McGinnis & Pearce note that e-discovery is already changing the discovery practice of large commercial litigation).

111. See, e.g., Katz, *supra* note 73, at 945 (“We now stand on the cusp of the next generation of e-discovery centered around ‘predictive coding’ technology, which should reduce costs to clients and in turn increase profits to high-performing law firms and legal product companies engaged in the enterprise.” (footnotes omitted)); McGinnis & Pearce, *supra* note 12, at 3048 (describing that only “large law firms have set up e-discovery units within their firms”).

112. See McGinnis & Pearce, *supra* note 12, at 3048 (“[L]awyers will face competition from companies outside the profession that want to offer discovery services to lawyers . . . [that] are likely more innovative, specialized, and less attached to traditional ways of thinking about the issue.”); Katz, *supra* note 73, at 944 (“[L]aw firms—and their clients—have not been uniformly innovative in response to the new world of e-discovery.”); Spahn, *supra* note 40 (“Using artificial intelligence can amount to ‘outsourcing’ work to the third-party artificial intelligence vendor.”).

113. See *infra* Subpart III.A.3.

114. See Vanderbilt University, *supra* note 9 and accompanying text.

cite in a brief.”¹¹⁵ Early legal AI companies like ROSS are transforming legal research with the help of IBM’s Watson technology, which, as McGinnis and Pearce explain, signals a significant shift from the use of keywords to semantics that match not only words, but similar concepts.¹¹⁶ Not only can these new systems match relevant cases, but they can gauge their relative persuasiveness based on how frequently other cases rely on it, and can do so within the context of certain courts or judges.¹¹⁷

As helpful as this technology will be in assisting lawyers with traditional practice processes, in the near future it will also likely fundamentally transform the way lawyers approach legal research. “Machine intelligence will not only uncover precedent but will also *guide lawyers’ judgments about the use of precedent*, as most lawyers can neither comprehensively evaluate the strength of precedent [n]or recall all possible precedents to mind.”¹¹⁸

Other commentators have noted that emerging services “purport not just to review documents and do word searches, but to give advice or something that is tantamount to advice.”¹¹⁹ Of course, AI can only guide a lawyer’s judgment based on the observational, “datafied” information it has at its disposal. To get the full picture of a set of facts, issue, or case, a lawyer will either have to account for un-datafied information wholly apart from the AI’s analysis, or begin to datafy that information for the AI to utilize.

The fact that AI tools will increasingly be able to help guide lawyers’ judgment in developing a case represents a monumental shift from the impact of previous technologies, which merely aided efficiency, and makes the design and responsible use of these systems even more critical.

Outcome prediction is what much of AI’s use in law is—and will continue to be—focused on.¹²⁰ Clients and lawyers both want to know whether to pursue a particular case, and if they do, they want to know what strategy has the greatest chance of success. As Katz explains in his article on “quantitative legal prediction:”

[Much of a lawyer’s work] can be substantially aided through the use of data, metrics, and models. Whether sourcing a particular legal matter, determining the outcome of a given piece of litigation, or forecasting the long-run implications of a given contract provision, the core questions involve matters of prediction.¹²¹

115. McGinnis & Pearce, *supra* note 12, at 3046.

116. *Id.* at 3049.

117. *See id.* (noting that machine intelligence will also make judgments about the strength of precedent and will help gauge the strength of legal precedent as it is tested in subsequent case law).

118. *Id.* at 3049–50 (emphasis added).

119. Connell, *supra* note 41, at 7.

120. *See* McGinnis & Pearce, *supra* note 12, at 3045 (“[A]ll machine-driven legal services will use sophisticated algorithms both to structure data in various forms, such as legal documents, and to make predictions about future events, like case outcomes.”).

121. Katz, *supra* note 73, at 948.

Others, such as McGinnis and Pearce, have described this process as “predictive analytics,” noting that “law, with its massive amounts of data from case law, briefs, and other documents, is conducive to machine data mining that is the foundation of this new predictive science.”¹²²

However, although some aspects of legal tasks are well-suited for outcome prediction and data mining, such practices have their limits. As McGinnis and Pearce acknowledge, “[l]egal data include fact patterns, precedents, and case outcomes,”¹²³ all of which can be mined. However, in reality, lawyers consider much more information in crafting a case, much of which is never documented and therefore not available for machine analysis.¹²⁴ Moreover, even the data that are available are often biased, or subject to the biases of the algorithms designed for certain types of practice or clients, in addition to the often unconscious biases of the algorithm designers themselves.¹²⁵

Individual lawyers, firms, and AI designers cannot confront these challenges on their own. It will take a profession-wide effort—one that involves lawyers and AI designers and takes into account the public’s needs and the preferences and expectations of clients—to maximize the benefits of AI in light of these risks. Although legal ethics oversight bodies have issued guidance in the last few years regarding certain forms of emerging technology, and even amended some rules to take into account the challenges posed by such technologies, these efforts will only be of limited use if applied to the unique challenges posed by AI. The following section examines the guidance to date regarding ethical obligations in light of new technologies, and identifies areas where ethics bodies and bar authorities should immediately strive to foster dialogue and issue additional guidance specific to the unique challenges of AI.

III. CONFRONTING AI’S CHALLENGES THROUGH A RENEWED COMMITMENT TO ETHICAL OBLIGATIONS

A. CURRENT ETHICAL GUIDANCE IS INSUFFICIENT TO ADDRESS THE UNIQUE CHALLENGES POSED BY AI IN LAW PRACTICE

Technology has always caused tension when reconciling lawyers’ ethical obligations. For many reasons, law firms have historically been slow to adopt new technologies.¹²⁶ The paramount obligation to protect confidential information, among other justifications, has kept lawyers from initially adopting many forms of technology, including email and computers. These conservative

122. McGinnis & Pearce, *supra* note 12, at 3052.

123. *Id.*

124. See Katz, *supra* note 73 and accompanying text.

125. See *supra* notes 66–68 and accompanying text.

126. See Mark A. Cohen, *Lawyers and Technology: Frenemies or Collaborators?*, FORBES (Jan. 15, 2018, 5:56 AM), <https://www.forbes.com/sites/markcohen1/2018/01/15/lawyers-and-technology-frenemies-or-collaborators/#17e53ace22f1> (arguing that lawyers have a “curious ambivalence” towards technology and are often reticent to embrace it professionally).

tendencies run counter to the parallel duties of zealously and competently representing clients. There is an emerging consensus, especially within the context of cybersecurity, that lawyers cannot take an “ostrich-with-its-head-in-the-sand” approach to technology—both in terms of using and not using various forms.¹²⁷

Over time, much of the profession has embraced, at times reluctantly, various forms of technology, and in doing so, has confronted various ethical dilemmas that can result from its use. In 2012, the American Bar Association (“ABA”), after a resolution of its Commission on Ethics 20/20, amended the black letter and commentary of several key model rules in order to take into account the increased role of technology in the profession.¹²⁸ This guidance will be modestly helpful, but ultimately insufficient, to address the challenges posed by AI in law practice. Even so, understanding the substance of the rules and the reasoning behind recent amendments is critical for context when determining the appropriate course for confronting the unique challenges posed by AI, as well as ensuring its potential to improve access to justice.

1. Competence

The meaning of “competent practice” fundamentally changes when a lawyer uses AI that performs increasingly sophisticated tasks, especially when that lawyer does not understand how the underlying technology works. Lawyers are not alone when it comes to failing to comprehend what is happening in the “black box” of AI.¹²⁹ Even many developers do not fully understand the AI they are designing.¹³⁰ But unlike individuals in other professions, lawyers have an ethical obligation that should be interpreted to require them, to some degree, to

127. See, e.g., JILL D. RHODES & VINCENT I. POLLEY, *THE ABA CYBERSECURITY HANDBOOK: A RESOURCE FOR ATTORNEYS, LAW FIRMS, AND BUSINESS PROFESSIONALS* 64 (2013) (“In short, a lawyer cannot take the ‘ostrich’ approach of hiding his head in the sand and hoping that his office or firm will not suffer a data breach that compromises client information. [Instead,] lawyers must implement administrative, technical, and physical safeguards to meet their obligation to make reasonable efforts to protect client information.”).

128. See generally *ABA Commission on Ethics 20/20*, A.B.A. (Mar. 18, 2013), [https://www.americanbar.org/groups/professional_responsibility/committees_commissions/standingcommittee_on_professionalism2/resources/ethics2020homepage/?q=&fq=\(id%3A%5C%2Fcontent%2Faba-cms-dotorg%2Fen%2Fgroups%2Fprofessional_responsibility%2F*\)&wt=json&start=0](https://www.americanbar.org/groups/professional_responsibility/committees_commissions/standingcommittee_on_professionalism2/resources/ethics2020homepage/?q=&fq=(id%3A%5C%2Fcontent%2Faba-cms-dotorg%2Fen%2Fgroups%2Fprofessional_responsibility%2F*)&wt=json&start=0) (“Created by then ABA President Carolyn B. Lamm in 2009, the Commission will perform a thorough review of the ABA Model Rules of Professional Conduct and the U.S. system of lawyer regulation in the context of advances in *technology* and the global legal practice developments.” (emphasis added)).

129. Charles McLellan, *Inside the Black Box: Understanding AI Decision-Making*, ZDNET (Dec. 1, 2016), <http://www.zdnet.com/article/inside-the-black-box-understanding-ai-decision-making/> (“Artificial intelligence algorithms are increasingly influential in peoples’ lives, but their inner workings are often opaque.”).

130. Simon, *supra* note 49 (“Even many of the experts who develop these products don’t fully understand them.” (citing Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES MAG. (Nov. 21, 2017), <https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html> (“As machine learning becomes more powerful, the field’s researchers increasingly find themselves unable to account for what their algorithms know—or how they know it.”))).

find out.¹³¹ There is currently no formal guidance for practicing competently in light of these emerging services.¹³²

Under the ABA's pre-2012 Model Rules, and still in some states, the competence rule language and accompanying commentary are simple. The rule merely states: "[c]ompetent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation,"¹³³ with commentary adding that, "[t]o maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice."¹³⁴ Although this might implicitly include keeping abreast of technology's benefits and risks, the ABA's 2012 resolution expressed that "it is important to make this duty explicit because technology is such an integral—and yet, at times invisible—aspect of contemporary law practice."¹³⁵ The ABA, accordingly, explicitly amended the commentary language to read that, "[t]o maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, *including the benefits and risks associated with relevant technology*."¹³⁶ This explicit addition of technology to the rule was and is significant. So too, though, is the fact that many states have chosen not to adopt this amendment¹³⁷—a sign that some states may also resist any rule changes that explicitly take into account AI, out of fear of being too prescriptive. Some commenters, on the other hand, have called for an even more prescriptive competence rule in light of AI's unique challenges.¹³⁸

Under the amended language, and arguably even the original language, the competence rule has been interpreted as saying that lawyers must understand not only the technical aspects of the technology they adopt, but also the related ethical implications. In the context of e-discovery, predictive coding, and computer assisted review, one commenter has noted that, practically speaking, "[t]his provision will require lawyers to better understand any advances in

131. See Baker, *supra* note 67, at 558 (arguing that the ethical "Duty of Technology Competence" should extend to the use of algorithms in law).

132. See *id.* ("A technology that has not yet been formally interpreted to apply to the Duty of Technology Competence is the use of algorithms in law.").

133. MODEL RULES OF PROF'L CONDUCT r. 1.1 (AM. BAR ASS'N 2002).

134. MODEL RULES OF PROF'L CONDUCT r. 1.1 cmt. 6 (AM. BAR ASS'N 2002).

135. ABA Resolution 105A, A.B.A. (Aug. 6–7, 2012), http://www.americanbar.org/content/dam/aba/directories/policy/2012_hod_annual_meeting_105a.doc [hereinafter *ABA Resolution 105A*].

136. Compare MODEL RULES OF PROF'L CONDUCT r. 1.1 cmt. 8 (AM. BAR ASS'N 2016) (emphasis added to show added language), with MODEL RULES OF PROF'L CONDUCT r. 1.1 cmt. 6 (AM. BAR ASS'N 2002) (the 2012 amendments added the "including the benefits and risks associated with relevant technology" language).

137. Robert Ambrogio, *Make That 30 States, as Another Adopts Ethical Duty of Technology Competence*, LAW SITES BLOG (Mar. 14, 2018), <https://www.lawsitesblog.com/2018/03/make-30-states-another-adopts-ethical-duty-technology-competence.html> (noting that, as of March 2018, twenty states had not adopted the ABA's amended language instituting a "technological competence" obligation); Baker, *supra* note 67, at 561–62 (noting that "thirty-one states have adopted the Duty of Technology Competence by amending the respective Duty of Competence" by citing each state's respective action).

138. See, e.g., Medianik, *supra* note 101, at 1515 ("[S]tate competency rules shadowing the Model Rules, [], remain too ambiguous to lend an adequate sense of direction for lawyers using AI technology.").

technology that genuinely relate to competent performance of the lawyer's duties to a client."¹³⁹ The most obvious of these other ethical duties is maintaining confidentiality.¹⁴⁰

As AI becomes more prevalent, lawyers might implicitly be required, as the rule has been understood, to exercise "continued vigilance and learning as technology advances, in order to comply with a lawyer's duties under ethics rules."¹⁴¹ Although this rule imposes a positive duty on lawyers to understand the technology they use, their knowledge must only be sufficient to competently *use* the technology. Guidance thus far has not required or even suggested that lawyers should be involved in the design phase of new technologies they use, for example, the way DoNotPay involved lawyers during its recent expansion.¹⁴² Rather, the rule and guidance merely imply that outside experts can help a lawyer become competent or act competently in certain circumstances.¹⁴³

Competence in an era of AI should require a lawyer to either be involved in the design of the AI systems they are using, or at the very least, to understand (with the help of an expert, if needed) certain underlying characteristics that affect the AI's bias (including that of the design, designer, and data),¹⁴⁴ its limits (including the limits of observational data and exclusion of information which has not been "datafied"),¹⁴⁵ and its confidentiality concerns.¹⁴⁶

2. Confidentiality

The emergence of AI in law practice should fundamentally change the way lawyers think about, talk about, and take measures to protect client confidentiality. This is due in large part to the new ways that client information will be generated, used, stored, and in some cases, comingled with that of other clients.

Confidentiality, especially when it comes to new technology, is at the core of a lawyer's ethical obligations.¹⁴⁷ With limited exceptions, confidentiality

139. John M. Barkett, *More on the Ethics of E-Discovery: Predictive Coding and Other Forms of Computer-Assisted Review* (2012) (unpublished manuscript), https://judicialstudies.duke.edu/sites/default/files/centers/judicialstudies/TAR_conference/Panel_5-Original_Paper.pdf.

140. See *infra* Subpart III.A.2; see also RHODES & POLLEY, *supra* note 127, at 65 ("[A] lawyer's ethical obligation of competence requires that the lawyer become and remain competent about the technology they use so as to be able to protect client confidential information.").

141. RHODES & POLLEY, *supra* note 127, at 66.

142. See *supra* notes 22–23 and accompanying text.

143. RHODES & POLLEY, *supra* note 127, at 66 ("If a lawyer is not competent to decide whether use of a particular technology (e.g., cloud storage, public Wi-Fi) allows reasonable measures to protect client confidentiality, the ethics rules require that the lawyer must get help, even if that means hiring an expert information technology consultant to advise the lawyer.").

144. See *supra* notes 66–68 and accompanying text.

145. See *supra* note 76 and accompanying text.

146. See *infra* Subpart III.A.2.

147. See David G. Ries, *Cyber Security for Attorneys: Understanding the Ethical Obligations*, L. PRAC. TODAY (Mar. 2012), http://www.americanbar.org/content/dam/aba/publications/law_practice_today/cyber-security-for-attorneys-understanding-the-ethical-obligations.authcheckdam.pdf.

rules typically provide that “[a] lawyer shall not reveal information relating to the representation of a client unless the client gives informed consent.”¹⁴⁸ In recent years, especially in the context of cybersecurity, commenters have stressed the importance of renewed commitment to confidentiality, including in the *ABA Cybersecurity Handbook*, which explains that the “obligation to maintain confidentiality of all information concerning a client’s representation, no matter the source, is paramount.”¹⁴⁹

The language and interpretation of confidentiality rules have trended in a stricter direction in recent years. Until 2012, and as is still the case in some states, Rule 1.6’s black letter only contained a negative obligation to avoid actively revealing client information.¹⁵⁰ The commentary, on the other hand, has suggested a more positive obligation, explaining that “[a] lawyer must *act competently* to safeguard information relating to the representation of a client . . . against inadvertent or unauthorized disclosure,”¹⁵¹ and that “[w]hen transmitting a communication that includes information relating to the representation of a client, the lawyer must *take reasonable precautions* to prevent the information from coming into the hands of unintended recipients.”¹⁵² When it comes to communicating client information via email or a cloud-based service, rule commentary also notes that, absent special circumstances, no special security measures are needed if the communication method affords a “reasonable expectation of privacy,” which is determined by “the sensitivity of the information and the extent to which the privacy of the communication is protected by law or by a confidentiality agreement.”¹⁵³ One challenge with AI will be determining what a client’s and lawyer’s reasonable expectations of privacy are with such a rapidly developing technology, especially in light of the “black box” within which the intelligence often operates.¹⁵⁴ This makes a lawyer’s competent understanding, and ability to communicate that understanding,¹⁵⁵ all the more critical.

In 2012, the ABA wanted the affirmative obligations to safeguard client information to be more explicit within the confidentiality rule, expressing in its resolution that “technological change has so enhanced the importance of this duty that it should be identified in the black letter and described in more detail in [the commentary].”¹⁵⁶ Accordingly, Rule 1.6 of the ABA Model Rules, and the adopted rule in some states, now provides: “A lawyer *shall make reasonable*

148. MODEL RULES OF PROF’L CONDUCT r. 1.6(a) (AM. BAR ASS’N 2016).

149. RHODES & POLLEY, *supra* note 127, at 62.

150. See *ABA Resolution 105A*, *supra* note 135, at 4 (Model Rule 1.6(c) now imposes an affirmative obligation to maintain confidentiality).

151. MODEL RULES OF PROF’L CONDUCT r. 1.6 cmt. 18 (AM. BAR ASS’N 2016) (emphasis added).

152. MODEL RULES OF PROF’L CONDUCT r. 1.6 cmt. 19 (AM. BAR ASS’N 2016) (emphasis added).

153. *Id.*

154. See *supra* notes 129–130 and accompanying text.

155. See *supra* Subpart III.A.1; *infra* Subpart III.A.4.

156. *ABA Resolution 105A*, *supra* note 135, at 14.

efforts to prevent the inadvertent or unauthorized disclosure of, or unauthorized access to, information relating to the representation of a client,”¹⁵⁷ with Comment 18 now laying out the “[f]actors to be considered in determining the reasonableness of the lawyer’s efforts,” which include:

the sensitivity of the information, the likelihood of disclosure if additional safeguards are not employed, the cost of employing additional safeguards, the difficulty of implementing the safeguards, and the extent to which the safeguards adversely affect the lawyer’s ability to represent clients (e.g., by making a device or important piece of software excessively difficult to use).¹⁵⁸

Whereas with technology like cloud computing some sensitive information can be withheld from third-party storage if the lawyer determines that it is better suited for storage on the firm’s premises or in paper form, AI relies on constant access to critical information.¹⁵⁹ Therefore, withholding certain data from AI systems could undermine the effectiveness of a service assisting with tasks that assist with case development, legal research, or argument development and drafting.

On a more fundamental level, the post-2012 rules, interpretations, and guidance are almost entirely focused on security. While important, security alone does not represent the full extent of confidentiality concerns with AI. Even under the 2012 amendments, the ethics rules have been interpreted to focus on *disclosure* of information, contemplating something like a breach of a cloud service. The emergence of AI will certainly magnify security challenges,¹⁶⁰ but it will also change the way client information is gathered, datafied, formatted, and used, such that keeping unwanted eyes off of a stored document will no longer be sufficient to ensure that a client’s confidences are protected in the ways that they would expect.

Protecting confidentiality in an era of AI must go beyond merely ensuring security and must include competently understanding how AI systems work, communicating with clients (and former clients)¹⁶¹ to understand their expectations and preferences, and ensuring that the designers and managers of AI systems, including third parties, understand the critical importance of confidentiality.

157. MODEL RULES OF PROF’L CONDUCT r. 1.6(c) (AM. BAR ASS’N 2016) (emphasis added).

158. MODEL RULES OF PROF’L CONDUCT r. 1.6 cmt. 18 (AM. BAR ASS’N 2016).

159. See Simon, *supra* note 49 (“Most AI products, such as . . . cite-checking products . . . require access to your confidential data. (A draft memo itself is confidential information, for example.) This raises a lot of questions about confidentiality.”).

160. See *id.* (“What happens to your confidential data once the AI vendor gains access to it? Who has access to it at the AI vendor? Does the AI vendor share your confidential information with other third-party vendors? If so, do you know who those third-party vendors are, and have you checked them out? Do they have a contractual duty of confidentiality? What happens to your client’s data if the AI vendor is sold, merges, retires, or goes bankrupt? If the AI vendor is subpoenaed, is the vendor contractually obligated to give you notice so that you can intervene to challenge the subpoena?”).

161. See *infra* Subpart III.A.6.

3. *Supervising Third Parties*

AI services will frequently be, or at the very least involve, third parties. Indeed, the increased role of non-lawyers could play a major role in helping to improve access to justice. As Bill Henderson has noted:

Stated bluntly, the legal profession is becoming a subset of a larger legal industry that is increasingly populated by nonlawyers, technologists, and entrepreneurs. . . . Virtually every other aspect of a legal problem can be broken down into its component parts, reengineered, streamlined, and turned into a legal input or legal product that is better, cheaper, and delivered much faster.¹⁶²

The increased risks and interconnected nature of new technologies in the practice of law have prompted some review of the obligations of lawyers to supervise both the other lawyers with which they are associated, as well as third-party non-lawyers.

Model Rules 5.1 and 5.3 require supervisory attorneys to “make reasonable efforts to ensure that the firm has in effect measures giving reasonable assurance that,” under 5.1, “all lawyers in the firm conform to the Rules of Professional Conduct,”¹⁶³ and under 5.3, that the conduct of non-lawyers employed by, retained by, or associated with the lawyer, “is compatible with the professional obligations of the lawyer.”¹⁶⁴ The ABA recognized in 2012 that third-party assistance no longer involves just people, and in 2012 changed Rule 5.3’s title from “Responsibilities Regarding Nonlawyer *Assistants*,” to “Responsibilities Regarding Nonlawyer *Assistance*,”¹⁶⁵ with commentary now referencing “cloud computing” as a specific example of such a third-party service.¹⁶⁶ Similarly, the use of the phrase “nonlawyer,” as opposed to “person,” indicates “that the rule is intended to have reach beyond human assistants, to other nonlawyers, human or not, involved in the representation of a client.”¹⁶⁷ Within this context, “Artificial intelligence products are effectively non-human nonlawyers.”¹⁶⁸

While cloud computing is a valuable example of an emerging technological service that has undergone some helpful ethical scrutiny, AI’s role in a lawyer’s

162. William D. Henderson, *A Blueprint for Change*, 40 PEPP. L. REV. 461, 462–63 (2013).

163. MODEL RULES OF PROF’L CONDUCT r. 5.1 (AM. BAR ASS’N 2016).

164. MODEL RULES OF PROF’L CONDUCT r. 5.3(a)–(b) (AM. BAR ASS’N 2016).

165. Compare MODEL RULES OF PROF’L CONDUCT r. 5.3 (AM. BAR ASS’N 2002) (emphasis added), with MODEL RULES OF PROF’L CONDUCT r. 5.3 (AM. BAR ASS’N 2016) (emphasis added).

166. See MODEL RULES OF PROF’L CONDUCT r. 5.3 cmt. 3 (AM. BAR ASS’N 2016) (allowing a lawyer to use “an Internet-based service to store client information,” but when using such services outside the firm, “a lawyer must make reasonable efforts to ensure that the services are provided in a manner that is compatible with the lawyer’s professional obligations”).

167. David L. Gordon & Rebecca L. Ambrose, *The Ethics of Artificial Intelligence*, JACKSON LEWIS (May 11, 2017), https://www.jacksonlewis.com/sites/default/files/docs/Final_The%20Ethics%20of%20Artificial%20Intelligence_Gordon%20and%20Ambrose.pdf; see also Medianik, *supra* note 101, at 1522 (“[I]n representations involving AI technology, lawyers too have a responsibility to adequately supervise ROSS’s work since it carries out consequential tasks for client representation. If, however, lawyers blindly rely on ROSS’s outputs, they should be disciplined . . . because they would be breaching their fundamental obligations to their clients for failing to properly supervise a nonlawyer assistant.” (footnote omitted)).

168. Simon, *supra* note 49.

practice, including its ability to guide a lawyer's judgment, requires additional and urgent guidance. Even so, existing guidance regarding cloud computing is a useful starting point for understanding the context within which these important discussions must take place.

The *ABA Cybersecurity Handbook* defines cloud computing as "any system whereby a lawyer stores digital information on servers or systems that are not under the close control of the lawyer or the lawyer's firm."¹⁶⁹ This will undoubtedly encompass third-party AI services, including ROSS. State ethics boards' guidance on cloud computing is representative of the profession's general approach to striving for ethical use of new technology, which involves a baseline competence of how the technology works and a vetting of vendors for things like security in order to maintain confidentiality of client information.¹⁷⁰ Professor Simon suggests that under these rules, at the very least, lawyers implementing AI must "(1) hire an expert to vet the AI product; (2) learn what the AI product can (and can't) do; and (3) double-check the output of the AI product."¹⁷¹

However, a baseline technical understanding of AI, and ensuring that it is not malfunctioning, will not necessarily ensure that lawyers consider the myriad social, ethical, and moral issues that AI raises in the practice of law.¹⁷² In addition, although security is important, it will not be the only thing a lawyer needs to consider in evaluating whether an AI service will appropriately maintain client confidentiality. Moreover, current guidance does not raise critical ethical issues related to the duty to communicate with clients, the duties to exercise independent judgment and render candid advice, and lawyers' ongoing obligations to former clients, all of which are explained in greater detail below.

4. *Communicating with Clients*

Despite the enthusiasm of some attorneys when it comes to emerging AI, communicating with clients about AI in law practice is difficult,¹⁷³ especially considering how little most lawyers actually know about such services. The transformative role that AI will play in legal representation makes the communication between the lawyer and the client all the more essential to ensuring a productive, ethical representation.

169. RHODES & POLLEY, *supra* note 127, at 77.

170. *See id.* at 78 ("[State ethics opinions] make clear that a lawyer must have a basic understanding of the technical aspects of cloud computing, and should conduct a due diligence evaluation of the provider to ensure that they have adequate security measures.").

171. Simon, *supra* note 49.

172. *See infra* Subpart III.A.5.

173. *See* Marc Lauritsen, *Marketing Real Lawyers in the Age of AI*, L. PRAC., Jan./Feb. 2017, at 51 ("It's increasingly a no-brainer to use intelligent tools in law practice. Not so clear is how to talk about them with clients and prospective clients.").

Model Rule 1.4, which was unchanged in 2012, requires appropriate communication with clients “about the means by which the client’s objectives are to be accomplished.”¹⁷⁴ This has been interpreted to encompass communicating the ways in which a law practice utilizes technology, and even notifying clients when their information has been compromised.¹⁷⁵

Because AI in law—if it is to be used in a way that considers all of a client’s needs—will require gathering, datafying, formatting, and using especially sensitive client information in new ways, this communication with clients will be of paramount importance. Not only will lawyers need to discuss with clients the potential risks to their information, but also the fundamental nature of AI as a means of assisting with the representation—one that either has severe limitations (because many client needs are not datafied, and therefore not considered by the machine intelligence), or which makes very complex use—with third parties—of especially sensitive new data, not previously datafied. Because it is not yet clear what clients will prefer if faced with this choice, it is all the more important that lawyers are explicitly responsible for considering these realities before adopting a service, and for being able to competently discuss such implications with their clients.

It is important to note that, in the same way that certain ethical dilemmas should not be fatal to some forms of legal self-help,¹⁷⁶ mere tension between ethical obligations as a result of AI should not preclude a lawyer or firm from implementing a potentially transformative and beneficial AI service. What is critical is that such decisions weigh client needs and preferences in determining how to proceed in light of the ethical tensions, and are made in consultations between the lawyer and client in which both parties are adequately informed about the specific nature of the AI involved in their case.

AI guidance should extend this principle as it has been articulated with regard to security in the commentary to most confidentiality rules, which states that, “[a] client may require the lawyer to implement special security measures not required by this Rule or may give informed consent to forgo security measures that would otherwise be required by this Rule.”¹⁷⁷ It is difficult to anticipate with confidence how clients will respond to an AI driven ecosystem. Will the risks, limits, complexities, and unknowns of AI be such that clients prefer their lawyers to forego its use in some or all of their legal matters? Or, will the increased efficiency and potential quality of service lead to a client’s ringing endorsement of such services? Lawyers will not know unless they ask.

174. MODEL RULES OF PROF’L CONDUCT r. 1.4(a)(2) (AM. BAR ASS’N 2016) (note that alterations to the comments were made).

175. See Ries, *supra* note 147 (expressing that Model Rule 1.4, Communications, requires keeping clients informed of any compromises of their confidential information). Of course, practically speaking, it is very difficult for lawyers to know when client information has been compromised. See generally Eli Wald, *Legal Ethics’ Next Frontier: Lawyers and Cybersecurity*, 19 CHAP. L. REV. 501 (2016).

176. See *supra* Subpart I.A.

177. MODEL RULES OF PROF’L CONDUCT r. 1.6 cmt. 18 (AM. BAR ASS’N 2016).

Especially in these critical years ahead, ethics guidance should specifically stress an obligation to foster these communications.

5. *Independent Judgment and Candid Advice*

As previously stressed, one of the major limitations of AI is its inability to take into account information beyond the observational data that it has at its disposal. Many pieces of information, including sensitive or embarrassing information concerning the client, the instinctual knowledge of the lawyer, and relevant non-legal factors that the AI might not have access to, are not currently or cannot be datafied. Guidance must stress that, consistent with the preferences of a client, this information, which drives a lawyer's professional judgment, must not be marginalized if AI is adopted.¹⁷⁸

This guidance should stress a lawyer's obligation under Model Rule 2.1, which was also unchanged in the ABA's Model Rules in 2012. The rule explains that, "[i]n representing a client, a lawyer shall exercise independent professional judgment and render candid advice," and that this might involve referring "not only to law but to other considerations such as moral, economic, social and political factors, that may be relevant to the client's situation."¹⁷⁹ This means, among other things, that lawyers must consider and address clients' non-legal needs, as well as their legal ones.

On a more abstract level, as lawyers become increasingly reliant on intelligent systems, it draws into question the extent to which their professional judgment is "independent."¹⁸⁰ This is especially true if they do not fully understand and were not involved with the design of the system, and therefore cannot make independent judgments based on the AI's output. Although early adopters of ROSS report that users have double checked the service's results by comparing them with other legal research platforms, it has also been reported that users are beginning to rely on ROSS's results without any crosschecks.¹⁸¹ "Given the lack of transparency and other issues with blindly relying on

178. See Catherine Nunez, *Artificial Intelligence and Legal Ethics: Whether AI Lawyers Can Make Ethical Decisions*, 20 TUL. J. TECH. & INTELL. PROP. 189, 204 (2017) ("It is clear that ROSS has been exceedingly useful in the legal research department. . . . However, an attorney's role is not merely research. Attorneys must utilize their research skills in conjunction with their individual professional and moral judgment. Answers to questions requiring either of the two require a certain human quality of which ROSS is yet equipped.").

179. MODEL RULES OF PROF'L CONDUCT r. 2.1 (AM. BAR ASS'N 2016).

180. Medianik, *supra* note 101, at 1517 ("In terms of implementing the work of an AI lawyer to a case, when a lawyer relies solely on ROSS's outputs, independent professional judgment—as required by Model Rule 2.1—vanishes because reliance on such outputs turns into dependence on the judgments of a technological apparatus.").

181. See *id.* at 1511 (noting that lawyers rely primarily on the searches performed by ROSS and interpret the results, but do not go back and "quality check" to ensure the search was accurate) (citing E-mail from William Caraher, Chief Info. Officer & Dir. of Operations, von Briesen & Roper, to Katherine Medianik, Student, Benjamin N. Cardozo Sch. of Law (Sept. 8, 2016, 11:57 AM)); see also Baker, *supra* note 67, at 558 ("[T]he research habits of this generation show an apt to rely on algorithms to generate results with little evaluation of those results.").

algorithms, lawyers may be at a loss as to how to competently use this ubiquitous technology.”¹⁸²

Moreover, although AI will be very good at tracking and analyzing documented legal inputs and producing potentially helpful outcomes based on past observations, the effect of moral, social, and political factors will be difficult to analyze, or even account for or program into the system in the first place. Indeed, as Remus and Levey observe:

[R]educing legal advising to legal prediction could threaten to impede the law’s development. Predictability and stability are of course critical rule-of-law values, but so too is democratic participation in lawmaking. A core way in which citizens participate is through their lawyers, who translate their interests into persuasive and sometimes novel arguments as to how the law should apply to their clients’ circumstances. Lawyers can do so because our legal system is about reasons as well as outcomes—reasons, asserted by lawyers and memorialized in judicial opinions, which provide a continual opportunity through which to debate and potentially change the law. If lawyering is replaced by computer prediction, we will shift to a system that is more about outcomes than reasons—and outcomes that are inescapably “informed by the world as it was in the past, or, at best, as it currently is.”¹⁸³

Of course, over time, lawyers might experiment with ways in which AI might be able to take more of these factors into account, especially if law firms or third-party AI service providers begin tracking how such information has been handled—and to what success—in the past. This gives rise to the final obligation that should be urgently stressed in guidance regarding the adoption and use of AI—a lawyer’s obligations to former clients.

6. *Obligations to Former Clients*

ABA Model Rule 1.9(c), which has been adopted by most states, provides that duties such as confidentiality extend to the data of former clients.¹⁸⁴ AI will be powerful—indeed exponentially powerful—because it leverages information from many different cases from which new inputs can identify analogous points to create helpful outcomes, whether in the form of a suggested case to read or a suggested format of an argument or overall legal strategy based on past favorable outcomes. Whereas in days past lawyers might have shredded or deleted client information at some point, there is no longer an incentive—and in fact there is actually a disincentive—to dispose of any client information today. Because AI performs better with the more data it has access to,¹⁸⁵ client information could remain not only in existence, but remain in use, indefinitely.

182. Baker, *supra* note 67, at 572 (footnote omitted).

183. Remus & Levey, *supra* note 20, at 548–49 (footnote omitted).

184. MODEL RULES OF PROF’L CONDUCT r. 1.9(c) (AM. BAR ASS’N 2016).

185. Enrique Dans, *From Data to Artificial Intelligence*, MEDIUM (Feb. 5, 2017), <https://medium.com/enrique-dans/from-data-to-artificial-intelligence-491bdd92400> (“Data is the gasoline that powers artificial intelligence. Data allows us to develop the best algorithms, and above all, to improve them over time so that they produce better results and adapt to changing conditions. . . . The biggest mistake that can be

Ethics rules are designed, in part, to ensure that clients feel they can be candid with their attorneys throughout the course of a representation. The idea that an increasing amount of sensitive information will not only be collected, but also used, in perpetuity, could threaten this coveted comfort and trust that fosters critical openness. Many things must happen to preserve the trust between clients and attorneys in the age of AI, including making sure that this information is secure. But if clients are going to trust their attorneys, they are also going to have to trust AI as a tool, highlighting again the need for competence, communication, and the rest of the obligations that have been outlined.

It is time for lawyers to confront these challenges, wrestle with the ethical tensions, and through their ethics oversight bodies and bar associations issue guidance that will help the profession responsibly and ethically integrate AI into practice in a way that will improve the effectiveness of lawyers across the industry and will increase access to justice.

B. NEEDED GUIDANCE CONCERNING THE DESIGN, ADOPTION, AND USE OF AI IN LAW PRACTICE

The best place to begin to address these challenges is in forums that issue guidance through formal or informal ethics opinions. Some commentators advocate for amendments to the ABA Model Rules, or their commentary, that take into account the unique challenges posed by AI.¹⁸⁶ Indeed, if AI continues to progress in the profession without guidance, more prescriptive oversight might be necessary to respond to the possible negative consequences articulated in this Article. However, in the near term, one of the advantages of guidance over rule amendments is that guidance can be issued more quickly than actual changes to the black letter or commentary of ethics rules or, in more extreme case, changes to the law.¹⁸⁷

made in artificial intelligence is to try to judge an algorithm by its results the moment we get it, without taking into account the progress that can be made by using more and better data.”).

186. See, e.g., Bigda, *supra* note 40, at 412 (“Due to the increased use of artificial technology [sic] within the legal community, new laws and rules of professional conduct must be written to regulate the use of artificial intelligence in replacing lawyers.”); Medianik, *supra* note 101, at 1502 (advocating for, among other things, “the addition of several comments that incorporate AI technology and account for technological advancement”).

187. Some AI legislation currently being discussed in academic literature would likely affect emerging legal AI services. For example, the proposed Artificial Intelligence Development Act “would create a federal agency tasked with certifying the safety of AI systems,” and “would create a liability system under which the designers, manufacturers, and sellers of agency-certified AI programs would be subject to limited tort liability, while uncertified programs that are offered for commercial sale or use would be subject to strict joint and several liability.” Medianik, *supra* note 101, at 1508–09 (quoting Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 393 (2016)). Services like ROSS would likely be subject to such oversight. *Id.* at 1508.

1. *The Need for Urgency*

AI is likely to be adopted at a much faster,¹⁸⁸ but perhaps less uniform, rate than previous technologies that spurred calls for ethical guidance in the legal profession. Part of this is due to what has been described as AI's "exponential growth [that] confounds our intuition and expectations."¹⁸⁹ Katz explains the converging role of "the synergy of Moore's Law, Big Data, and the AI Revolution" in the legal profession by noting that "[w]ith each doubling of processor speed, halving of data storage costs, and major advances in machine learning, the possibility frontier is opening up and doing so at a drastically nonlinear rate."¹⁹⁰

This rate of development sets AI apart from other technologies adopted by the legal profession in the past,¹⁹¹ and makes design, proactive consideration of challenges, and ethical guidance all the more critical. It is likely that lawyers will have less time to confront ethical implications of AI than they have had with other technologies, because legal markets simply will not wait.¹⁹² Moreover, because of the potential ability of AI to help close the access to justice gap, the profession and indeed society cannot afford to wait years for amendments to rules or changes to law.¹⁹³

2. *Guidance Is Needed During Critical Design Stages of Early AI*

The coming years will be critical ones for the design of increasingly advanced AI that will continue to make its way into the legal profession. While many challenges presented by AI will depend on responsible use of these systems by lawyers, another critical front is ensuring that, as much as possible, the ethical values that guide lawyers are designed into the AI systems themselves. There are both practical and theoretical conceptions of how to go about designing values into AI, and all involve lawyers first understanding what those values are, specifically in light of AI's challenges. Oxford philosopher and AI expert Nick Bostrom stated:

188. See Katz, *supra* note 73, at 949 ("Whether the questions surround the financing of lawsuits or engaging in . . . predictions . . . it does not matter what you think ought to happen; it only matters what the relevant market will embrace. The market will (or already has) embraced this sort of technology and there is likely much more coming down the pipeline.").

189. ERIK BRYNJOLFSSON & ANDREW MCAFEE, RACE AGAINST THE MACHINE: HOW THE DIGITAL REVOLUTION IS ACCELERATING INNOVATION, DRIVING PRODUCTIVITY, AND IRREVERSIBLY TRANSFORMING EMPLOYMENT AND THE ECONOMY 19 (2011).

190. Katz, *supra* note 73, at 922.

191. See McGinnis & Pearce, *supra* note 12, at 3041 ("[C]ontinuous technological acceleration in computational power is the difference between previous technological improvements in legal services and those driven by machine intelligence.").

192. See Katz, *supra* note 73, at 949 (discussing the speed of adoption of AI technology in the legal community).

193. Some states are still considering, but have not yet adopted, the ABA's amendments to its Model Rules that take into account new technologies, which it adopted in 2012. See *supra* note 137.

We would want the AI we build to ultimately share our values, so that it can work as an extension of our will. It does not look promising to write down a long list of everything we care about. It looks more promising to leverage the AI's own intelligence to learn about our values and what our preferences are.¹⁹⁴

Unlike many sectors, the legal profession has in fact already “writ[ten] down a long list of everything we care about,” in the form of its rules of professional conduct. In order to incorporate these values into AI as much as possible, lawyers must first maximize their understanding of these values within the specific context of AI. In addition, because it might someday be possible for AI to determine and automatically implement human values within systems,¹⁹⁵ lawyers must ensure that they have thought about and are living these values every day. All of these fronts will be aided by guidance from robust discussion, debate, and guidance.

3. *The Need for Proactive, but Not Prescriptive, Guidance*

The most important component of guidance from ethics bodies concerning the design, adoption, and use of AI, is that it be proactive. Some scholars have acknowledged that AI, and particularly its predictive functions, will require some form of pre-deployment “validations.” Daniel Katz, in his article titled *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, cautions that, “As the field moves forward into greater use of prediction models, it is critical for [] validation efforts to be undertaken and demanded prior to their actual deployment in any real world application.”¹⁹⁶ Ethical guidance is the first step to ensuring that lawyers know what needs to be validated before undertaking these efforts.

State ethics oversight bodies are not averse to issuing such guidance when a transformative technology comes along. Again, the legal profession's treatment of cloud computing is a useful starting point in charting a possible path forward for confronting the challenges present by AI. The ABA provides an online guide to “Cloud Ethics Opinions Around the U.S.”¹⁹⁷ and *The ABA Cybersecurity Handbook* contains an appendix of “Ethics Opinions on Lawyer Confidentiality Obligations Concerning Cloud Computing.”¹⁹⁸

194. *What's Next for Artificial Intelligence*, *supra* note 69 (quoting Nick Bostrom, founding director of the Future of Humanity Institute at Oxford University).

195. Steven Kotler, *The Uncanniest Valley: What Happens When Robots Know Us Better than We Know Ourselves?*, FORBES (July 20, 2014, 1:22 PM), <https://www.forbes.com/sites/stevenkotler/2014/07/20/the-uncanniest-valley-what-happens-when-robots-know-us-better-than-we-know-ourselves/#1e070bd66d1d>.

196. Katz, *supra* note 73, at 942.

197. *Cloud Ethics Opinions Around the U.S.*, A.B.A., https://www.americanbar.org/groups/departments_offices/legal_technology_resources/resources/charts_fyis/cloud-ethics-chart.html (last visited Nov. 21, 2018).

198. RHODES & POLLEY, *supra* note 127, at 245.

Many state ethics opinions regarding cloud computing merely accept that the use of cloud services is ethical as long as lawyers competently select an appropriate vendor, preserve confidentiality, safeguard client property, provide reasonable supervision of cloud vendors, and communicate with the client as appropriate (in other words, it is ethical if it is ethical).¹⁹⁹ However, others have mandated that lawyers take significant steps that require substantial research and consideration before adopting services that might, for example, make client confidential information vulnerable to exposure. For example, Iowa requires lawyers to “[d]etermine the degree of protection the vendor provides to its clients’ data” before adopting a service, New Jersey requires lawyers to “[m]ake sure that vendors are using available technology to guard against foreseeable infiltration attempts,” and North Carolina requires lawyers to “[e]valuate the vendor’s security and backup strategy.”²⁰⁰ As one commentator has acknowledged in light of these various requirements, “It is probably safe to say that this subject matter does not form part of the curriculum at law schools.”²⁰¹ Nevertheless, under these jurisdictions’ guidance, the burden on lawyers to learn about the intricacies of the technology they are adopting and to consider the resulting ethical implications is outweighed by the unique challenges posed by the technology, and the importance of the legal ethics principles that the jurisdiction believes should not be undermined by the adoption of certain forms of technology. Guidance regarding the outlined challenges of AI is even more imperative.

Some guidance exists regarding how to design technology more broadly in a way that improves access to justice. For instance, Katherine Alteneder and Linda Rexer, in their article *Consumer Centric Design: The Key to 100% Access*, advocate for closing access gaps with “a consumer-centric approach in which consumers can be efficiently and effectively directed to the type and level of help they need” by maximizing “self-help services,” “building connections with providers,” employing methods of “simplification,” and “minding the digital divide.”²⁰² They argue that this “can maximize many emerging developments such as non-lawyer practice, enhanced unbundled legal services, Alternative Dispute Resolution (“ADR”) and online dispute resolution (“ODR”), remote legal services, and other innovations that give promise to a robust and integrated justice system.”²⁰³ This model provides a valuable starting point for improving the design of many legal technology services. However, AI’s unique challenges,

199. See, e.g., Ohio State Bar Ass’n, Informal Advisory Op. 2013-03 (July 25, 2013), <https://www.ohiobar.org/ForPublic/LegalTools/Documents/OSBAInfAdvOp2013-03.pdf>.

200. Drew T. Simshaw, *Legal Ethics and Data Security: Our Individual and Collective Obligation to Protect Client Data*, 38 AM. J. TRIAL ADVOC. 549, 565 (2015) (alterations in original).

201. Adam Cohen, *Lawyers Between a Rock (Social Media) and a Hard Place (The Cloud)*, INSIDE COUNSEL (Apr. 16, 2014), Proquest, Doc. No. 1516417190.

202. Katherine Alteneder & Linda Rexer, *Consumer Centric Design: The Key to 100% Access*, 16 J.L. SOC’Y 5, 7 (2014).

203. *Id.*

outlined in this Article, will require additional design guidance that focuses not only on connecting consumers with providers, but also on the inevitable challenges that will persist throughout all phases of the representation.

In issuing guidance regarding the design, adoption, and use of AI in an era where the role of humans lawyers will to some degree be marginalized, it will be essential to elevate the “humanistic” nature of lawyering, especially in legal education.²⁰⁴ Indeed, the proposed ethical guidance in this paper will not solve the larger moral questions or fundamental limitations of AI in the law. For instance, there are certain things that machines, no matter how well designed, will not be able to do as well as humans, such as create emotional bonds with clients that lead to better legal representation.²⁰⁵ However, in light of the immediate needs surrounding current rapid development and implementation, and especially in light of AI’s potential to help increase access to justice, issuing guidance concerning emerging AI services will enable the profession to address these larger issues as the sophistication of lawyers and clients regarding AI continues to grow.

CONCLUSION

Overturning parking tickets, improving lawyer efficiency, and reducing costs for law firm clients is just the beginning of AI’s potential in the legal profession. AI has the ability to expand access to legal services to parts of society that have historically been shut out. The demand for AI in the law is great, and the potential benefits are undeniable.

However, AI’s transformation of the legal profession will not be without challenges. Because the future of legal services is one in which lawyers, AI services, and third parties likely will all be involved at some point in a large majority of cases, the legal profession must take a comprehensive approach to ensuring that AI is integrated responsibly and ethically into all forms of legal services. For the reasons outlined in this Article, part of this approach must entail restraint from imposing or advocating for arguably self-serving restrictions on emerging legal self-help solutions. Such restraints could stunt the development of the larger AI revolution in law in a way that would ultimately favor large firms over other legal services and the broader public interest.

With an eye toward the broad challenges facing the profession, legal communities should urgently initiate robust dialogue and issue guidance concerning the ethical challenges stemming from the emergence of AI systems

204. See Kevin P. Lee, *The Citizen Lawyer in the Coming Era: Technology Is Changing the Practice of Law, but Legal Education Must Remain Committed to Humanistic Learning*, 40 OHIO NORTHERN U. L. REV. 1, 30–36 (2013) (defending humanistic education as necessary for the formation of citizen lawyers who are the artisans of democratic citizenship).

205. See, e.g., McGinnis & Pearce, *supra* note 12, at 3042 (“[C]ounselors who must persuade unwilling clients to do what is in their self-interest will . . . continue to have a role [in legal services], since machines will be unable to create the necessary emotional bonds with clients.”).

in law. The appended list provides a starting point for this dialogue and eventual guidance. It summarizes the new challenges concerning existing obligations, as outlined in this Article, and identifies new tensions between certain obligations, which must be confronted proactively.

It is time for lawyers to confront these challenges, wrestle with the ethical tensions, and issue guidance that will help the profession responsibly, morally, and ethically integrate AI into practice. With a more fully informed appreciation of the unique nature of AI and the associated ethical challenges, the profession can more thoughtfully confront questions concerning the unauthorized practice of law as AI's effect on the profession becomes more cognizable. There may be certain areas of the law (for example capital criminal cases or sensitive deportation cases) where the profession ultimately decides that AI is not an appropriate tool. In addition, there may be certain tasks (such as actual brief writing) that are not suitable for automation or AI, or which make lawyers less effective in their representation. This Article offers a framework for evaluating these questions in light of lawyers' ethical obligations. Lawyers, clients, third parties, and decision makers must all rise to these challenges if the AI revolution is to continue in a way that will improve the effectiveness of lawyers across all parts of the industry and ultimately increase access to justice.

APPENDIX

Competence—What it means to practice competently fundamentally changes when a lawyer uses AI that performs increasingly sophisticated tasks, especially if the lawyer lacks a full appreciation for how the underlying technology works. Competence in the era of AI should require a lawyer to either be involved in the design of the AI systems they are using, or at the very least, to understand—with the help of an expert, if needed—certain underlying characteristics that affect (1) the AI's bias, including conscious bias manifested in the design, unconscious bias of the designer, and bias of the underlying data; (2) AI's limits, including the limits of observational data and limits resulting from the exclusion of information which has not been datafied; and (3) AI's confidentiality concerns.

Confidentiality—The emergence of AI in law practice must fundamentally change the way lawyers think about, talk about, and protect client confidentiality in light of the new ways that client information will be generated, used, stored, and in some cases, comingled with that of other clients. The emergence of AI in law practice magnifies security challenges associated with other less sophisticated technologies. Further, because of the changes to the way client information is gathered, datafied, formatted, and used, keeping unwanted eyes off of passively stored documents will no longer be sufficient to ensure that clients' confidences are protected in the ways that they would expect. This recognizes inevitable tension between existing ethical obligations. AI relies on access to critical, sometimes sensitive information, and withholding certain data from the system's analysis could undermine the effectiveness of a service assisting with tasks that involve case development, legal research, or argument development. Protecting confidentiality in the era of AI must go beyond merely ensuring security and must include (1) competently understanding how AI systems work; (2) communicating with clients and former clients to understand their expectations and preferences; and (3) ensuring that the designers and managers of AI systems, including third parties, understand the critical importance of confidentiality in this new ecosystem.

Supervising Third Parties—AI services will frequently be, or at the very least involve, third parties. AI's role in a lawyer's practice, including its ability to guide a lawyer's judgment based on past outcomes, requires additional diligence beyond that which has been advised in prior guidance concerning other technologies. A baseline technical understanding of AI is not sufficient to ensure that lawyers consider the myriad social, ethical, and moral issues that AI raises in the practice of law. Although security is important, it is not the only thing a lawyer needs to consider in evaluating whether an AI service will effectively maintain client confidentiality, among other obligations. The increased role of third parties also heightens the importance of (1) the duty to communicate with clients, (2) the duties to exercise independent judgment and render candid advice, and (3) lawyers' ongoing obligations to former clients.

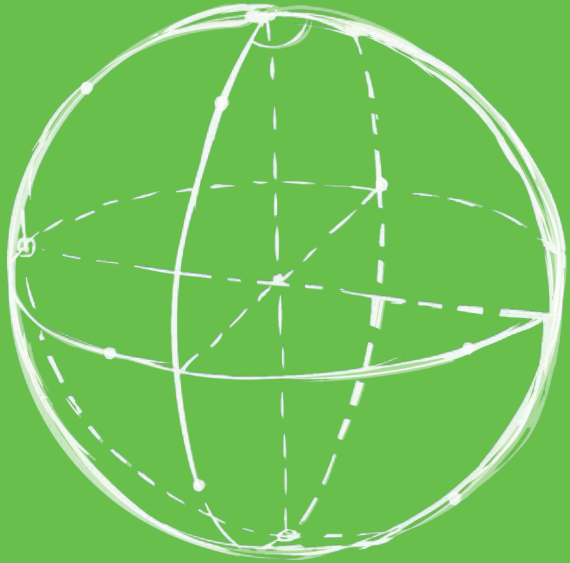
Communicating with Clients—AI will play a transformative role in how a lawyer handles a client's case, which makes the communication between the lawyer and the client all the more essential to ensuring an effective, ethical representation. If AI is to be used in a way that accounts for all of a client's needs, it will require gathering, datafying, formatting, and using especially sensitive client information in new ways. Therefore, communication with clients is of paramount importance. Lawyers must discuss with clients the potential risks to their information, as well as the fundamental nature of AI as a means of assisting with the representation—one that either has severe limitations (because many of a client's needs are not datafied, and therefore not accounted for by the machine intelligence), or which makes very complex use—with third parties—of especially sensitive new data, not previously datafied. However, mere tension between these ethical obligations should not alone preclude a lawyer or firm from responsibly implementing a potentially transformative and beneficial AI service, just so long as such decisions weigh client needs and preferences in light of these ethical tensions, and are made in consultations between the lawyer and client in which both parties are sufficiently informed about the nature of the AI they are dealing with.

Independent Judgment & Candid Advice—One of the major limits of AI is its inability to take into account information beyond the observable data that it has at its disposal. Many pieces of information, including sensitive or embarrassing information concerning the client, the instinctual knowledge of the lawyer, and relevant non-legal factors that the AI might not have access to, might not be datafied. Consistent with the preferences of a client, this information that drives a lawyer's professional judgment must not be marginalized if AI is adopted.

Obligations to Former Clients—AI's power is derived in part from its ability to leverage information from many different data points, from which new inputs can identify analogous points to create helpful outcomes. In law practice, these data points touch many cases, authorities, and clients, and will yield everything from a suggested case to read, to a suggested format of an argument or overall legal strategy based on past favorable outcomes. Unlike days past when lawyers might have shredded or deleted client information at some point, there is a disincentive to dispose of any client information today, meaning client information could remain not only in existence, but remain in use, indefinitely. In order to preserve the trust between clients and lawyers in the age of AI, lawyers must make sure that this information is secured and that their clients trust AI as a tool in their cases, reinforcing the need for competence, communication, and other obligations that have been outlined in this guidance.

AI and Robots in the Workplace

F



AI and Robots in the Workplace

January 9, 2020



AMERICAN **BAR** ASSOCIATION

Presented by



NATALIE A. PIERCE

Shareholder

Littler Mendelson, San
Francisco, CA



R. JASON STRAIGHT

Senior Managing Director

Ankura, New York, NY



AUSTIN TARANGO

Product Counsel

Google AI, San Francisco,
CA

Agenda

- AI-Powered HR Systems: Their Arrival, Expanding Applications and Necessity
- Legal Challenges: The Call for Transparency and Coming Regulatory Climate, Privacy Requirements, and the Emergence of Algorithmic Bias & Biometric Class Actions
- Navigating AI and Machine Learning Adoption Scenarios and Solutions
- Benchmarking Legal/Ethical Compliance for AI-Powered HR Systems and Practical Recommendations



November 6, 2019 Washington Post:

The Electronics Privacy Information Center: Filed Formal
Complaint With Federal Trade Commission alleging:
“Unproven AI Recruiting Technology”
“Unfair and Deceptive Trade Practice”

AI's Unstoppable March Transforming the Workplace

- August 9, 2019 Illinois Governor J. B. Pritzker signed the Illinois Artificial Intelligence Video Interviewing Act
 - Signals The Arrival of AI Reaching The Core of the Talent Assessment Process
 - Foreshadows the Coming Legislation, Regulations, and Litigation
- Accompanying AI's Unstoppable March Transforming the Workplace



What are these HR AI systems...?

- Video Interviewing and Assessments
- Blind Resumes
- Ensure all applicants are getting a “thank you for applying” email
- Utilizing chat bots to give candidates access to information they need at anytime
- Quickly getting open job information out on Social Media
- Having algorithms screen resumes for you



HR is Buzzing About AI

7 Ways Artificial Intelligence is Reinventing Human Resources

3 Ways That A.I. Is Transforming HR and Recruiting

5 Ways (AI) in

The Rise of AI in Recruitment | AI for Recruitment is Here

Is the March of AI into HR inevitable?

- HR is preparing to embrace AI tools without knowing exactly what they will deliver or change.
- For many companies, the first pilots of AI are in talent acquisition.
 - Can see significant, measurable, and immediate results
 - Reduces time to hire
 - Increases productivity for recruiters
 - Delivers an enhanced candidate experience that is seamless and simple
- Many feel if they don't adopt AI systems, they will fall behind.
 - Will they? Are they?



AI and HR's New Digital Mandate

- No facet of HR remains untouched and the potential is just emerging: Bain & Company
 - online recruiting,
 - AI-assisted interviewing,
 - and machine learning predicting who will resign,
 - chatbots handling out-sick notices,
 - smart redesign algorithms evaluating performance and identifying employees for layoff.
- The Promise: Cost savings, speed, quality and accuracy – all essential to remain competitive.

Brief

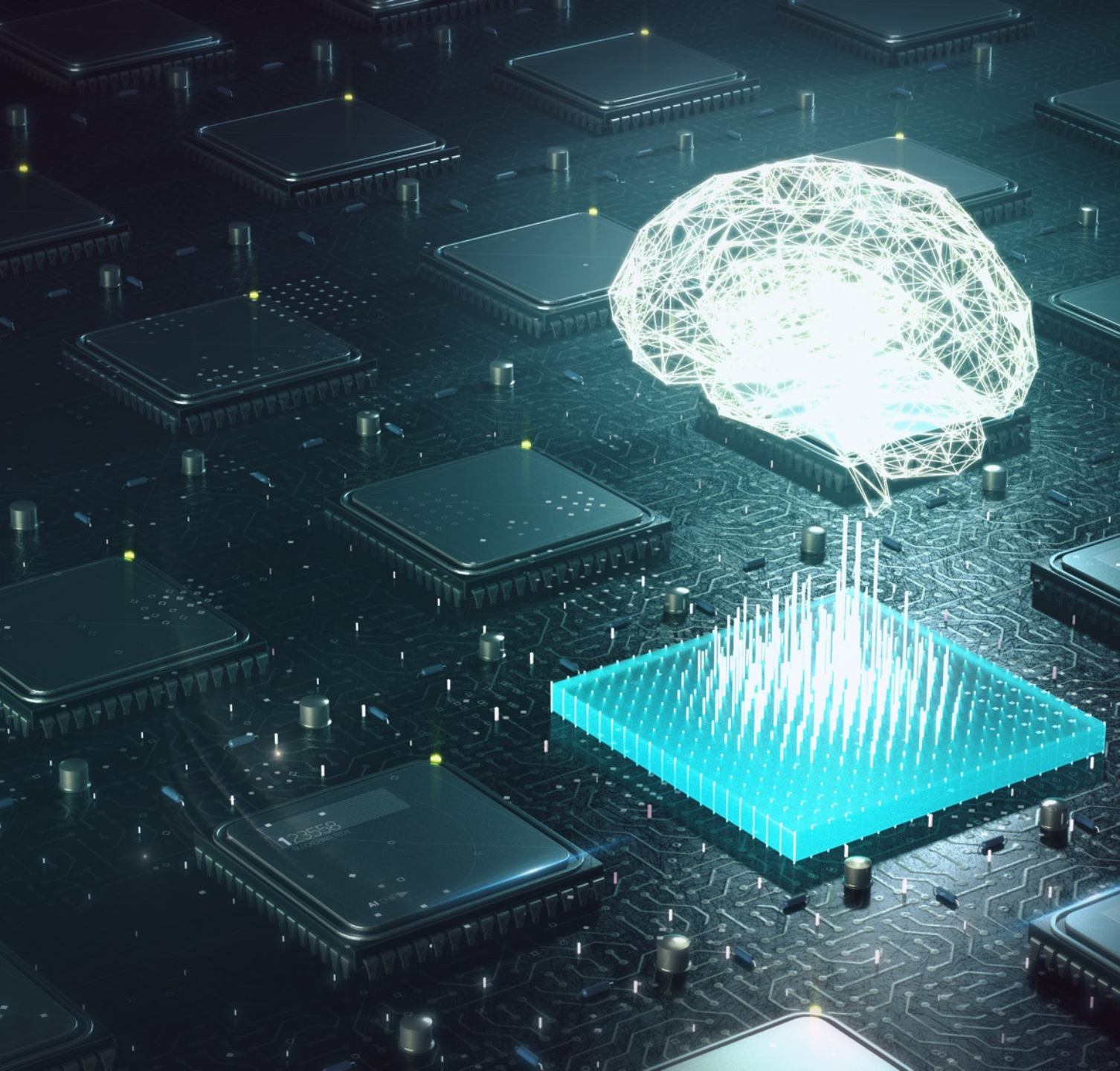
HR's New Digital Mandate

Digital technologies have become essential for HR to engage top talent and add value to the business.

By Michael Heric

October 10, 2018 • 12 min read





The Dark Side: The Call for Transparency & Coming Regulatory Climate

The Call for Transparency & Coming Regulatory Climate

Worker Acceptance and the Call For Transparency



The Call for Transparency & Coming Regulatory Climate

Lessons from the Illinois AI Video Interviewing Act



The Call for Transparency & Coming Regulatory Climate

GDPR Mandates and Creating Expectations



The Call for Transparency & Coming Regulatory Climate



California Consumer Privacy Act and AB 25

The Call for Transparency & Coming Regulatory Climate

Tiered Transparency and the Downside of Too Much Technical Details

The Call for Transparency & Coming Regulatory Climate

The Black Box vs. Glass Box Debate



The Dark Side: The Emergence of Algorithmic Bias & Biometric Class Actions

The Dark Side of AI-Powered HR Systems

The New York Times

Opinion

Beware of Automated Hiring

It won't end employment discrimination. In fact, it could make it worse.

By Ifeoma Ajunwa

Dr. Ajunwa is an expert on employment and labor law.

Oct. 8, 2019



The Emergence of Algorithmic Bias & Biometric Class Actions

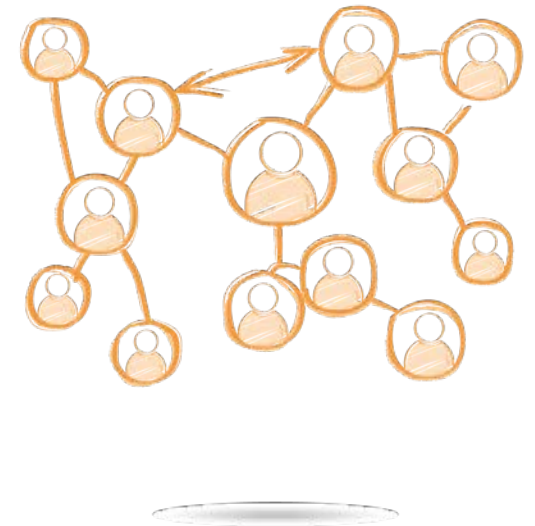
- AI-powered talent acquisition includes both recruiting and assessment.
- Such systems are often described and viewed as more objective because of the technology. This can be true or misleading, and sometimes even false.
- AI uses algorithms which are created by humans and can reflect human biases— both intentionally and unintentionally created.
- Using AI-powered HR systems requires awareness of the potential of Algorithmic Bias, which can include biases that are unlawful such as discriminating against talent on basis of protected categories such as age, sex, race, color, disabilities, religion, national origin, and several others, depending on applicable laws.
- Intentional discrimination is possible, such as alleged age restriction on social media advertising.

The Emergence of Algorithmic Bias & Biometric Class Actions

- Algorithms increasingly deploy pattern recognition known as machine learning. Such systems learn from test data that often is collected regarding the most successful existing employees. If most of these employees are men, for example, patterns can be identified that favor men over women.
- Examples of disparate impact include zip codes, language, names, and even seemingly objective criteria, such as years of experience.
- One of the greatest challenges facing employers is algorithmic bias class actions.
 - EEOC increases scrutiny and review.
 - Plaintiff's counsel are actively looking for ways to bring such suits.
 - Several demand letters have issued.
 - Class Actions are now being filed (Disability and Age cases).

The Emergence of Algorithmic Bias & Biometric Class Actions

- Algorithmic Bias and Coming Class Actions
 - AI's Debated Promise To Make Talent Recruitment and Assessment More Objective, Reducing Both Intentional and Unconscious Human Bias.
 - Human Bias Can Influence Algorithm Design and Operation, e.g, employment gaps, cultural fit, and years of experience.
 - AI's Machine Learning, Deep Learning, & Neural Networks and How They Learn: Potential For Disparate Impact Based Discrimination.
 - Anticipated Surge in Algorithmic Bias Class Actions: A Bipolar Disorder Diagnosed College Student with Near-Perfect SAT Score Rejected By Several Employers Using An AI-Powered Personality Test; Manufacturing Company Using AI Selection Tool For Layoffs Alleged To Violate Age and Pensions Acts.
 - From Lie Detector Statutes to the Federal Credit Reporting Act: Unanticipated Legal Compliance Challenges



The Emergence of Algorithmic Bias & Biometric Class Actions

Biometric Identification Requirements, Challenges, and Class Actions

- Informed Consent State Statutes
(Over 200 Class Actions Under Illinois Biometric Information Privacy Act)
- UK High Security Firm Failed To Use Encryption And Stored Actual Retina and Fingerprint Scans.

Major breach found in biometrics system used by banks, UK police and defence firms

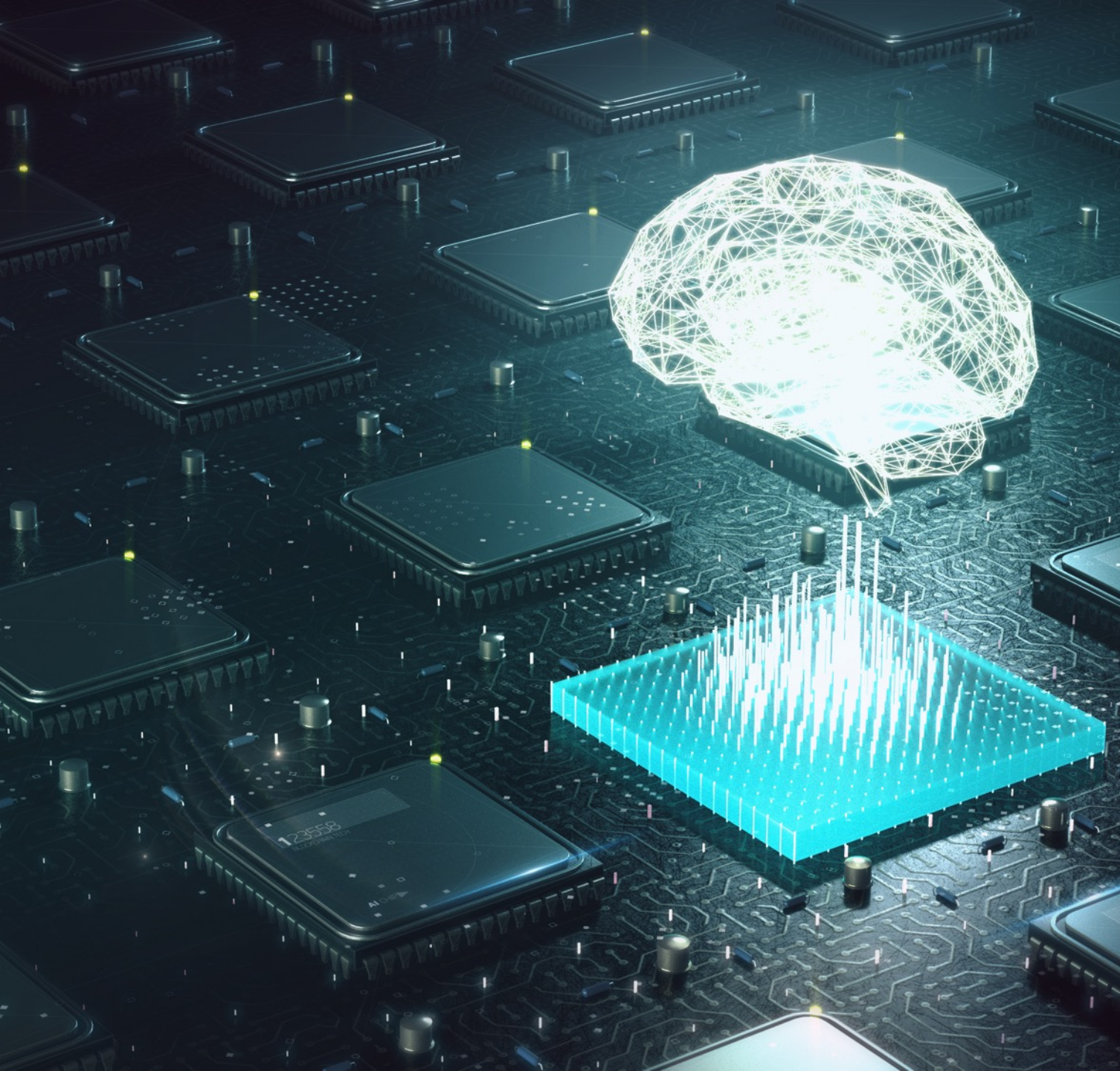
[theguardian.com/technology/2019/aug/14/major-breach-found-in-biometrics-system-used-by-banks-uk-](https://theguardian.com/technology/2019/aug/14/major-breach-found-in-biometrics-system-used-by-banks-uk)

Josh Taylor

August 14,
2019

The fingerprints of over 1 million people, as well as facial recognition information, unencrypted usernames and passwords, and personal information of employees, was discovered on a publicly accessible database for a company used by the likes of the UK Metropolitan police, defence contractors and banks.

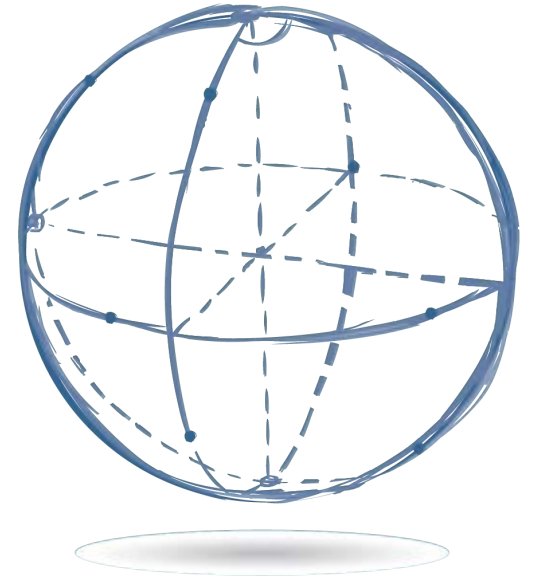




Benchmarking Legal Compliance for AI- Powered HR Systems & Practical Recommendations

Benchmarking Legal Compliance for AI-Powered HR Systems

- Benchmarking Legal Compliance for AI-Powered HR Systems
 - Vital Role of In-House Legal Compliance Attorney as Part of the Overall Multi-Disciplinary and Diverse Compliance Team. Catalogue and Understand Existing Laws and Regulations Regarding Algorithmic intelligence, Data, and Privacy Rights, Including Regulatory Environment.
 - Incorporate Legal Compliance Review and Documentation Into All Phases of Development of AI-Powered HR Systems: Design, Building, Implementation, and Monitoring.
 - Determine Whether HR System Includes Automated AI Decision Making; If So, Include Human Oversight and Ability To Change Decisions.



Benchmarking Legal Compliance for AI-Powered HR Systems

- Adopt Audit Standards, and Documentation Methods and Process.
- Assess the Transparency and Explainability Levels of the AI System For Users, Developers, Legal Compliance Team, and Anticipated Response From Regulators and Litigation: Lessons from Prisoner Sentencing Algorithms to Predictive Coding Will Help Determine Expected Disclosure Requirements Mandated By Regulators and the Litigation Process.
- Make Training Data Selections and Testing One of the Most Critical Legal Compliance Benchmarks: This Data Selection Will Be Important In Avoiding Algorithmic Bias and Will Likely Be Sought By Regulators and In Litigation.
- Audit the Origin, Quality, and Fitness of Your Data.
- Repeated Use of Data Analytics to Audit Throughout Development, and Especially For Monitoring: Single Audits vs. Continuous Audits.

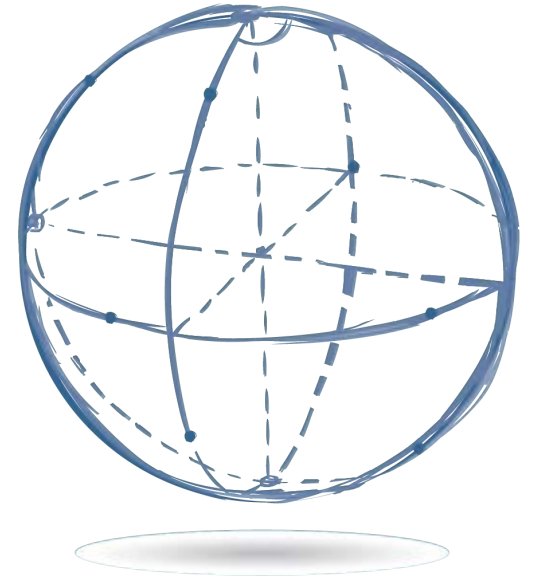
Benchmarking Legal Compliance for AI-Powered HR Systems

- Customize Legal Compliance Check-List Based On Type of AI-Powered HR Systems and Jurisdictions.
- Evaluate and Potentially Use AI-Powered Legal Compliance Assessment Tools, Including Proven Industry Toolkits and Qualified Independent Researchers.
- Conduct a Similar Legal Compliance Review Of Third-Party Systems, Including Indemnity Agreements.
- Determine Likely Application of Attorney-Client Privilege, Trade Secret Protections, and Other Privileges, Including Limitation of Attorney-Client Privilege Over Underlying Data.



Compliance Review

- Consider Conducting Your Compliance Review Such That It Qualifies Under The Self-Evaluation Requirements for a Safe Harbor Under the Massachusetts Act to Establish Pay Equity (MEPA).
- Your Jurisdiction Is Likely Outside of Massachusetts But The Self-Evaluation Requirements Are Likely To Be Used In Future Safe Harbor Legislation and Are Base-Line Standards for a Legal Compliance Review of Any AI-Powered HR System.
- The MEPA's Self-Evaluation Must Be "Reasonable in Detail and Scope," Conducted in Good Faith, Include an Adequate Number of Jobs and Employees (Positions and Applicants) Based On All Relevant and Available Information, and Be Reasonably Sophisticated in Analysis of the Tasks Performed.
- While the Purpose of the Self-Evaluation of AI-Powered HR Systems Is To Provide Reasonable Levels of Legal Compliance, the MEPA Only Requires the Elimination of Disparities in a Reasonable Amount of Time.



Data Privacy, Data Security, and Information Governance

G

NATIONAL INSTITUTES

AI Data Privacy, Data Security, and Information Governance

January 10, 2020 / 8:30 am, Santa Clara, CA



THE PREMIER SOURCE FOR CLE

Speakers

- **Ian C. Ballon**, Litigation Shareholder, Greenberg Traurig LLP, Silicon Valley, CA; **Ballon@gtlaw.com**
- **Kristin J. Madigan**, Counsel, Crowell & Moring LLP; former Attorney, Division of Privacy & Identity Protection, Bureau of Consumer Protection, FTC; Washington, DC; **kmadigan@crowell.com**
- **Lucy L. Thomson**, Founding Principal, Livingston PLLC; Past Chair, ABA Science & Technology Law Section; Washington, DC; **lucythomson1@mindspring.com**
- **Ruth Hill Bro**, Privacy/Cybersecurity Attorney; Co-Chair, ABA Cybersecurity Legal Task Force; Past Chair, ABA Science & Technology Law Section, Chicago IL (Moderator); **ruth.hill.bro@gmail.com**

Topics: The Ultimate Trilogy

3 Views from the Top of the Galaxy

- **Data Security: Data, data everywhere; the threat landscape (Lucy Thomson)**
- **Data Privacy: California, here I come; GDPR (Kristin Madigan)**
- **A Litigation Perspective (Ian Ballon)**

Panel Discussion: 3 Questions and Q&A

Top 3 Takeaways

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



A short time ago, in a galaxy not far away....

AI PRIVACY/SECURITY

A NEW WORLD

It is the information age. An entire generation has been raised on the Internet. Customers use technology 24/7. The line between work and home has blurred. AI is being used everywhere and collecting data about everything and everyone, with implications

Star Wars and AI Privacy/Security

- “I have a bad feeling about this.”
- “He’s more machine now than man, twisted and evil.”
- “Never underestimate a droid.”

1. View From the Top of the Galaxy

Data Security/Threat Landscape

Lucy Thomson

Expanded Cyber Risks in the Complex AI Technology World

Attack Surfaces

- Internet of Things (IoT)
- Big Data
- AI

Attack Objectives

Compromise (CIA)

Confidentiality

Integrity

Availability

+ Safety

Threat Actors

- Russia
- China
- North Korea
- Iran

Cyber Weapons

- Defensive AI
- Offensive AI
- Customized Attacks

2. View From the Top of the Galaxy

**Data Privacy:
California, GDPR**

Kristin Madigan

AI/Robotics and Compliance

Given the capabilities and threat landscape, consider AI/Robotics compliance in light of:

- CCPA
- GDPR
- Other possible legal frameworks or requirements
 - Federal privacy law
 - State laws modeled on CCPA

California Consumer Privacy Act: Overview

- Comprehensive U.S. data privacy law
 - First of its kind, signed into law on June 28, 2018
 - Effective beginning January 1, 2020
 - Enforceable on the earlier of July 1, 2020, or 6 months after the California Attorney General's office promulgates final CCPA regulations
- Designed to give consumers (defined broadly as natural persons who are California residents) control over the collection and use, including sale, of their personal data.

California Consumer Privacy Act: Scope

- Similar to the EU General Data Protection Regulation (GDPR) in some respects, including broad enforcement scope and giving individual consumers more control over their personal information even when held by third parties.
- CCPA provides extensive rights to consumers:
 - Right to notice of data collection
 - Right to access data collected and request deletion
 - Companies must respond to verified requests and provide information for prior 12 months
 - Right to opt out of the sale of personal information

California Consumer Privacy Act: Scope

- Applies to for-profit entities doing business in California that meet one of the criteria below:
 1. Annual gross revenue in excess of \$25M USD
 2. Annually buys, receives for commercial purposes, sells or shares for commercial purposes personal information of 50,000 or more consumers, households or devices
 3. Derives 50 percent or more of annual revenue from selling consumers' personal information

California Consumer Privacy Act: Enforcement

- Enforceable by the Attorney General
 - 30-day period to cure violations after notice
 - Civil penalties up to \$2,500 per negligent violation or \$7,500 per intentional violation
- Private civil actions (for data breaches only)
 - Statutory damages of \$100-\$750 per consumer per incident or actual damages if company failed to implement reasonable data security

AI/Robotics and Compliance

- Consider how AI applications function – both collecting and processing consumer information
 - Data collected and used for the benefit of consumers vs. selling/sharing or purchasing consumer data
 - Inferences based on data collected about consumers
 - Allowing for deletion of personal information from AI applications that learn from the personal information
- For compliance, focus will be on data mapping and compliance policies including data use and masking

3. View From the Top of the Galaxy

Litigation Perspective

Ian Ballon

PLANNING FOR – AND AVOIDING – CCPA LITIGATION

What litigation will we see?

- ▣ CCPA class action litigation over cybersecurity breaches brought in state court in California and federal court potentially anywhere but most likely in California
- ▣ Class action litigation over those provisions of the CCPA not actionable under California law, under the laws of other states (for companies that implement the CCPA nationally)
 - A violation of law may be an unfair trade practice under Massachusetts law and in some other jurisdictions
 - Failure to implement CCPA procedures nation-wide could be characterized as negligent – falling below perceived practices
 - Failing to comply with CCPA obligations incorporated by reference in a privacy statement could support a breach of contract claim
- ▣ Suits between or among businesses, service providers, and/or third parties for breach of contract and indemnification (including claims arising out of AG enforcement actions)
- ▣ Suits against insurers over coverage issues for litigation and AG enforcement actions

California Consumer Privacy Act

- California Consumer Privacy Act (effective Jan. 1, 2020)
 - preempted in the future by federal legislation??
- AG Regulations on or by January 1, 2020 and enforcement by the AG by July 1, 2020 or sooner (Draft regulations issued in October 2019)
- Private cause of action – good news/ bad news
- Applies to California residents, not just *consumers*
- Applies to *businesses* with (1) annual gross revenue > \$25 M; (2) that buy, sell or receive for commercial purposes personal information of 50,000 or more consumers, households or devices, and (3) businesses that derive 50% or more of their annual revenue from selling consumers' personal information (excludes entities subject to federal regulation)
- Regulates *businesses, third parties* and *service providers*
- Consumer rights to
 - Notice of the personal information collected and the purpose of collection at or before collection
 - Request disclosure up to 2x every 12 months (generally free of charge, generally 45 days)
 - Opt out of collection (for minors 16 years and under, opt-in consent is required)
 - Deletion of personal information
- *Personal information* is very broadly defined.
 - Even deidentified information could become *personal information* if a business fails to undertake certain protective measures
 - Inferences drawn about a consumer (ie, likes to dive) are *personal information*
- Broad: Rather than regulating the use, collection and dissemination of information obtained *by companies from consumers*, as past consumer laws did, the CCPA focuses on information *about* state residents
- Nondiscrimination/ financial incentives
- Required Privacy Policy disclosures – but a Privacy Policy alone is not enough

CCPA Putative Class Action Litigation

- California Consumer Privacy Act (effective Jan. 1, 2020)
 - if not first preempted by federal legislation
- The private right of action narrowly applies only to security breaches and the failure to implement reasonable measures, not other aspects of the statute
- However, plaintiffs may recover statutory damages of between \$100 and \$750
- The CCPA creates a private right of action for consumers “whose **nonencrypted or nonredacted** personal information . . . is subject to an unauthorized access and exfiltration, theft, or disclosure as a result of the business’s violation of the duty to **implement and maintain reasonable security procedures and practices**”
- What is *reasonable* will be defined by case law and potentially guidance from the California Attorney General
 - Regulations to be issued on or before January 1, 2020, effective by July 1, 2020 (or earlier)
- \$100 - \$750 “per consumer per incident or actual damages, whichever is greater, injunctive or declaratory relief, and any other relief that a court deems proper
- In assessing the amount of statutory damages, the court shall consider “any one or more of the relevant circumstances presented by any of the parties to the case, including, but not limited to, the nature and seriousness of the misconduct, the number of violations, the persistence of the misconduct, the length of time over which the misconduct occurred, the willfulness of the defendant’s misconduct, and the defendant’s assets, liabilities, and net worth”
- 30 day notice and right to cure as a precondition to seeking statutory damages
 - Modeled on the Consumer Legal Remedies Act
 - Can one “cure” a breach?
 - If cured, a business must provide “an express written statement” (which could later be actionable)

How will this litigation play out – and how can we avoid it??

- ▣ CCPA class action litigation over cybersecurity breaches –
- ▣ **Three relevant touchstones:**
 - California CLRA litigation (30 day notice & cure provision)
 - Cybersecurity class action litigation over the past decade
 - TCPA class action litigation (class action suits where plaintiffs can recover statutory damages regardless of injury or damage)
 - ▣ 3,803 new suits filed in 2018
 - ▣ 2,300 in 2019 through August 30 (webrecon.com)
- ▣ Class action litigation over those provisions of the CCPA not actionable under California law, under the laws of other states (for those companies that are rolling out the CCPA nationally)
- ▣ **How to avoid class action litigation?**
 - Encrypt your data and comply with the CCPA (or make sure to avoid its application)....
 - Craft a binding and enforceable arbitration provision and include it in every contract with consumers under the FAA (not state law), avoiding or complying with AAA requirements
 - Make sure your online and mobile consumer contract formation process conforms to the law in the worst jurisdictions (currently the First and Ninth Circuits)
 - Where you don't have privity of contract, make sure you are an intended beneficiary of an arbitration clause in a contract with a business partner who does have privity (because you will be sued!)
 - Explore insurance coverage
- ▣ Suits between or among businesses, service providers, and/or third parties for breach of contract and indemnification (including claims arising out of AG enforcement actions)
 - Play close attention to indemnification provisions, encryption obligations, notice obligations and intended beneficiary clauses where there is no privity of contract with consumers
- ▣ Suits against insurers over coverage issues for litigation and AG enforcement actions
 - Check your insurance coverage NOW
 - Make sure you can hire counsel of your choosing

CCPA Putative Class Action Litigation

- ▣ \$100-\$750 “per consumer per incident or actual damages, whichever is greater
- ▣ Suits will be brought as putative class action suits
 - 100,000 consumers → up to \$75,000,000
 - 1,000,000 state residents → up to \$750,000,000 and *at least* \$100,000,000
- ▣ 30 day advance notice and the right to cure
- ▣ Compare to Cal. Civil Code § 1798.84(b)
- ▣ Standing
 - *In re Zappos.com, Inc.*, 888 F.3d 1020, 1023-30 (9th Cir. 2018) (holding that plaintiffs, whose information had been stolen by a hacker but who had not been victims of identity theft or financial fraud, nevertheless had Article III standing to maintain suit in federal court)
 - *Cahen v. Toyota Motor Corp.*, 717 F. App'x 720 (9th Cir. 2017) (affirming the lower court's ruling finding no standing to assert claims that car manufacturers equipped their vehicles with software that was susceptible to being hacked by third parties)
 - *Antman v. Uber Technologies, Inc.*, Case No. 3:15-cv-01175-LB, 2018 WL 2151231 (N.D. Cal. May 10, 2018) (dismissing, with prejudice, plaintiff's claims, arising out of a security breach, for allegedly (1) failing to implement and maintain reasonable security procedures to protect Uber drivers' personal information and promptly notify affected drivers, in violation of Cal. Civ. Code §§ 1798.81, 1798.81.5, and 1798.82; (2) unfair, fraudulent, and unlawful business practices, in violation of California's Unfair Competition Law, Cal. Bus. & Prof. Code § 17200; (3) negligence; and (4) breach of implied contract, for lack of Article III standing, where plaintiff could not allege injury sufficient to establish Article III standing); *see generally infra* § 27.07 (analyzing claims raised in security breach litigation).
- ▣ Contractual suits between a business, service provider and/or third party

WHAT CAN WE LEARN
FROM CLRA, TCPA, AND
SECURITY BREACH
PUTATIVE CLASS ACTION
LITIGATION TO DATE?

Anticipating CCPA Class Action Litigation

- ▣ CLRA
 - You will have 30 days to plan to be sued if a plaintiff wants to recover damages
 - Some may sue you anyway claiming notice would be futile and the lawsuit constitutes notice, so plan ahead and retain counsel now
- ▣ TCPA
 - There will be an avalanche of lawsuits – likely multiple suits for every cybersecurity breach, as class action lawyers jostle for lead position
 - Some companies will overpay to settle these cases (pushed by insurers or out of concern for potentially large damage awards), fueling even more litigation
 - Relief eventually may come from Congress, but not before one or more companies are hit with punitive awards
 - ▣ Golan v. FreeEats.com, Inc., 930 F.3d 950, 962-63 (8th Cir. 2019) (statutory min. damages \$1.6 Billion)
- ▣ Cybersecurity class action suits
 - The life cycle of a case – and how to win!
 - Standing (caveat – for CCPA cases you may end up in California state court)/ MTD/ SJ/ Class certification/ Settlement/ No trials
 - Settlement values – and how to value your case and your exposure
 - Statutory damages under the CCPA will skew settlement numbers nationally
- ▣ Standing: To establish injury in fact, a plaintiff must have suffered “an invasion of a legally protected interest” that is “concrete and particularized” and “actual or imminent, not conjectural or hypothetical”
 - Frank v. Gaos, 139 S. Ct. 1041, 1046 (2019) (remanding a 9th Circuit order to address “whether any named plaintiff” had alleged injuries “sufficiently concrete and particularized to support standing” under *Spokeo*)
 - Clapper v. Amnesty International USA, 568 U.S. 398 (2013) (5-4)
 - Spokeo, Inc. v. Robins, 136 S. Ct. 1540 (2016) (Alito) (compromise 6-2)
- ▣ Circuit split on the risk of future harm under *Clapper*

Security Breach Litigation

- Circuit split – Low threshold: 6th, 7th, 9th, DC vs. high threshold: 2d, 4th, 8th (3d)
- Remijas v. Neiman Marcus Group, 794 F.3d 688 (7th Cir. 2015)
- Lewert v. P.F. Chang's China Bistro Inc., 819 F.3d 963 (7th Cir. 2016)
- Dieffenback v. Barnes & Noble, Inc., 827 F.3d 826 (7th Cir. 2018)
- Galaria v. Nationwide Mut. Ins. Co., 663 F. App'x 384 (6th Cir. 2016) (2-1)
- Reilly v. Ceridian Corp., 664 F.3d 38 (3d Cir. 2011), *cert. denied*, 566 U.S. 989 (2012)
- Beck v. McDonald, 848 F.3d 262 (4th Cir. 2017)
 - Allegation that data breaches created an enhanced risk of future identity theft was too speculative to constitute an injury-in-fact
 - Rejected evidence that 33% of health related data breaches result in identity theft
 - Rejected the argument that offering credit monitoring services evidenced a substantial risk of harm (rejecting *Remijas*)
 - Mitigation costs in response to a speculative harm do not qualify as injury in fact
- Whalen v. Michael's Stores, Inc., 689 F. App'x. 89 (2d Cir. 2017)
 - The theft of plaintiff's financial information was not sufficiently concrete or particularized to satisfy *Spokeo*
 - breach of implied contract, N.Y. Gen. Bus. L. § 349
 - Plaintiff made purchases via a credit card at a Michaels store on December 31, 2013
 - Michaels experienced a breach involving credit card numbers but no other information such as a person's name, address or PIN
 - plaintiff alleged that her credit card was presented for unauthorized charges in Ecuador on January 14 and 15, 2014, but she did not allege that any fraudulent charges were actually incurred by her prior to the time she canceled her card on January 15
- Attias v. Carefirst, Inc., 865 F.3d 620 (D.C. Cir. 2017), *cert. denied*, 138 S. Ct. 981 (2018)
 - following *Remijas v. Neiman Marcus Group, LLC* in holding that plaintiffs, whose information had been exposed but who were not victims of identity theft, had plausibly alleged a heightened risk of future injury to establish standing because it was plausible to infer that a party accessing plaintiffs' personal information did so with "both the intent and ability to use the data for ill."
- In re U.S. Office of Personnel Management Data Security Breach Litig., 928 F.3d 42 (D.C. Cir. 2019) (21mil records)
- In re SuperValu, Inc., Customer Data Security Breach Litig., 870 F.3d 763 (8th Cir. 2017)
 - affirming dismissal for lack of standing of the claims of 15 of the 16 plaintiffs but holding that the one plaintiff who alleged he suffered a fraudulent charge on his credit card had standing to sue for negligence, breach of implied contract, state consumer protection and security breach notification laws and unjust enrichment
 - defendants experienced two separate security breaches, which they announced in press releases may have resulted in the theft of credit card information, including their customers' names, credit or debit card account numbers, expiration dates, card verification value (CVV) codes, and personal identification numbers (PINs). Plaintiffs alleged that hackers gained access to defendants' network because defendants failed to take adequate measures to protect customers' credit card information
 - Rejected cost of mitigation (*Clapper*) (Cf. *P.F. Chang's*)
- In re Zappos.com, Inc., 888 F.3d 1020 (9th Cir. 2018), *cert. denied*, 139 S. Ct. 1373 (2019)
 - merely having personal information exposed in a security breach constitutes sufficient harm to justify Article III standing in federal court, regardless of whether the information in fact is used for identity theft or other improper purposes
 - **Bootstrapping** - Because other plaintiffs alleged that their accounts or identities had been commandeered by hackers, the court concluded that the appellants in *Zappos* – who did not allege any such harm – could be subject to fraud or identity theft
- **Causation/ damages**
- **Settlement value**

RELATED
CYBERSECURITY
AND DATA PRIVACY
CLAIMS THAT
COULD BE JOINED IN
A CCPA SUIT

CCPA Class Action Litigation – Related Claims

▣ Related cybersecurity claims

(not preempted by the CCPA if not based on a violation of the CCPA)

- Breach of contract (if there is a contract)
- Breach of the covenant of good faith and fair dealing (if the contract claim isn't on point)
- Breach of implied contract (if there is no express contract)
- Breach of fiduciary duty
- Negligence
- Fraud
- State cybersecurity statutes (especially those that provide for statutory damages and attorneys' fees)

▣ Related data privacy claims

- Electronic Communications Privacy Act
 - ▣ Wiretap Act
 - ▣ Stored Communications Act
- Computer Fraud and Abuse Act
 - ▣ \$5,000 minimum injury
- Video Privacy Protection Act
- State laws
 - ▣ Illinois Biometric Information Privacy Act (recently adopted in other states)
 - ▣ Michigan's Preservation of Personal Privacy Act
 - ▣ California laws including the California Consumer Privacy Act (CCPA) which takes effect Jan 1, 2020
- Breach of contract/ privacy policies

▣ Regulatory enforcement – important to coordinate litigation strategy with California AG (and potentially FTC) enforcement actions

- Experience from other cases

▣ Internet of Things

- Who is responsible
- How does a company establish privity of contract to reduce exposure and obtain assent to an arbitration agreement?

CCPA Class Action Litigation – Related Claims

- ▣ California's Internet of Things (IoT) Law (effective Jan. 1, 2020)
 - Cal. Civil Code §§ 1798.91.04 to 1798.91.06
 - Effective January 1, 2020, the new law will require a manufacturer of a connected device to equip the device with a reasonable security feature or features that are appropriate to the nature and function of the device, appropriate to the information it may collect, contain, or transmit, and designed to protect the device, and any information it contains, from unauthorized access, destruction, use, modification, or disclosure
- ▣ Michigan's Preservation of Personal Privacy Act
- ▣ Illinois Biometric Information Privacy Act (BIPA)
 - A private cause of action for "any person aggrieved by a violation" of BIPA
 - ▣ Rosenbach v. Six Flags Entertainment Corp., 129 N.E.3d 1197 (Ill. 2019) (holding that a person need not have sustained actual damage beyond violation of his or her rights under the statute to be *aggrieved* by a violation)
 - A plaintiff may recover the greater of (1) actual damages or (2) \$1,000 in liquidated damages for negligent violations or \$5,000 for intentional or reckless violations
 - The statute also authorizes recovery of attorneys' fees
 - ▣ Patel v. Facebook, 932 F.3d 1264 (9th Cir. 2019) (affirming certification of a class of Illinois users of Facebook's website for whom the website created and stored a face template during the relevant time period)
 - ▣ In re Facebook Biometric Information Privacy Litig., Case No. 3:15-cv-0373-JD, 2018 WL 2197546 (N.D. Cal. May 14, 2018) (denying cross motions for summary judgment)
 - ▣ Santana v. Take-Two Interactive Software, Inc., 717 F. App'x 12 (2d Cir. 2017) (affirming the lower court's finding of no standing in a BIPA case based on mere procedural violations)

ONLINE AND MOBILE
CONTRACT FORMATION
– YOUR BEST DEFENSE
AGAINST THE CCPA

Online and Mobile Contract Formation

- ▣ **Trend: Continued hostility to implied contracts but some good news**
 - Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1175-79 (9th Cir. 2014)
 - declining to enforce an arbitration clause where the website provided terms of use via a link accessible on every page of the website but provided no notice to users or prompts to demonstrate express assent to those terms; “where a website makes its terms of use available via a conspicuous hyperlink on every page of the website but otherwise provides no notice to users nor prompts them to take any affirmative action to demonstrate assent, even close proximity of the hyperlink to relevant buttons users must click on – without more – is insufficient to give rise to constructive notice”
 - Tompkins v. 23andMe.com. Inc., 840 F.3d 1016 (9th Cir. 2016)
 - While a unilateral right to amend may itself be unconscionable, the presence of such a clause, without more, doesn’t make the entire contract unconscionable.
 - unilateral modification provision was limited by the duty of good faith and fair dealing and plaintiff had not shown how the arbitration clause was unconscionable based on a unilateral amendment
 - Nicosia v. Amazon.com, Inc., 834 F.3d 220 (2d Cir. 2016)
 - reversing the lower court's order dismissing plaintiff's complaint, holding that whether the plaintiff was on inquiry notice of contract terms, including an arbitration clause, presented a question of fact where the user was not required to specifically manifest assent to the additional terms by clicking "I agree" and where the hyperlink to contract terms was not "conspicuous in light of the whole webpage."
 - Meyer v. Uber Technologies, Inc., 868 F.3d 66 (2d Cir. 2017)
 - (1) Uber’s presentation of its Terms of Service provided reasonably conspicuous notice as a matter of California law and (2) consumers’ manifestation of assent was unambiguous
 - “when considering the perspective of a reasonable smartphone user, we need not presume that the user has never before encountered an app or entered into a contract using a smartphone. Moreover, a reasonably prudent smartphone user knows that text that is highlighted in blue and underlined is hyperlinked to another webpage where additional information will be found.”
 - “[T]here are infinite ways to design a website or smartphone application, and not all interfaces fit neatly into the clickwrap or browwrap categories.”
 - Cullinane v. Uber Technologies, Inc., 893 F.3d 53 (1st Cir. 2018)
 - Displaying a notice of deemed acquiescence and a link to the terms is insufficient to provide reasonable notice to consumers



Review your order

By placing your order, you agree to Amazon.com's [privacy notice](#) and [conditions of use](#).

Shipping address [Change](#)

[Redacted]
[Redacted]
[Redacted]
[Redacted]
[Redacted]



Or try Amazon Locker
20 locations near this address

Payment method [Change](#)

VISA

Gift Card

Billing address [Change](#)

Same as shipping address

Gift cards & promotional codes

Order Summary

Items:

Shipping & handling:

Total before tax:

Estimated tax to be collected:

Total:

Gift Card:

Order total:[How are shipping costs calculated?](#)**FREE** TWO-DAY SHIPPING

FREE Two-Day Shipping on this Order: [Redacted] you can save \$5.48 on this order by selecting "FREE Two-Day Shipping with a free trial of Amazon Prime" below.

» [Sign up for a free trial](#)

Estimated delivery: Sept. 25, 2014 - Sept. 26, 2014



Choose a delivery option:

- ☐ FREE Two-Day Shipping with a free trial of [Redacted] **Prime** —get it Wednesday, Sept. 24
- ☐ One-Day Shipping —get it tomorrow, Sept. 23
- ☐ Two-Day Shipping —get it Wednesday, Sept. 24
- ☒ Standard Shipping —get it Sept. 25 - 26
- ☐ FREE Shipping —get it Sept. 28 - Oct. 2

*Why has sales tax been applied? [See tax and seller information](#)

Do you need help? Explore our [Help pages](#) or [contact us](#)

For an item sold by Amazon.com: When you click the "Place your order" button, we'll send you an email message acknowledging receipt of your order. Your contract to purchase an item will not be complete until we send you an email notifying you that the item has been shipped.

Colorado, Oklahoma, South Dakota and Vermont Purchasers: [Important information regarding sales tax you may owe in your State](#)

Within 30 days of delivery, you may return new, unopened merchandise in its original condition. Exceptions and restrictions apply. See [Amazon.com's Returns Policy](#)

[Go to the Amazon.com homepage](#) without completing your order.

< Register



GOOGLE+



FACEBOOK

OR

First Name

Last Name

name@example.com



(201) 555-5555

Password

NEXT

< Payment

PROMO CODE



Credit Card Number



SCAN

MM

YY

CVV



U.S.

ZIP

REGISTER

OR

PayPal



Google Wallet


By creating an Uber account, you agree to the
[TERMS OF SERVICE & PRIVACY POLICY](#)


CANCEL

LINK PAYMENT

GO BACK

 1234 5678 9012 3456

 scan your card

 enter promo code

OR

PayPal

By creating an Uber account, you agree to the


[Terms of Service & Privacy Policy](#)


CANCEL

LINK PAYMENT

GO BACK

 1234 5678 9012 3456

 scan your card

 enter promo code

By creating an Uber account, you agree to the

[Terms of Service & Privacy Policy](#)

1

2

ABC

3

DEF

4

GHI

5

JKL

6

MNO

7

PQRS

8

TUV

9

WXYZ

0



Online and Mobile Contract Formation

▣ Trend: Continued hostility to implied contracts but some good news

- Nguyen v. Barnes & Noble Inc., 763 F.3d 1171, 1175-79 (9th Cir. 2014)
 - declining to enforce an arbitration clause where the website provided terms of use via a link accessible on every page of the website but provided no notice to users or prompts to demonstrate express assent to those terms; “where a website makes its terms of use available via a conspicuous hyperlink on every page of the website but otherwise provides no notice to users nor prompts them to take any affirmative action to demonstrate assent, even close proximity of the hyperlink to relevant buttons users must click on – without more – is insufficient to give rise to constructive notice”
- Tompkins v. 23andMe.com, Inc., 840 F.3d 1016 (9th Cir. 2016)
 - While a unilateral right to amend may itself be unconscionable, the presence of such a clause, without more, doesn't make the entire contract unconscionable.
 - unilateral modification provision was limited by the duty of good faith and fair dealing and plaintiff had not shown how the arbitration clause was unconscionable based on a unilateral amendment
- Nicosia v. Amazon.com, Inc., 834 F.3d 220 (2d Cir. 2016)
 - reversing the lower court's order dismissing plaintiff's complaint, holding that whether the plaintiff was on inquiry notice of contract terms, including an arbitration clause, presented a question of fact where the user was not required to specifically manifest assent to the additional terms by clicking "I agree" and where the hyperlink to contract terms was not "conspicuous in light of the whole webpage."
- Meyer v. Uber Technologies, Inc., 868 F.3d 66 (2d Cir. 2017)
 - (1) Uber's presentation of its Terms of Service provided reasonably conspicuous notice as a matter of California law and (2) consumers' manifestation of assent was unambiguous
 - “when considering the perspective of a reasonable smartphone user, we need not presume that the user has never before encountered an app or entered into a contract using a smartphone. Moreover, a reasonably prudent smartphone user knows that text that is highlighted in blue and underlined is hyperlinked to another webpage where additional information will be found.
 - “[T]here are infinite ways to design a website or smartphone application, and not all interfaces fit neatly into the clickwrap or browsewrap categories.”
- Cullinane v. Uber Technologies, Inc., 893 F.3d 53 (1st Cir. 2018)
 - Displaying a notice of deemed acquiescence and a link to the terms is insufficient to provide reasonable notice to consumers
- Starke v. Squaretrade, Inc., 913 F.3d 279 (2d Cir. 2019)
 - Denying motion to compel arbitration where the consumer did not have reasonable notice because the post-sale T&C were not provided in a clear and conspicuous way. An Amazon purchase page said plaintiff would receive a “service contract” by email. Plaintiff then received an email advising he would receive a “service agreement.” He then received an email saying his “contract” was enclosed, but it came in the form of a link and none of the communications put him on notice that his “service contract” would come via a link.
 - (1) no notice it would be a link; (2) the link was buried in an email that primarily comprised a chart (3) more similar to *Nicosia* than *Meyer*

Online and Mobile Contract Formation

▣ Arbitration and Class Action Waivers

- AT&T Mobility LLC v. Concepcion, 131 S. Ct. 1740 (2011)
- Henry Schein, Inc. v. Archer & White Sales, Inc., 139 S. Ct. 524 (2019)
- American Express Co. v. Italian Colors Restaurant, 133 S. Ct. 2304 (2013)
- Tompkins v. 23andMe.com. Inc., 840 F.3d 1016 (9th Cir. 2016)
 - Abrogating or limiting earlier Ninth Circuit cases that applied pre-*Concepcion* California unconscionability case law, which had treated arbitration clauses differently from other contracts
 - Venue selection, bilateral attorneys' fee and IP carve out provisions not unconscionable
 - Enforcing delegation clause
- Baltazar v. Forever 21, Inc., 62 Cal. 4th 1237, 200 Cal. Rptr. 3d 7 (2016) (abrogating earlier precedent that held certain provisions to be unconscionable when included in arbitration agreements)
- Larsen v. Citibank FSB, 871 F.3d 1295 (11th Cir. 2017) (compelling arbitration; unilateral amendment provision modified by the duty of good faith and fair dealing under either Ohio or Washington law)
- National Federation of the Blind v. Container Store, 904 F.3d 70 (1st Cir. 2018)
 - Holding T&Cs illusory under TX law, and declining to enforce the included arbitration clause
 - Rejecting the argument that a unilateral amendment clause was not illusory because modified by the duty of good faith and fair dealing or based on the severability clause

▣ Drafting tips

- Rent-A-Center, West, Inc. v. Jackson, 130 S. Ct. 2772 (2010)
 - Challenge to the enforceability of an agreement (arbitrable) vs. challenge to the agreement to arbitrate
 - Clause: arbitrator, not a court, must resolve disputes over interpretation, applicability, enforceability or formation, including any claim that the agreement or any part of it is void or voidable
- Rahimi v. Nintendo of America, Inc., 936 F. Supp. 2d 1141 (N.D. Cal. 2013)
- Henry Schein, Inc. v. Archer & White Sales, Inc., 139 S. Ct. 524 (2019)
- Tompkins v. 23andMe.com. Inc., 840 F.3d 1016 (9th Cir. 2016)
- Spirit Airlines, Inc. v. Maizes, 899 F.3d 1230 (11th Cir. 2018)
 - Disagreeing with four other circuits, holding that incorporation by reference of AAA rules delegates the issue of whether arbitration may proceed on a class-wide basis to the arbitrator, not the court, if the contract is otherwise silent about whether it provides for individual or class arbitration
 - **But see Stolt-Nielsen S.A. v. AnimalFeeds Int'l Corp.**, 559 U.S. 662 (2010)
 - **But see Lamps Plus, Inc. v. Varela**, 139 S. Ct. 1407 (2019)
- AAA – registration requirement
- Roll into your year end TOS/PP update?
- Review and update frequently – January 1, 2020 is the beginning, not the end of CCPA litigation

AI,
SCREEN SCRAPING,
DATABASE
PROTECTION

AI/ Screen Scraping/ Database Protection

■ AI

- AI and the law
- Licenses (contracts) – like a physical world agent, a software agent acts on behalf of a principal
 - To what extent is a principal liable for the unanticipated actions of an intelligent agent?
 - Indemnification/ waivers
- Ownership issues
 - If copyrighted – who owns the derivative works (address by license – valid assignments/WFH)
 - Sufficient creativity if created by the agent itself (cases on the output of programs)
- Rights of privacy and publicity (in connection with avatars)
- Software agents
 - Smorgasbord of remedies (similar to database protection/ screen scraping)
 - Liability for failing to adhere to contract (TOU), trespass/CFAA, misappropriation, unfair competition, DMCA anti-circumvention

AI/ Screen Scraping/ Database Protection

- Contract/TOU/PP restrictions
 - Assent
 - Indemnification/ Warranties (and how valuable are those)
- Copyright protection
 - Facts vs creative expression
 - Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340, 350 (1991)
 - Reed Elsevier, Inc. v. Muchnick, 559 U.S. 154 (2010).
 - In re Literary Works in Electronic Databases Copyright Litig., MDL No. 1379 (S.D.N.Y. June 10, 2014)
 - Protection for compilations if originality in the selection, arrangement or organization of a database
 - Even if protectable, a database may be entitled to only thin protection
 - Assessment Technologies of WI, LLC v. WIREDATA, Inc., 350 F.3d 640 (7th Cir. 2003)
 - Data mining as a transformative fair use : Author's Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014)
 - Fox News Network LLC v. TVEyes Inc., 883 F.3d 169 (2d Cir. 2018)
- Common law claims, such as misappropriation to the extent not preempted
 - Copyright preemption – 17 U.S.C. § 301
 - International News Service v. Associated Press, 248 U.S. 215 (1918)
 - National Basketball Ass'n v. Motorola, Inc., 105 F.3d 841 (2d Cir. 1997)
 - Barclays Capital Inc. v. Theflyonthewall.com, Inc., 650 F.3d 876 (2d Cir. 2011)
- Interference with contract or prospective economic advantage
- Unfair competition
- Trespass and Conversion
 - trespass to chattels may be based on unauthorized access (plus damage)
 - Register.com, Inc. v. Verio, Inc., 356 F.3d 393, 438-39 (2d Cir. 2004)
 - Intel Corp. v. Hamidi, 30 Cal. 4th 1342, 1 Cal. Rptr. 3d 32 (2003)
 - conversion usually requires a showing of dispossession or at least substantial interference
- Computer Fraud and Abuse Act
 - Federal anti-trespass computer crimes statute
 - Must establish \$5,000 in damages to sue
 - Split of authority on whether exceeding authorized access could be based on an access contract
 - hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985 (9th Cir. 2019)
 - Affirming entry of an injunction prohibiting LinkedIn from blocking hiQ's access, copying or use of public profiles on LinkedIn's website (information which LinkedIn members had designated as public) or blocking or putting in place technical or legal mechanisms to block hiQ's access to these public profiles, in response to a C&D letter sent by LinkedIn
- Anti-circumvention provisions of the DMCA, 17 U.S.C. §§ 1201 et seq.
- Removing, altering or falsifying copyright management information (CMI) - 17 U.S.C. § 1202
- Product liability – software and devices
- Platform liability and immunity

Panel Discussion

3 QUESTIONS

and Q&A...

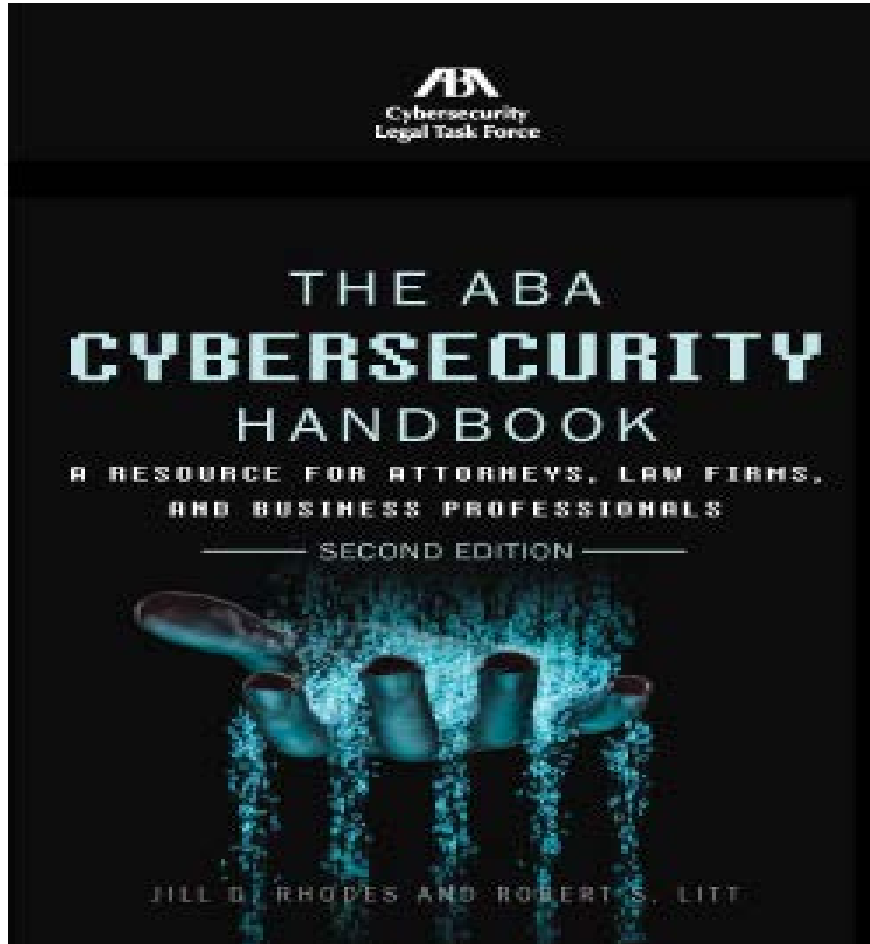
Additional Resources

- *Digital Dangers: A yearlong exploration of cybersecurity and the law*
- A joint production of the *ABA Journal* and the ABA Cybersecurity Legal Task Force
- “What do AI, blockchain and GDPR mean for cybersecurity?” (12/1/18 article)
- Articles (FREE) at abajournal.com/magazine/cyber



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



- Additional Resource: ABA Cybersecurity Legal Task Force's 2018 *ABA Cybersecurity Handbook*, at ambar.org/cybersecurity.
- Winner of 2018 ACLEA Best Publication Award.
- Two chapters from book included in today's materials.
- Inspired 5-part webinar series: "Cybersecurity Wake-Up Call: The Business You Save May Be Your Own" (register for all 5 webinars and get free e-copy of book); go to ambar.org/cyberwakeup.

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Parting AI Privacy/Security Insight

**MAY THE
FORCE BE
WITH YOU**

DATABASE PROTECTION, SCREEN SCRAPING AND THE USE OF BOTS AND ARTIFICIAL INTELLIGENCE TO GATHER CONTENT AND INFORMATION

Excerpted from Chapter 5 (Database Protection, Screen Scraping and the
Use of Bots and Artificial Intelligence to Gather Content and Information) of
E-Commerce and Internet Law: Legal Treatise with Forms 2d Edition
A 5-volume legal treatise by Ian C. Ballon (Thomson/West Publishing, www.IanBallon.net)

**ARTIFICIAL INTELLIGENCE AND ROBOTICS
NATIONAL INSTITUTE
AMERICAN BAR ASSOCIATION
SANTA CLARA, CA
JANUARY 9-10, 2020**

**Ian C. Ballon
Greenberg Traurig, LLP**

Silicon Valley: 1900 University Avenue, 5th Fl. East Palo Alto, CA 914303 Direct Dial: (650) 289-7881 Direct Fax: (650) 462-7881	Los Angeles: 1840 Century Park East, Ste. 1900 Los Angeles, CA 90067 Direct Dial: (310) 586-6575 Direct Fax: (310) 586-0575
---	--

Ballon@gtlaw.com
<www.ianballon.net>
LinkedIn, Twitter, Facebook: IanBallon



Ian C. Ballon

Shareholder

Internet, Intellectual Property & Technology Litigation

Admitted: California, District of Columbia and Maryland
Second, Third, Fourth, Fifth, Seventh, Ninth, Eleventh and Federal
Circuits

U.S. Supreme Court

JD, LL.M., CIPP/US

Ballon@gtlaw.com

LinkedIn, Twitter, Facebook: IanBallon

Silicon Valley

1900 University Avenue

5th Floor

East Palo Alto, CA 94303

T 650.289.7881

F 650.462.7881

Los Angeles

1840 Century Park East

Los Angeles, CA 90067

T 310.586.6575

F 310.586.0575

Ian Ballon is a litigator who is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property & Technology Practice and represents Internet, technology, mobile and other companies in intellectual property and internet- and mobile-related litigation, including the defense of data privacy, security breach, and TCPA class action suits. He is also the author of the leading treatise on Internet law, *E-Commerce and Internet Law: Treatise with Forms 2d edition*, the 5-volume set published by West (www.IanBallon.net). In addition, he is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009). He also serves as Executive Director of Stanford University Law School's Center for the Digital Economy, which hosts the annual Best Practices Conference where lawyers, scholars and judges are regularly featured and interact. A list of recent cases may be found at <http://www.gtlaw.com/Ian-C-Ballon-experience>.

Mr. Ballon was named the Lawyer of the Year for Information Technology Law in the 2019, 2018, 2016 and 2013 editions of Best Lawyers in America. In both 2018 and 2019 he was recognized as one of the Top 1,000 trademark attorneys in the world for his litigation practice by *World Trademark Review*. In addition, in 2019 he was named one of the top 20 Cybersecurity lawyers in California and in 2018 one of the Top Cybersecurity/Artificial Intelligence lawyers in California by the *Los Angeles and San Francisco Daily Journal*. He received the "Trailblazer" Award, Intellectual Property, 2017 from *The National Law Journal* and he has been recognized as a "Groundbreaker" in *The Recorder's* 2017 Litigation Departments of the Year Awards. In addition, he was the 2010 recipient of the State Bar of California IP Section's Vanguard Award for significant contributions to the development of intellectual property law (<http://ipsection.calbar.ca.gov/IntellectualPropertyLaw/IPVanguardAwards.aspx>). He is listed in Legal 500 U.S., The Best Lawyers in America (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also has been recognized by *The Daily Journal* as one of the Top 75 IP litigators in California in every year that the list has been published, from 2009 through 2019, and has been listed as a Northern California Super Lawyer every year from 2004 through 2018 and as one of the Top 100 lawyers in California. Mr. Ballon also holds the CIPP/US certification from the International Association of Privacy Professionals (IAPP).

Chapter 5

Database Protection, Screen Scraping and the Use of Bots and Artificial Intelligence to Gather Content and Information

- 5.01 Database Law—In General**
- 5.02 Copyright Protection for Databases**
 - 5.02[1] Scope of Protection**
 - 5.02[2] Enforcement of Database Copyrights and the Virtual Identity Standard**
- 5.03 Contractual and Licensing Restrictions**
 - 5.03[1] In General**
 - 5.03[2] Database Contract Case Law**
 - 5.03[3] The Scope of Contractual Restrictions**
 - 5.03[4] Forms**
 - 5.03[5] Interference with Contract or Prospective Economic Advantage**
 - 5.03[6] Unjust Enrichment**
- 5.04 Common Law Misappropriation and Unfair Competition**
 - 5.04[1] Misappropriation (including the “Hot News” Doctrine)**
 - 5.04[2] Unfair Competition**
 - 5.04[3] Patent Preemption of State Law Claims**
- 5.05 Trespass and Conversion**
 - 5.05[1] Trespass to Chattels**
 - 5.05[2] Conversion**
- 5.06 Computer Fraud and Abuse Act**
- 5.07 DMCA and BOTS Act Claims**
 - 5.07[1] DMCA Anti-Circumvention Provisions**
 - 5.07[2] Removing, Altering or Falsifying Copyright Management Information**

- 5.07[3] BOTS Act Anticircumvention**
- 5.08 Lanham Act Remedies**
- 5.09 Trade Secret Protection**
- 5.10 EU Database Directive**
 - 5.10[1] Overview**
 - 5.10[2] Copyright Protection for Databases**
 - 5.10[3] *Sui Generis* Protection**
 - 5.10[3][A] In General**
 - 5.10[3][B] Territorial Scope of Protection**
 - 5.10[3][C] Term of Protection**
- 5.11 Sample Injunction Order**
 - 5.11[1] Overview**
 - 5.11[2] FORM**
- 5.12 Anti-Scraping Measures Pursuant to the Cybersecurity Information Sharing Act**
- 5.13 Checklist of Potential Ways to Protect Database Content**
- 5.14 Checklist for Ethical Scraping Practices**
- 5.15 Managing the IP Risks of Artificial Intelligence By Contract**

KeyCite®: Cases and other legal materials listed in KeyCite Scope can be researched through the KeyCite service on Westlaw®. Use KeyCite to check citations for form, parallel references, prior and later history, and comprehensive citator information, including citations to other decisions and secondary materials.

5.01 Database Law—In General

Data and information contained in databases or stored online may be protected from third party use in the United States by a smorgasbord of state and federal laws that may or may not apply, depending on the nature of the database, the type of information copied or used by a third party, how it was accessed, and what is being done with it. The specific requirements to state a claim under potentially applicable laws, and an array of defenses—including fair use, statutory exceptions and various preemption doctrines—potentially limit a database owner’s ability to protect its data and information from unwanted third party use. For these reasons,

screen scraping¹—or the practice of automatically extracting unprotectable facts or other data from third party websites or other locations—if done correctly, is permissible in certain instances. The variety of fact-specific claims and defenses, and evolving circuit splits in some areas of potentially applicable law, make database protection and screen scraping an area where close adherence to the law, and how it is developing, is important. It is also an area where missteps, by either database owners or screen scrapers, can have significant consequences. This chapter addresses the various claims and defenses potentially applicable and provides practical checklists² for database owners and those engaged in lawful scraping. It also addresses E.U. law on database protection.³ Privacy and security issues related to personal information in databases are separately analyzed in chapters 26 and 27, respectively.

This chapter also addresses the practice of using bots or intelligent agents to scrape data or other information from databases. The flipside to thin database protection is that competitors and others may, subject to the various laws analyzed in this chapter, freely access data that is not protected through the various means outlined in this chapter. Businesses typically use bots or intelligent agents to automatically search for and retrieve particular data. The legal regime is largely the same regardless of whether the software agents are preprogrammed to perform a routine task, intelligent agents that are programmed make decisions, or agents using AI. In all cases, the company that deployed the agent is likely to be liable for any misconduct, much in the same way that a business is responsible for any misconduct by an employee acting within the scope of employment. This liability may be varied by contract or subject to indemnification obligations or insurance, but the party that deployed an agent in most cases will be liable for the actions it directed, preprogrammed, or enabled through AI.

Databases are electronic compilations of information.

[Section 5.01]

¹*Scraping* is the programmed extraction of data, usually by a bot or intelligent agent software (often referred to as a *screen scraper*), as opposed to manual copying.

²See *infra* §§ 5.13, 5.14.

³See *infra* § 5.10.

Companies such as Reuters, Reed-Elsevier, Inc., Dow Jones and Dun & Bradstreet spend large sums compiling original databases of useful information that are typically made available to subscribers for monthly or other periodic access fees or charges for specific content in the database, such as a reprint of a single article. The legal protection afforded database content in the United States, however, is limited. Unless a database is comprised of material that itself is independently entitled to copyright protection (such as photographs, articles, music files or video clips), the level of copyright protection for a compilation of otherwise unprotectable material (such as a factual database) likely will be thin. In such cases, database owners may need to rely on a patchwork of other remedies, each one of which typically provides only narrow and limited relief which may or may not protect a database owner in a given case, as laid out in this chapter.

The 1976 Copyright Act created a split copyright interest in compilations.⁴ The owner of a database, which is a compilation of facts or other content, potentially may be entitled to a copyright in the database itself, provided there is sufficient creativity in the selection, arrangement, or organization of the database to merit copyright protection. In addition, the owner or owners of the contributions to the collective work—which could be the same as the owner of the database or could be different people—retain a copyright in their individual contributions, if protectable. Thus, for example, freelance authors who contributed articles to a newspaper may retain their individual copyrights in the articles while at the same time the newspaper may register a copyright for the compilation that includes those articles

⁴See 17 U.S.C.A. § 201(c); *see generally supra* § 4.05[3]. A *compilation* is a work “formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship. The term ‘compilation’ includes collective works.” 17 U.S.C.A. § 101. A *collective work* is “a work, such as a periodical issues, anthology, or encyclopedia, in which a number of contributions, constituting separate and independent works in themselves, are assembled into a collective whole.” *Id.*

Works first published prior to January 1, 1978, are not subject to the 1976 Copyright Act and therefore are not be subject to the split copyright created by section 201(c) for works created on or after January 1, 1978. For a more complete discussion of the scope of section 201(c) and potential *Tasini* issues, *see infra* § 17.03.

(for example, the December 5th edition of *The Miami Herald*). In the absence of an express transfer of the copyright to a contribution to a collective work, the owner of the collective work (in this case, the database) is presumed to have acquired only the privilege of reproducing and distributing the contribution as part of that particular collective work (the December 5th edition), any revision of that collective work (such as the afternoon edition) and any later collective work in the same series.⁵ In *New York Times Co. v. Tasini*,⁶ the U.S. Supreme Court ruled that digitized versions of a newspaper were not “revisions” as that term is used in section 201(c) of the Copyright Act. Thus, freelance authors who had granted permission to the *New York Times* to include their works in the original editions of the paper and, by operation of law, “any revision . . . , and any later collective work in the same series” but who had retained their separate copyright in their articles (or “contributions” to the collective work), were deemed not to have authorized *The New York Times* to reproduce digitized versions of their articles in electronic databases. *Tasini* underscores that different parties may own rights to creative content in a compilation such as a database—such as articles, photographs, music files and the like—and, if so, the owner of the compilation may need to obtain express permission from the owner of a contribution to a collective work when the collective work is reused in new media (such as when preexisting works are included in an electronic database).⁷

By contrast, when a database is comprised of factual information or content that is otherwise unprotectable—such as U.S. government publications or court opinions—there is only one copyright potentially at issue. A database that is a compilation of unprotectable data or information may be subject to copyright protection if there is creativity in the selection, arrangement or organization of the database.⁸ The level of copyright protection, however, is thin, and may not

⁵17 U.S.C.A. § 201(c).

⁶*New York Times Co. v. Tasini*, 533 U.S. 483 (2001).

⁷See *supra* § 4.05[3] (discussing *Tasini* and subsequent cases); *infra* § 17.03 (discussing licensing issues arising out of *Tasini*).

⁸A database is a software application and therefore the application itself also may be entitled to copyright protection (as well as potentially patent or trade secret protection), but this legal protection for the software application will not protect the components of a database (although this additional copyright potentially could be used to prevent unauthorized ac-

protect the owner against all forms of copying.

Since the U.S. Supreme Court's rejection of the "sweat of the brow" doctrine in 1991 in *Feist Publications, Inc. v. Rural Telephone Service Co.*,⁹ the fact that a company may have invested significant time and money creating a database no longer assures it that its work will be protected by copyright law. Many commercial databases are literally large collections of unprotectable factual data efficiently organized to facilitate rapid search and retrieval.¹⁰ Copyright protection may extend to the arrangement and selection of the data, if sufficiently creative, but not to the underlying data itself. While copyright law therefore may protect a database owner from piracy—when the entire database is literally copied and incorporated in a new work or posted at a different location—it typically does not prevent competitors from reviewing a database and copying unprotectable facts or data (so long as the extent of copying does not rise to the level where the competing work is substantially similar or virtually identical¹¹ to the work copied). Indeed, in many instances, it is possible to construct a minimally creative, noninfringing database that incorporates unprotectable facts copied from a rival.

In the process of extracting unprotected data from a website, a screen scraper sometimes must copy protectable content—such as software, photos or creative text or other expression. While the extraction process undoubtedly involves the unauthorized reproduction of protected material, it also could be viewed as fair use intermediate copying,¹² depending on the ultimate use of the material extracted. Other fair use principles also may apply when

cess to the database). *See infra* § 5.03[1].

A database also may include personal information subject to privacy and security laws which are addressed in, respectively, chapters 26 and 27.

⁹*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991); *see generally supra* § 4.02; *infra* § 5.02.

¹⁰*See* Merriam-Webster's Collegiate Dictionary Deluxe Electronic Edition (1995) (defining a database as "a usually large collection of data organized especially for rapid search and retrieval (as by a computer).").

¹¹A number of courts require a showing of virtual identity or heightened substantial similarity in order to establish copyright infringement where the work is a compilation of primarily unprotectable elements and therefore entitled to thin copyright protection. *See infra* § 5.02.

¹²*See Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520–28

material is copied from a database.¹³

In Europe, unlike the United States, databases are entitled to *sui generis* protection analogous to copyright law.¹⁴ U.S.

(9th Cir. 1992); *Nautical Solutions Mktg., Inc. v. Boats.com*, Copy. L. Rep. (CCH) P28,815 (M.D. Fla. 2004) (holding that “momentary copying of open . . . public Web pages in order to extract yacht listings facts unprotected by copyright law constitutes a fair use.”); *Ticketmaster Corp. v. Tickets.com, Inc.*, CV99-7654-HLH (VBKx), 2003 WL 21406289, at *5 (C.D. Cal. Mar. 7, 2003) (“Taking the temporary copy of the electronic information [from the Ticketmaster.com website database] for the limited purpose of extracting unprotected public facts leads to the conclusion that the temporary use of the electronic signals was ‘fair use’ and not actionable.”); see also *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640 (7th Cir. 2003) (citing *Sega* and other cases for the proposition that intermediate copying is a fair use where the only effect of enjoining it would be to give the copyright owner control over noninfringing material produced by a competitor, which was stated in *dicta* as a warning to the plaintiff to not attempt to circumvent the court’s order by reconfiguring its product to make it impossible for customers to extract data without making unauthorized copies); see generally *supra* § 4.10[1] (analyzing intermediate copying as potentially but not always a fair use); *infra* § 5.02.

¹³See generally *infra* § 5.02[2]; *supra* § 4.10[1] (analyzing copyright fair use more extensively).

¹⁴In the mid-1990s, PTO Commissioner Bruce Lehman and others spearheaded efforts to create *sui generis* protection for databases much in the same way that Congress had created a new form of intellectual property protection for semiconductor chips in the early 1980s in the Semiconductor Chip Protection Act. See 17 U.S.C.A. §§ 901 *et seq.* In February 1996, the European Community submitted a proposal to WIPO for international harmonization of database laws, based on the EU Database Directive. See Jack E. Brown, “Proposed International Protection of Electronic Databases,” *The Computer Lawyer*, Jan. 1997, at 17, 19. Similar protection would have been afforded to databases under the Database Investment and Intellectual Property Antipiracy Act, H.R. 3531, 104th Cong. 2d Sess. (1996), which was introduced in Congress in 1996, but not enacted. Critics of both initiatives argued that they would not merely reverse *Feist* with respect to databases, but grant broad *sui generis* protection to collections of otherwise unprotectable facts without the same fair use safeguards otherwise available under U.S. copyright law. See *supra* § 4.10 (analyzing copyright fair use). Critics also charged that the Clinton Administration had sought to have the essential provisions of the failed 1996 database bill made part of U.S. law through negotiation of an international database treaty which would then have been submitted to Congress a *fait accompli*. See, e.g., Pamela Samuelson, “Big Media Beaten Back,” *Wired*, Mar. 1997, at 61. The proposed treaty, however, was never approved by WIPO. In 1998, the Collection of Information Antipiracy Act, H.R. 2652, 105th Cong. 2d Sess. (1998), which included fair use provisions, was passed by the U.S. House of Representatives, but failed to win passage in the U.S. Senate. See David Mirchin, “Putting An End to Database Piracy,” *Boston Globe*, June 2, 1998, at C4. Efforts to create *sui*

residents, however, typically cannot take advantage of this protection for their works in Europe, although it is possible to structure operations such that a European entity could obtain database protection for a work under EU law.¹⁵

In the absence of strong intellectual property protection, U.S. database owners seek to protect the content of their databases by contract. Contractual restrictions in subscription agreements or access licenses, however, only work where there is privity of contract (although potentially, claims could be asserted for interference with contract or prospective economic advantage, to the extent that a third party provides tools to allow a user to breach the database or website access agreement).¹⁶ In addition, where purported licensing restrictions are merely posted on a site, assent may be deemed lacking and no contract formed.¹⁷

Database owners may be able to prevent screen scraping or copying from their databases under state unfair trade statutes or common law theories, such as misappropriation, to the extent not preempted by the Copyright Act, the Patent Act, the Lanham Act, or the Uniform Trade Secrets Act. State law claims based on copying information from a database will be preempted by the Copyright Act (and therefore be unavailable, even if copyright law itself provides no measure of relief because the material copied is merely factual data) unless the database owner can allege at least one additional element (beyond what would be required to state a claim for copyright infringement). If not preempted, databases that provide “hot news”—such as stock quotes or sports scores that have value for their timeliness—may be entitled to protection based on common law misappropriation if the act of copying information also involves the theft of lead time.¹⁸ Where the only state law claim a party may assert is that data was merely copied, the claim likely will be deemed preempted by the Copyright Act.¹⁹ In more limited circumstances, where the content of a database is also a trade secret, other state law claims may be preempted in

generis database protection in the United States ultimately faded over time.

¹⁵See *infra* § 5.10.

¹⁶See *infra* § 5.03[5].

¹⁷See *infra* § 5.03.

¹⁸See *infra* § 5.04.

¹⁹See *infra* § 5.04.

states that have enacted section 7 of the Uniform Trade Secrets Act.²⁰ In even more limited circumstances, state law claims also could be preempted by the Patent Act, where state law presents an obstacle to the execution and accomplishment of patent laws or offers patent-like protection to intellectual property that is inconsistent with the federal scheme.²¹ Claims against platforms, intermediaries or other interactive computer service providers based on user misconduct, or asserted against users for republishing or restricting access to or removing third party material, also may be preempted by the Communications Decency Act.²²

A claim for trespass may be asserted where access to a database is unauthorized. Whether access is unauthorized may turn on the Terms of Use of a site or whether notice was provided. Most courts, however, require a showing of actual injury, such as diminishment of server capacity, which may be difficult to establish.²³ A database owner also may be able to sue for unauthorized access under the Computer Fraud and Abuse Act, an anti-hacking statute, if a minimum of \$5,000 in damages may be shown²⁴ (or aggregated in a class action suit).²⁵ Where access restrictions are circumvented or copyright management information removed, claims potentially could also be brought under the Digital Millennium Copyright Act (DMCA).²⁶

In narrow circumstances involving online event ticket sales, the Better Online Ticket Sales Act (or BOTS Act)²⁷ makes it unlawful to “circumvent a security measure, access control system, or other technological control or measure on an Internet website or online service that is used by the ticket issuer to enforce posted event ticket purchasing limits or to maintain the integrity of posted online ticket purchasing order rules”²⁸

Where material scraped or copied includes logos or other

²⁰See *infra* §§ 5.09, 10.17.

²¹See *infra* § 5.04[3].

²²See *generally infra* § 37.05.

²³See *infra* § 5.05.

²⁴See *infra* § 5.06.

²⁵See *infra* § 44.08.

²⁶See *infra* § 5.07.

²⁷See 15 U.S.C.A. § 45c.

²⁸15 U.S.C.A. § 45c(a)(1)(A); see *generally infra* § 5.07[3].

branding, a claim potentially may be asserted under the Lanham Act.²⁹ Some Lanham Act claims may be preempted, however, under the U.S. Supreme Court's decision in *Dastar Corp. v. Twentieth Century Fox Film Corp.*,³⁰ which could preclude claims where branding information or credits are edited out of material from a factual database or other compilation that is unprotectable, if the Lanham Act claim (or potentially even a state law unfair competition cause of action) amounts to a disguised copyright claim.³¹

A claim also potentially could be brought for trade secret misappropriation, but only where the contents of a database are secret.³² As previously noted, however, where the information at issue is a trade secret, other state law claims, such as common law misappropriation, may be preempted under the laws of those states that have enacted section 7 of the Uniform Trade Secrets Act.³³

Claims asserted against third parties for merely republishing or hosting material (such as advertisements or promotional material for screen scraping tools), as opposed to those directly responsible, or for restricting access to objectionable material, may be preempted by the Good Samaritan Exemption to the Telecommunications Act of 1996 (called colloquially the Communications Decency Act, or CDA).³⁴

In short, the remedies potentially available to database owners under U.S. law tend to be narrow and shallow. Absent copyright protection for the components of a database, an owner's copyright in the database itself, as a factual compilation, will only restrict literal copying that rises to the level of substantial similarity or virtual identity. Protection may be augmented by contract, but only if the actual agreement is binding and not unconscionable. Some courts

²⁹See *infra* § 5.08.

³⁰*Dastar Corp. v. Twentieth Century Fox Film Corp.*, 539 U.S. 23 (2003); see generally *infra* § 5.08.

³¹See *infra* § 5.08.

³²See *infra* § 5.09.

³³See *infra* § 5.09.

³⁴47 U.S.C.A. § 230(c); see generally *infra* § 37.05. As analyzed in section 37.05[5][B], certain claims pertaining to intellectual property, among others, are excluded from the scope of the exemption.

While the CDA preempts claims based on content originating with a third party, it does not insulate a party from its own direct liability for any act or omission.

also are reluctant to enforce either copyright or contract rights that effectively prevent access to material in the public domain. Contract claims likewise may be unavailable in the absence of privity of contract, although a party engaging in screen scraping (or providing its users with the tools to do so) potentially could be sued for tortious interference with contract or prospective economic advantage in some circumstances. Common law misappropriation provides very specific grounds for relief, but only where a claim is not preempted, such as when based on the timeliness of the delivery of information, rather than mere copying. Trespass may be a viable claim where a party accesses a site without authorization, but generally requires a showing of damage to the chattel (not merely a business or competitive injury), such as diminishment of server capacity. The Computer Fraud and Abuse Act may provide relief premised, like trespass, on unauthorized access, but only if a minimum of \$5,000 in damage may be shown. The anti-circumvention provisions of the Digital Millennium Copyright Act may afford additional remedies, but only where copy protection or access controls are circumvented or copyright management information is removed. The Better Online Ticket Sales (BOTS) Act proscribes circumventing a security measure, access control system, or other technological control or measure that is used by a ticket issuer to enforce posted event ticket purchasing limits or to maintain the integrity of posted online ticket purchasing order rules. A claim may be brought under the Lanham Act where a brand is copied or tarnished in connection with screen scraping, but the Lanham Act does not prohibit copying of underlying data. Similarly, relief may be obtained for misappropriation of trade secrets, but only in those instances where a database is comprised of confidential information treated as a trade secret. While one or more of these remedies may suffice to provide relief in a given case, in many instances a screen scraper can structure its conduct to avoid liability under this patchwork of remedies.

Issues raised by the Cybersecurity Information Sharing Act (CISA)³⁵ are summarized in section 5.12 and analyzed in more depth in section 27.04[1.5] in chapter 27.

A checklist of potential issues for rights owners and others is set forth in section 5.13. A form for proposed injunctive relief is set forth in section 5.11. A checklist for ethical scrap-

³⁵6 U.S.C.A. §§ 1501 to 1510; *infra* § 27.04[1.5].

ing practices is set forth in section 5.14.

While this chapter addresses database protection and screen scraping, many databases include personally identifiable information (PII). The laws governing the collection, use and transfer to third parties of PII are analyzed in chapter 26. Laws governing the security of information in databases, in turn, are analyzed in chapter 27.

5.02 Copyright Protection for Databases

5.02[1] Scope of Protection

While factual data generally is not entitled to copyright protection, scraping protected content, even in small quantities, may amount to copyright infringement if the portion copied is not a fair use¹ and the copying is not otherwise permissible. For example, in *The Associated Press v. Meltwater U.S. Holdings, Inc.*,² the court held that a news aggregator's use of the lede, or introductory section of a news story that summarizes the story, which was automatically scraped from targeted sources, including Associated Press licensees, and included in Meltwater's subscription news summaries, was infringing and not a fair use or otherwise justified based on implied license, estoppel, copyright misuse or other defenses.³

Using bots to extract data from websites or databases may raise an array of issues. In *Ticketmaster LLC v. Prestige*

[Section 5.02[1]]

¹Copyright fair use is analyzed in section 4.10[1].

²*The Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537 (S.D.N.Y. 2013).

³In rejecting Meltwater's fair use defense, the court held that the use of AP articles in Meltwater's news summaries was not transformative and that the summaries were substitutes for the genuine works, with subscribers clicking through to the AP articles only 0.08% of the time. The court found that the amount and substantiality of the portion taken also weighed against a finding of fair use because Meltwater's scraping tool automatically took the lede from every AP story (in either 140 or 300 character excerpts) which, depending on the length of the article, amounted to between 4.5% and 61% of a genuine work.

The court rejected Meltwater's implied license defense because consent to copy the lede could not reasonably be inferred from the AP's failure to affirmatively block crawlers using a robots.txt file. As the court explained, "what Meltwater is suggesting would shift the burden to the copyright holder to prevent unauthorized use instead of placing the burden on the infringing party to show it had properly taken and used content."

Entertainment West, Inc.,⁴ however, Ticketmaster sued the defendant for secondary copyright infringement, arguing that the defendant's use of bots to extract data from the Ticketmaster site was so sophisticated that it necessarily had to have copied extensive portions of Ticketmaster's web content and code. Specifically, Ticketmaster alleged that (1) third party Bot Developers developed bots that proved highly capable of purchasing large quantities of tickets on the Ticketmaster.com website, and (2) Ticketmaster's website and mobile app are complex platforms that each contain several layers of protection and security measures. Ticketmaster alleged that developing such capable bots would necessarily require deep study and analysis of the pages and code of Ticketmaster's website and mobile app, which meant that the Bot Developers must have downloaded and stored literal or non-literal elements of Ticketmaster's website and mobile app on their local systems in the course of developing these bots.⁵ The court, in denying defendant's motion to dismiss, agreed that Ticketmaster had at least alleged a plausible claim for copyright infringement.⁶

When material is scraped or copied from a database, as discussed in section 5.01, there potentially may be two sepa-

The court likewise rejected Meltwater's estoppel defense because the AP had no duty to restrict general access to its online content by requiring its licensees to put AP content behind a paywall nor did it have any duty to notify Meltwater that it objected to Meltwater's scraping before filing suit.

⁴*Ticketmaster LLC v. Prestige Entertainment West, Inc.*, 315 F. Supp. 3d 1147 (C.D. Cal. 2018).

⁵More specifically, Ticketmaster alleged that the bots enabled defendants to launch thousands of concurrent and recurring reserve requests for tickets for specific events. The bots were calibrated such that when the reserve request expired, the bot was able to regenerate a new ticket reserve request at a far greater speed than any legitimate human could manage, thus preventing any human from reserving or purchasing the ticket. Moreover, the bots could trade information so that purchases coming from multiple computers would appear to be coming from the same computer, allowing the bots to hide themselves by more closely mirroring what Ticketmaster's algorithms considered normal human use. Finally, the bots were able to escape detection when ordering tickets through the mobile app interface, which Ticketmaster alleged was not possible without first obtaining certain lines of code called security tokens embedded deep within the code of the Ticketmaster mobile app. *See Ticketmaster LLC v. Prestige Entertainment West, Inc.*, 315 F. Supp. 3d 1147, 1163-63 (C.D. Cal. 2018).

⁶*Ticketmaster LLC v. Prestige Entertainment West, Inc.*, 315 F. Supp. 3d 1147, 1159-65 (C.D. Cal. 2018).

rate copyright owners of the contents of the database⁷—the owner of the collective work and, if different, the owner of contributions to that collective work.⁸ A copyright in a compilation generally does not extend to preexisting works included in the compilation, such as articles, photos, or other material. These components, if protectable, are separately copyrightable.⁹ A copyright in a database, as a compilation, merely protects the selection, arrangement, or organization of the material in the database, to the extent sufficiently creative to be entitled to copyright protection.¹⁰ Many databases are comprised of material that is not separately protectable under U.S. copyright law, such as factual information or data, court opinions, government records or other components in the public domain. As discussed in this section, protection for a database that constitutes a compilation of facts or otherwise separately unprotectable material is quite limited under U.S. copyright law, but is potentially available where there is creativity in the selection, arrangement, or organization of the compilation. A form for applying for copyright protection for a database is included in the appendix to chapter 4.

“Facts are not copyrightable, because they lack any degree of creativity. . . . Facts exist and are not created.”¹¹ Purely factual compilations involving no creativity in the selection, arrangement, or organization of data (such as telephone white page directories) are not entitled to copyright protec-

⁷A database program, like the contents of a database, may also be entitled to copyright protection if original and creative. *See supra* § 4.07. As noted by one court, a “database is not simply a shoe box into which all information is thrown. It is, rather, a very structured hierarchy of information.” *Positive Software Solutions, Inc. v. New Century Mortgage Corp.*, 259 F. Supp. 2d 531, 532 n.1 (N.D. Tex. 2003). A copyright in the underlying database software, however, generally will not protect against copying of contents stored in a database.

⁸*Tasini* problems—where the owners of contributions to the collective are different from the owner of the collective work and where adequate electronic rights have not been obtained—are addressed in section 5.01.

⁹*See* 17 U.S.C.A. § 103. Potential registration issues with database compilations are separately analyzed in section 4.08[2].

¹⁰*See* 17 U.S.C.A. § 103.

¹¹*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1181 (9th Cir. 2018).

tion, since “raw facts may be copied at will.”¹² Thus, for example, in *ProCD, Inc. v. Zeidenberg*,¹³ the Seventh Circuit agreed with the lower court’s analysis that a factual database that the defendant made available over the Internet was not entitled to copyright protection, where the database had been copied entirely from the plaintiff’s CD-ROMs, which in turn contained telephone directory listings from all of the white pages published in the United States.¹⁴

Similarly, in *National Basketball Association v. Motorola, Inc.*,¹⁵ the Second Circuit held that the transmission by pager of continuously updated basketball scores did not constitute copyright infringement because the defendants reproduced only facts from the protected broadcasts (the actual scores), “not the expression or description of the game that constitutes the broadcast.”¹⁶ Randomly generated codes¹⁷ and the volume and page numbers assigned to otherwise unprotectable information¹⁸ likewise have been held to lack sufficient originality to be deemed protectable.

Likewise, in *Bikram’s Yoga College of India, L.P. v. Evolu-*

¹²*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 350 (1991). In *Feist*, the Supreme Court rejected the “sweat of the brow” doctrine, holding that hard work in creating a compilation is insufficient to confer copyright protection if the work does not contain the requisite level of creativity to meet statutory requirements. Facts, the Supreme Court emphasized, “do not owe their origin to an act of authorship . . .” and are “not ‘original’ in the constitutional sense.” *Id.* at 347-48. Facts, whether “scientific, historical, biographical, or news of the day” are merely “discovered” or “record[ed]” and are not copyrightable. *Id.* at 347. While facts are unprotectable, a compilation may be protectable if it has “a minimal degree of creativity” in the “selection and arrangement” of the facts, but the level of protection for a factual compilation is “thin.” *Id.* at 348-49.

¹³*ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996).

¹⁴The Seventh Circuit found for the plaintiff based on the enforceability of a consumer software license, and therefore did not extensively address the copyright issue. For an analysis of the case, see *infra* §§ 21.02, 21.03.

¹⁵*National Basketball Ass’n v. Motorola, Inc.*, 105 F.3d 841 (2d Cir. 1997).

¹⁶*National Basketball Ass’n v. Motorola, Inc.*, 105 F.3d 841, 847 (2d Cir. 1997).

¹⁷*See Mitel, Inc. v. Iqtel, Inc.*, 124 F.3d 1366, 1373 (10th Cir. 1997).

¹⁸*See Matthew Bender & Co. v. West Publishing Co.*, 158 F.3d 693 (2d Cir. 1998) (star pagination unprotectable), *cert. denied*, 526 U.S. 1154 (1999). But see *Oasis Publishing Co. v. West Publishing Co.*, 924 F. Supp. 918 (D. Minn. 1996).

tion Yoga, LLC,¹⁹ the Ninth Circuit held that copyright protection could not be obtained for Bikram Yoga's sequence of twenty-six asanas and two breathing exercises, arranged in a particular order, which the plaintiff called the "Sequence." The court explained that "[c]opyright protects only the expression of this idea—the words and pictures used to describe the Sequence—and not the idea of the Sequence itself. Because the Sequence is an unprotectible idea, it is also ineligible for copyright protection as a 'compilation'."²⁰

While facts generally are unprotectable, in limited circumstances, "creative facts"—or "facts" derived from original, creative expression—may be found independently protectable, at least in the Second Circuit.²¹ Study questions used in a film course likewise have been held to meet the minimal threshold of originality to be deemed protectable.²²

Final values, or the products of formula or calculation, also potentially may be protected in limited circumstances. In holding that settlement prices—or the value at the end of the trading day of a particular futures contract for a particular commodity for future delivery at a particular time—were

¹⁹*Bikram's Yoga College of India, L.P. v. Evolution Yoga, LLC*, 803 F.3d 1032 (9th Cir. 2015).

²⁰*Bikram's Yoga College of India, L.P. v. Evolution Yoga, LLC*, 803 F.3d 1032, 1036 (9th Cir. 2015).

²¹*See Castle Rock Entertainment, Inc. v. Carol Publishing Group, Inc.*, 150 F.3d 132 (2d Cir. 1998) (holding defendants liable for copyright infringement for creating the SAT (Seinfeld Aptitude Test), a trivia quiz book which tested readers' recollection of facts from the fictional television series Seinfeld; "unlike the facts in a phone book, which 'do not owe their origin to an act of authorship,' . . . each 'fact' tested by the SAT is in reality fictitious expression created by Seinfeld's authors [The] characters and events spring from the imagination of Seinfeld's authors").

²²*See Faulkner Press, LLC v. Class Notes, LLC*, 756 F. Supp. 2d 1352, 1357 (N.D. Fla. 2010). The court explained:

Although the fact statements are taken from the various films Dr. Moulton showed in class and his questions track the sequence of the films, Dr. Moulton picked only a few facts from each film to include in his film study questions. There may be nothing innovating or surprising about his selection. His selection was possibly random and made solely to ensure that his students were paying attention to the films. Even so, the selection was original because it was not a mechanical or routine arrangement. Dr. Moulton's selection was unique to himself and unlikely to be duplicated by someone else tasked with compiling film study questions. Some creativity was involved. His selection therefore qualifies for copyright protection.

Id.

not entitled to protection because, based on the merger doctrine,²³ the idea of the fair market value of contracts and their settlement value were essentially the same thing, the Second Circuit, in *New York Mercantile Exchange, Inc. v. IntercontinentalExchange, Inc.*,²⁴ addressed, without deciding, the threshold issue of whether settlement values could, in the first place, even be found sufficiently original to be deemed protectable, explaining that:

["]The first person to find and report a particular fact has not created the fact: he or she has merely discovered its existence." . . . [For] example, census takers are not authors of the census data. Census takers merely discover the appropriate population figure; "in a sense, they copy these figures from the world around them." . . . The question then, is one of characterization: does the Committee create the settlement prices, or is it more accurate to view the Committee's task as like that of a census taker, copying the market's valuation of futures contracts? While the line between creation and discovery is often clear-cut, we recognize that it is a difficult line to draw in this case.²⁵

Summarizing earlier case law, Judge Karas of the South-

²³Under the merger doctrine, material will be deemed unprotectable where there are so few ways to express an idea that the idea and expression may be said to have merged. See, e.g., *New York Mercantile Exchange, Inc. v. IntercontinentalExchange, Inc.*, 497 F.3d 109, 116–18 (2d Cir. 2007); *Lathan v. City of Whittier Alaska*, Case No. 3:10-cv-00070, 2011 WL 13115649, at *10-11 (D. Alaska Aug. 4, 2011) (granting summary judgment where plaintiff's method for estimating the power output of a proposed hydropower project from raw data, even if sufficiently original to be potentially entitled to copyright protection, was unprotectable under the merger doctrine); see generally *supra* §§ 4.02 (analyzing the merger doctrine and protectability under the Copyright Act), 4.07 (copyright protection for software code).

²⁴*New York Mercantile Exchange, Inc. v. IntercontinentalExchange, Inc.*, 497 F.3d 109 (2d Cir. 2007), *cert. denied*, 522 U.S. 1259 (2008).

²⁵*New York Mercantile Exchange, Inc. v. IntercontinentalExchange, Inc.*, 497 F.3d 109, 114 (2d Cir. 2007), quoting *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 344, 347 (1991). In *Mercantile Exchange*, the plaintiff had obtained a copyright registration certificate for its database but could not, as a result of the court's holding, prevent an Internet competitor from copying individual settlement values.

The court did not decide whether settlement values contained sufficient originality to be protectable (holding that even if they did, any claim to protection was barred by the merger doctrine). The court explained the process of arriving at settlement values as follows:

A futures contract requires the delivery of a commodity at a specified price at a specified future time, though most contracts are liquidated before physical delivery occurs [S]ettlement prices are used to value the open positions Unlike on a securities exchange, the settlement price may not be the

ern District of New York explained, in *BanxCorp v. Costco Wholesale Corp.*,²⁶ that the final value produced by a formula is unlikely to be entitled to copyright protection where (1) the raw data used to create the final value was comprised of unprotectable facts; (2) the method of converting raw data into the final value was an industry standard, or otherwise widely accepted as an objective methodology;²⁷ and (3) the

final trade, for two reasons. First, because of the nature of trading, it is not always clear which trade was the closing trade Second, . . . [f]or the “outer” months, those further from the trading date, there is often little or no trading on a particular day For high-volume months, settlement prices are based on a formula: “a weighted average of all trades done within the closing range.” . . . For low-volume months, the extent of the . . . creative judgment is disputed.

497 F.3d at 110–11. The majority wrote, in *dicta*, that there was “a strong argument” that settlement prices were unprotectable facts, 497 F.3d at 114, although that argument was weaker for low-volume months, noting that if there was no real market in those months the settlement prices appeared closer to creations (or predictions of expected value). *See* 497 F.3d at 116. In high volume months, by contrast, settlement values were “determinations of how the market values a particular futures contract . . . not how the market should value them or will value them So characterized, there is one proper settlement price; other seemingly-accurate prices are mistakes which actually overvalue or undervalue the futures contract.” 497 F.3d at 115; *see also Woods v. Resnick*, 725 F. Supp. 2d 809 (W.D. Wis. 2010) (holding finance formulas used for financing automobiles were not copyrightable under the merger doctrine and as *scènes à faire*).

²⁶*BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596 (S.D.N.Y. 2010).

²⁷The court explained that

if a scientist knew an object’s mass and the force acting upon the object, this raw data could be converted into the object’s acceleration due to that force by using the “formula” known as Newton’s Second Law of Motion. This use of a formula would merely discover an “empirical reality.” On the other hand, “formulae” that purport to identify the best baseball player based on some weighted composition of batting average, on-base percentage, defensive efficiency, and a myriad of other selective factors, are not discovering “empirical realities.” The difference lies in the originality of the method used to compile or analyze the data.

BanxCorp v. Costco Wholesale Corp., 723 F. Supp. 2d 596, 603 (S.D.N.Y. 2010). The court emphasized that significant weight should be attached to the degree of consensus and objectivity that attaches to the formula. *Id.* “Though at first counter-intuitive, . . . the more acceptance a financial measure obtains (i.e., the more successful it is), the more ‘fact-like’ it becomes. Just as scientific theories start as mere speculation and eventually gain a patina of objectivity, economic indicators that we now rely upon, such as CPI, were once just glimmers in the eyes of economists.” *Id.* at 605 n.7. At the same time, “the formula chosen can be generally accepted and objective enough to constitute a ‘fact’ without being completely

final value attempted to measure an empirical reality.²⁸

Stated differently, Judge Karas explained that to demonstrate that the final values produced by raw data are protectable under copyright law, a plaintiff must show one of the following three things: (1) the raw data used to create the final value was protectable; (2) the method of converting the raw data into a final value was an original (but not necessarily novel) process that is neither widely accepted as objective, nor an industry standard; or (3) the final value did not attempt to measure an empirical result.²⁹

accurate.” *Id.* at 604 n.4 (noting that even Newton’s Second Law of Motion is inaccurate because it fails to account for Einstein’s theory of relativity).

²⁸*BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 604–05 (S.D.N.Y. 2010), citing *New York Mercantile Exchange, Inc. v. IntercontinentalExchange, Inc.*, 497 F.3d 109 (2d Cir. 2007), *cert. denied*, 522 U.S. 1259 (2008), and *RBC Nice Bearings, Inc. v. Peer Bearing Co.*, 676 F. Supp. 2d 9 (D. Conn. 2009).

²⁹*BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 605 (S.D.N.Y. 2010). Based on these tests, Judge Karas held that BanxCorp’s National Average Money Market and CD Rates were unprotectable but that the method of converting raw data to final values involved sufficient minimal originality to be entitled to copyright protection.

Judge Karas found that there were potentially three levels of generality at which plaintiffs could be alleging copyright infringement—raw data, the product of the raw data (the actual averages listed in the BanxQuote Indices, or final value) and the arrangement and presentation of the final values (which he called the “arrangement.”). *Id.* at 602. He characterized both the final value and arrangement as compilations. *See id.* As an illustration, Judge Karas described a hypothetical involving three banks, banks A, B and C, which charged interests rates of 4%, 5% and 6%, respectively, on a given type of account. “These facts are the raw data. The average of these rates, 5%, is the final value. The table or graph containing this, and other, final values, is the arrangement.” *Id.* at 602 n.2.

The court found that the underlying raw data was comprised of unprotectable facts about interest rates charged by certain banks and a variety of economic indicators, akin to the actual trade values at issue in *New York Mercantile* and the physical characteristics of ball bearings in *RBC Nice Bearings*, and therefore unprotectable. Similarly, the court found that the BanxQuote Indices were intended to measure, among other things, rates paid by investors on negotiable certificates of deposit and high yield savings accounts, which were objective facts about the banking market and akin to the attempt to measure the value of settlement prices as they are (not as they should or will be) in *New York Mercantile* or the radial strength of ball bearings in *RBC Nice Bearings*, and therefore the final values were unprotectable.

By contrast, the court found that the plaintiffs had stated a claim based on the method of converting the raw data to final values because,

In a subsequent opinion in the case, Judge Karas held

for purposes of a motion to dismiss, it was plausible to infer that the BanxQuote Indices did not contain simple mathematical averages, but were instead created through judgment being applied to disparate indicators. Among other things, plaintiffs alleged that they exercised discretion over exactly which values to use from within certain categories of indicators (such as “leading banks.”). For purposes of stating a claim, the court held that the allegations were “sufficient to get Plaintiffs to first base.” *Id.* at 607. Finally, the court found that the merger doctrine did not bar plaintiffs’ claim because it was not implausible that the BanxQuote Indices were sufficiently subjective that a wide range of potential final values were possible. *Id.* at 608–09.

Unlike *New York Mercantile* and *RBC Nice Bearings*, *BanxCorp* was decided on a motion to dismiss, rather than a motion for summary judgment, so the court focused on facts alleged, rather than actual evidence.

In a subsequent opinion, however, in considering the evidence in ruling on competing motions for summary judgment, Judge Karas held that final values were unprotectable as facts, tables of weekly averages of interest rates offered by banks were not copyrightable as compilations and the merger doctrine rendered plaintiff’s list of national average rates of interest offered by banks for given financial products unprotectable under the Copyright Act. *BanxCorp v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 292–312 (S.D.N.Y. 2013).

Based on the evidence presented, the court determined that BanxCorp’s national interest rate averages were calculated using simple mathematical averages of reported rates from major banks, with no weighting or other calculations involved. *Id.* at 294. The output of the calculation was a single number that was “the exact mathematical average of the inputted rates as of a particular date.” *Id.* at 298. The averages were called “benchmark rates” or “national average rates,” were described as “current,” “accurate,” and “true,” and were represented to consumers, customers, and the financial media . . . [as] objective facts about average national interest rates as of a particular date.” *Id.* at 298–99.

Citing *dicta* in *New York Mercantile*, Judge Karas explained that “when confronted with raw data that have been converted into a final value through the use of a formula, courts should put significant weight on the degree of consensus and objectivity that attaches to the formula to determine whether the final value is fundamentally a ‘fact.’” *Id.* at 300. Judge Karas elaborated that “[i]f the data purports to represent actual objective prices of actual things in the world—the actual price of an actual settlement contract on a particular day—it is an unprotectable fact; if the data purports to represent an estimated price of a kind of idealized object—for instance, what a hypothetical mint condition 2003 Ford Taurus with approximately 60,000 miles might be worth—then the hypothetical price might be eligible for some form of copyright protection in the right circumstances.” *Id.* at 301. In the case at hand, Judge Karas wrote that “on a spectrum from fact to estimate suffused with judgment and opinion . . . , Plaintiff’s data is legally equivalent to the unprotectable load ratings in *RBC Nice Bearings*, the likely unprotectable settlement prices in *New York Mercantile*, and the likely unprotectable analyst recommendations in *Barclays* . . . [and] unlike the protectable list of estimated prices

of hypothetical used cars at issue in *Maclean Hunter*.” *Id.* at 303. In so ruling, the court rejected the argument that the averages were estimates because they are not based on information from every single financial institution, noting that

no white pages directory lists every single person living in a particular area, or gets every address, phone number, and name exactly right—indeed, the white pages at issue in *Feist* even contained four fictitious listings, inserted to detect copying—but that does not make the white pages a work of opinion regarding who lives in a given area. *See Feist*, 499 U.S. at 344. Likewise, in a case about the census that did not address copyright issues, the Supreme Court acknowledged that no population census can possibly capture everything about the population it surveys with complete accuracy. *See Dep’t of Commerce v. U.S. House of Representatives*, 525 U.S. 316, 322 (1999) (describing the Census Bureau’s methods for compensating for the “undercount,” which is the portion of the population not directly surveyed either in person or by mail). And yet the Supreme Court stated in *Feist* that “[c]ensus data . . . do not trigger copyright” because “[c]ensus takers . . . do not ‘create’ the population figures that emerge from their efforts; in a sense, they copy these figures from the world around them.” *Feist*, 499 U.S. at 347. So too here. Each average at issue in this case is a fact about the world—an “empirical reality”—even though it is in some sense an imperfect representation of some platonic ideal of a “national average bank rate.”

978 F. Supp. 2d at 304. Judge Karas also rejected the argument that the fact that there were several competing companies that measured average rates, all of which regularly computed slightly different final values, meant that the output was anything other than “fundamentally factual in nature.” *Id.* While different companies may consider different indicia relevant to consumers—such as, for example, the interest rate large banks pay on CDs with a \$10,000 minimum deposit—“[t]hese differences do not undermine the conclusion that Plaintiff’s data is fundamentally an attempt to represent an empirical fact about the world.” *Id.* While plaintiff and its competitors used “slightly different inputs” to produce their “national average rate,” “the level of judgment that goes into this decision is both minimal and, more relevant, of a type that does not render the output copyrightable. The differences in output come from the company’s slightly different views about how best to represent empirical, historical reality, given time and resource constraints and the need to simplify reporting and analysis for some audiences.” *Id.* at 305. Minimal judgment based on resource constraints, Judge Karas explained, does not merit copyright protection. *See id.*

With respect to the table of averages, the court held that plaintiff’s list of averages, which was organized by date, lacked sufficient creativity in the selection and arrangement of the data to be protectable as a compilation. *See id.*

The court also held that even if plaintiff’s averages were entitled to be treated as protected expression, the merger doctrine rendered them unprotectable. *See id.* at 308-12. Judge Karas explained that “the range of expression is not wide enough such that, if considered expressions, Plaintiff’s averages would be distinct enough from their idea to prevent application of the merger doctrine” *Id.* at 310. Although average rates compiled by plaintiff and its competitors could vary by as much as 0.59 percentage points, “the crucial point is that their expressive variation

that a series of percentages of national interest rate averages were unprotectable as facts, tables of weekly averages of interest rates offered by banks were not copyrightable as compilations and the merger doctrine rendered plaintiff's list of national average rates of interest offered by banks for given financial products unprotectable under the Copyright Act.³⁰

On the other hand, parts systems—even when elaborately compiled and highly complex—will not be deemed protectable where the parts numbers are factual and the catalogue or database is logical or functional, rather than reflecting creativity in the selection, arrangement, or organization of the parts.³¹ Needless to say, a parts system that organizes

is very low, even negligible, because the purpose of computing and publishing a national average rate is to give the consumer or the customer insight into the fact of what is going on in a national market.” *Id.* at 310-11.

³⁰See *BanxCorp v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 292-312 (S.D.N.Y. 2013). BanxCorp's national interest rate averages were calculated using simple mathematical averages of reported rates from major banks, with no weighting or other calculations involved. *Id.* at 294. The output of the calculation was a single number that was “the exact mathematical average of the inputted rates as of a particular date.” *Id.* at 298. The averages were called “benchmark rates” or “national average rates,” were described as “current,” “accurate,” and “true,” and were represented to consumers, customers, and the financial media. . . [as] objective facts about average national interest rates as of a particular date.” *Id.* at 298-99. Additional details about the case and court's rulings may be found in the preceding footnote.

³¹See, e.g., *Southco, Inc. v. Kanebridge Corp.*, 258 F.3d 148, 152 (3d Cir. 2001) (denying a motion for preliminary injunction where the plaintiff sought to enjoin copying of a nine-digit part numbers assigned pursuant to an elaborate numbering system whereby each fastener was given a unique number, with each digit describing a specific physical parameter of the fastener, which the court held lacked the “modicum of creativity” required for copyright protection because a given number merely resulted from “the mechanical application of the numbering system.”; although Southco had “devoted time, effort, and thought to the creation of the numbering system” the very existence of the system “made it impossible for the numbers themselves to be original.”); *Southco, Inc. v. Kanebridge Corp.*, 390 F.3d 276 (3d Cir. 2004) (*en banc*) (reaffirming this principle and holding, in affirming summary judgment for the defendant, that copyright protection was not available because the parts numbers were rigidly dictated by the rules of the numbering system, and therefore not creative, and analogous to short phrases or titles of works, which lack sufficient creativity to be protectable); *R&B, Inc. v. Needa Parts Mfg. Inc.*, Copy. L. Rep. (CCH) ¶ 28,478 (3d Cir. 2002) (affirming the denial of plaintiff's motion for a preliminary injunction because plaintiff's parts numbers were dictated by its product classification scheme, and not the result of creativ-

like parts together will not be protectable, while one that is creative—for example, combining unrelated parts based on aesthetically pleasing arrangements or to form anagrams—would be unlikely to ever be copied because it would not be useful for its intended purpose (to organize and locate parts).

In other cases, courts have found used car valuations,³² compiled property listings,³³ wholesale prices of used coins,³⁴ Craigslist.org’s compilation of user-submitted classified ad-

ity); *R&B, Inc. v. Needa Parts Mfg., Inc.*, 418 F. Supp. 2d 684 (E.D. Pa. 2005) (accord) (granting defendant’s motion for partial summary judgment); *ATC Distribution Group, Inc. v. Whatever It Takes Transmissions & Parts, Inc.*, 402 F.3d 700 (6th Cir. 2005) (holding that an automobile transmission parts catalog and the individual part numbers identified in the catalog were not copyrightable because (1) the part numbers were unprotectable due to merger and because the process of allocating numbers was not creative, and (2) the catalog did not qualify as a protectable compilation because the arrangement of the parts was based on a prior catalog and the parts were listed in a commonplace and practically inevitable manner).

³²See *CCC Information Services, Inc. v. Maclean Hunter Market Reports, Inc.*, 44 F.3d 61, 67 (2d Cir. 1994) (holding the valuations protectable because they were original creations of Maclean Hunter “based not only on a multitude of data sources, but also on professional judgment and expertise.”).

³³See *Metropolitan Regional Information Systems, Inc. v. American Home Realty Network, Inc.*, 888 F. Supp. 2d 691 (D. Md. 2012) (entering a preliminary injunction based on the court’s finding that the plaintiff was likely to prevail on its copyright infringement claim); see also *Metropolitan Regional Information Systems, Inc. v. American Home Realty Network, Inc.*, 904 F. Supp. 2d 530 (D. Md. 2012) (modifying the injunction and requiring the plaintiff to post a \$10,000 bond).

³⁴See *CDN Inc. v. Kapes*, 197 F.3d 1256, 1259-1261 (9th Cir. 1999) (holding wholesale prices contained in collectible coin guides protectable because, unlike the telephone listings at issue in *Feist*, which were simply provided by the phone company, CDN’s coin valuations were “wholly creative”). As explained by the Ninth Circuit,

CDN’s process to arrive at wholesale prices begins with examining the major coin publications to find relevant retail price information. CDN then reviews this data to retain only that information it considers to be the most accurate and important. Prices for each grade of coin are determined with attention to whether the coin is graded by a professional service (and which one). CDN also reviews the online networks for the bid and ask prices posted by dealers. It extrapolates from the reported prices to arrive at estimates for prices for unreported coin types and grades. CDN also considers the impact of public auctions and private sales, and analyzes the effect of the economy and foreign policies on the price of coins. As the district court found, CDN does not republish data from another source or apply a set formula or rule to generate prices. The prices CDN creates are compilations of data that represent its best estimate of the value of the coins.

vertisements,³⁵ and healthcare ratings and awards given to hospitals³⁶ entitled to copyright protection, while copyright

Id. at 1260. In so ruling, the panel was careful to explain the difference between protectable facts and protectable compilations of unprotectable facts:

Appellant's attempt to equate the phone number listings in *Feist* with CDN's price lists does not withstand close scrutiny. First, Kapes conflates two separate arguments: (1) that the listing, selection, and inclusion of prices is not original enough to merit protection; and (2) that the prices themselves are not original creations. Whether CDN's selection and arrangement of the price lists is sufficiently original to merit protection is not at issue here. CDN does not allege that Kapes copied the entire lists, as the alleged infringer had in *Feist*. Rather, the issue in this case is whether the prices themselves are sufficiently original as compilations to sustain a copyright. Thus Kapes' argument that the selection is obvious or dictated by industry standards is irrelevant.

Id. at 1259; see also *National Football Scouting, Inc. v. Rang*, 912 F. Supp. 2d 985, 990 (W.D. Wash. 2012) (following *CDN* for the proposition that a "numeric expression of a professional opinion can be copyrightable" in holding that football player grades were copyrightable as a compilation of facts, but granting summary judgment for the defendant based on fair use).

Applying *CDN*, a district court held that even where numbers are protectable, copyright law does not extend to protect the fact that an entity won an award or ranked in the top tenth percentile based on the copyrighted ranking system. *Comparion Medical Analytics, Inc. v. Prime Healthcare Services, Inc.*, Case No. 2:14-CV-3448 SVW (MANx), 2015 WL 12746228, at *5-6 (C.D. Cal. Apr. 14, 2015).

CDN has been criticized to the extent it could be read to mean that the prices themselves were compilations. The price estimates may have involved originality and they may be elements of a compilation, "but they are not themselves compilations." *BanxCorp v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 306 n.12 (S.D.N.Y. 2013), quoting James Grimmelmann, *Three Theories of Copyright in Ratings*, 14 Vand. J. Ent. & Tech. L. 851, 862 n.71 (2012).

³⁵See *Craigslist Inc. v. 3Taps Inc.*, 942 F. Supp. 2d 962, 971 (N.D. Cal. 2013). In *3Taps*, the court held that the database of user-submitted classified advertisements maintained at Craigslist.org was protectable where both the compilation and the individual advertisements were minimally creative, but that Craigslist could maintain suit for infringement only for a period of time during which it obtained an exclusive license from users to their classified advertisements, pursuant to its Terms of Use agreement, to which all users were required to assent.

³⁶See *Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1234 (D. Colo. 2009) (denying defendant's motion to dismiss where "Health Grades' healthcare ratings for RWJ and other medical providers are a product of Health Grades' collection of data and information from a variety of sources, which it then analyzes and weighs using its own proprietary methodologies to produce a Health Grades' rating of 1, 3 or 5 stars and/or awards for each healthcare provider reviewed. These ratings and awards are not, therefore, facts 'discovered'

protection was found unavailable for government building codes,³⁷ bearing load data,³⁸ a mathematical model based on the laws of physics,³⁹ “traffic conditions, speed restrictions

by Health Grades . . . , but rather are expressions These ratings only exist because Health Grades has selected, weighed and arranged facts it has discovered to present the collected data in a form . . . that can be used more effectively by the reader to make judgments about providers.”).

In *Comparison Medical Analytics, Inc. v. Prime Healthcare Services, Inc.*, Case No. 2:14-CV-3448 SVW (MANx), 2015 WL 12746228, at *5-6 (C.D. Cal. Apr. 14, 2015), the court conceded that while a system of numeric grades for various hospitals may be protectable as a compilation of facts, an award based on those facts or a finding that an entity ranks in the top tenth percentile based on that ranking did not amount to copyrightable expression. In that case, a company that “grants to hospitals awards, and then sells them the right to publicize the awards . . .” sued a recipient of its awards for “posting news of the awards on its website . . .” without purchasing a proffered license to do so. In so ruling, the court distinguished *Robert Wood Johnson* as a case that focused on plaintiff’s ratings.

³⁷See *Veeck v. Southern Bldg. Code Congress Int’l, Inc.*, 293 F.3d 791, 802 (5th Cir. 2002) (*en banc*) (ruling in favor of a website operator who had pasted the text of model building codes on his site, and holding, subject to a strong dissenting opinion, that these codes were not protectable).

³⁸See *RBC Nice Bearings, Inc. v. Peer Bearing Co.*, 676 F. Supp. 2d 9, 22 (D. Conn. 2009) (granting summary judgment for the defendant; “Although the state of the law regarding the copyrightability of numbers remains unclear, . . . the Court finds the bearing load data at issue in this case to be unprotectable facts [L]oad ratings are mainly a function of the geometry of the bearing and material [C]ertain other ‘life factors’ influence how load ratings are determined for a particular bearing, including tolerances, material cleanliness, lubrication, hardness, and operating temperature, and . . . these factors are enumerated in published industry guidelines [T]he bearing load ratings are essentially a numerical representation of the physical characteristics of a particular bearing While there may be some level of judgment involved in selecting which particular ‘life factors’ to utilize in adjusting the standard load rating calculation, . . . such judgment is very minimal given that the relevant life factors are published in industry guidelines.”).

³⁹*Ho v. Taflove*, 648 F.3d 489, 498-500 (7th Cir. 2011) (affirming summary judgment of non-infringement in favor of a defendant who copied plaintiffs’ mathematical model applying a law of physics). As the court explained:

The Model is an idea. In Professor Ho and Ms. Huang’s own words, the Model “mimic[s] . . . certain behaviors of millions of particles in a photonic device.” Appellants’ Br. 4. That is, the Model attempts to represent and describe reality for scientific purposes. This scientific reality was not created by the plaintiffs. Rather, the Model embodies certain newly discovered scientific principles. Granted, as the plaintiffs note, the Model makes certain hypothetical assump-

and police-monitors” in the Waze crowd-sourced GPS and traffic app,⁴⁰ and average money market and certificate of deposit (CD) rates.⁴¹

While these determinations may be made on summary judgment⁴² or at trial, there are an increasing number of

tions, but those hypothetical assumptions do not render the Model fictitious. Rather, the Model strives to describe reality, and, as conceded at oral arguments, the value of the Model is its ability to accurately mimic nature. See *Gates Rubber Co. v. Bando Chem. Indus., Ltd.*, 9 F.3d 823, 842–43 (10th Cir. 1993) (“The constants in the Design Flex program represent scientific observations of physical relationships concerning the load that a particular belt can carry around certain sized gears at certain speeds given a number of other variables. These relationships are not invented or created; they already exist and are merely observed, discovered and recorded. Such a discovery does not give rise to copyright protection.”). As the Supreme Court put it in *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 347, 111 S. Ct. 1282, 113 L. Ed.2d 358 (1991), “facts do not owe their origin to an act of authorship. The distinction is one between creation and discovery: The first person to find and report a particular fact has not created the fact; he or she has merely discovered its existence.”

Id. at 498-99.

⁴⁰*Phantomalert, Inc. v. Google Inc.*, No. 15-CV-03986-JCS, 2015 WL 8648669, at *10, 14 (N.D. Cal. Dec. 14, 2015) (dismissing plaintiff’s copyright claim where the information that plaintiff alleged defendant copied was “inherently factual, involving ‘traffic conditions, speed restrictions, and police-monitors,’ that is, objective facts that can be discovered and reported”). *But see Phantomalert, Inc. v. Google Inc.*, No. 15-CV-03986-JCS, 2016 WL 879758, at *8-12 (N.D. Cal. Mar. 8, 2016) (granting defendants’ motion to dismiss plaintiff’s amended copyright infringement claim for failing to plausibly allege infringement but finding that, as amended, plaintiff plausibly alleged that “the location of some of the individual Points of Interest, as well as the[ir] overall arrangement . . . , are protectable (at least as a pleading matter).”). In its amended complaint, Phantomalert differentiated between actual driving conditions and “Points of Interest,” alleging that Points of Interest were placed other than at the actual locations, based on judgments made about drivers’ experience and what they would want to know, even if it related to hazards that did not directly affect them. See *id.* at *9. Phantomalert also alleged that its database used a system of categorization that the court found was characterized by some minimal degree of originality. *Id.* at *10. In fact, the categories described (with the possible exception of “dangerous intersections” and “dangerous curves” which potentially might involve some creativity in what locations to include or exclude) appear to be entirely factual (railroad crossings, speed traps, speed cameras, potholes, school zones and red light cameras).

⁴¹See *BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 606 (S.D.N.Y. 2010).

⁴²See, e.g., *NTE, LLC v. Kenny Construction Co.*, No. 14 C 9558, 2016 WL 1623290, at *4-6 (N.D. Ill. Apr. 25, 2016) (granting summary judgment for the defendant after finding that NTE’s copyright extended to the

copyrightability opinions decided in connection with motions to dismiss. Although a court may consider a claim to be *plausible*, which is what is required at the outset of a case when a court evaluates the allegations in connection with a motion to dismiss based on the pleadings, a more detailed analysis later in the case (based on evidence) may reveal a lack of originality in the selection, arrangement or organization of a factual compilation.

In contrast to facts or data, product descriptions, if sufficiently creative, may be entitled to copyright protection. In *MyWebGrocer, LLC v. Hometown Info, Inc.*,⁴³ the Second Circuit affirmed the denial of a preliminary injunction in a case where the plaintiff had created 18,000 product descriptions for a grocery store chain whose website it maintained, which were copied verbatim by the new web host after the grocery store switched hosting services. The Second Circuit concluded that whether the product descriptions met the “minimum level of creativity” required for copyright protection was an issue to be decided on remand by the trial court. Among other things, the court emphasized that because the product descriptions differed from the ones used by the defendant on other grocery sites it hosted a “trier might conclude that MyWebGrocer made creative choices about what to include or exclude in its product descriptions,” thus

“selection, arrangement and coordination” of data in the NTE system, including the particular way in which NTE’s barcodes imbue the data with meaning, but that there was insufficient evidence that this (as opposed to unprotectable data) was copied).

Where summary judgment is sought, a plaintiff must clearly define what the alleged compilation is, where it is not clear from the face of a copyright registration what the work is. See *Cisco Systems Inc. v. Arista Networks, Inc.*, Case No. 14-cv-05344-BLF, 2016 WL 4440239, at *2-4 (N.D. Cal. Aug. 23, 2016) (denying the copyright holder’s motion for summary judgment in a software copyright dispute in part because the plaintiff had not presented evidence of where its alleged compilation had come from or how and when it was compiled, in a case where the plaintiff did not own a single registration for the work but claimed infringement of a compilation composed of pieces drawn from 26 different copyright registrations covering Cisco’s IOS, which the defendant characterized as “a lawyer created construct that simply mirrors. . . [the] copyright infringement allegations. . .” in plaintiff’s Complaint). See generally *supra* § 4.07 (analyzing software copyright protection). In other circuits, but not the Ninth Circuit, a plaintiff generally would need to establish that a work was covered by a registration certificate in order to even state a claim. See *supra* § 4.08[2].

⁴³*MyWebGrocer, LLC v. Hometown Info, Inc.*, 375 F.3d 190 (2d Cir. 2004).

allowing for a finding of infringement. However, the court held that this outcome was not sufficiently likely to merit a preliminary injunction.⁴⁴

On the other hand, in *Incredible Technologies, Inc. v. Virtual Technologies, Inc.*,⁴⁵ the Seventh Circuit affirmed a lower court finding that instructions for a videogame were not protectable or, if protectable, because the creativity was at most “slight” or “less than minimal,” could only have been infringed by a showing of “identical copying.” The appellate panel wrote that, “while there are arguably more ways than one to explain how the trackball system works, the expressions on the control panel . . . are utilitarian explanations of that system and are not sufficiently original or creative to merit copyright protection.”⁴⁶

Whether the contents of a database are separately protectable turns on the level of original, creative expression, as well as other factors analyzed in sections 4.02 and 4.07.

As noted above, where the selection,⁴⁷ arrangement or organization of otherwise unprotectable data (or preexisting material owned by a third party, but not the database owner) reflect some minimal level of creativity, a database will be protectable as a compilation. A compilation will qualify for protection to the extent “selected, coordinated, or arranged in such a way that the resulting whole constitutes an original work of authorship.”⁴⁸

Where a database is all-inclusive it will not be entitled to copyright protection. “In order to obtain copyright protection, a compilation must be guided by principles of selection other than all-inclusiveness. This is because the collection of

⁴⁴*MyWebGrocer, LLC v. Hometown Info, Inc.*, 375 F.3d 190, 193–94 (2d Cir. 2004).

⁴⁵*Incredible Technologies, Inc. v. Virtual Technologies, Inc.*, 400 F.3d 1007, 1013–14 (7th Cir. 2005).

⁴⁶*Incredible Technologies, Inc. v. Virtual Technologies, Inc.*, 400 F.3d 1007, 1013 (7th Cir. 2005); see also *Allen v. Academic Games League of America, Inc.*, 89 F.3d 614, 617–18 (9th Cir. 1996) (applying the merger doctrine to deny protection to expression in game manuals).

⁴⁷As a practical matter, the creativity inherent in the selection of articles or other more expressive works, if genuine selection or arrangement is involved, may be easier to establish than with purely factual data where the selection often serves logical or efficient purposes.

⁴⁸17 U.S.C.A. § 101.

‘all is not a selection.’”⁴⁹ Thus, for example, in *Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*,⁵⁰ the Ninth Circuit held that Experian’s list of compiled pairings of names and addresses in its Consumer View Database (CVD) was entitled to “limited protection.”⁵¹ Experian compiled its pairings from a variety of sources, such as catalogues, purchase data, cable company records, real estate deeds and warranty cards signed by consumers at retail stores. Experian excluded names and address pairings that it believed were not valuable to its clients, such as business addresses and addresses of individuals in prison and the very elderly. Experian also resolved conflicts between data sources, using thousands of “business rules” or algorithms to analyze data from each source and determine which name and address pairing should be included in CVD. Experian kept the data current and regularly updated its business rules. Experian estimated that it spends \$10 million annually to compile and update the CVD.

In holding that Experian’s pairings were entitled to limited copyright protection, the panel explained that “Experian’s selection process in culling data from multiple sources and selecting the appropriate pairing of addresses with names before entering them in the database involves a process of at least minimal creativity. The listings are compiled by first collecting and comparing multiple sources, and then sorting conflicting information through the creation of business rules that Experian created to select from among the conflicts.”⁵²

Creativity in the selection and arrangement of otherwise unprotectable data, according to the Second Circuit, “is a function of (i) the total number of options available, (ii) external factors that limit the viability of certain options and

⁴⁹*Silverstein v. Penguin Putnam, Inc.*, 522 F. Supp. 2d 579, 599 (S.D.N.Y. 2007), quoting *Silverstein v. Penguin Putnam, Inc.*, 368 F.3d 77, 85 (2d Cir.), cert. denied, 543 U.S. 1039 (2004). In *Silverstein*, the district court, on remand, held that a collection of poems was not entitled to copyright protection where essentially all poems by a given author were included in the work. There was thus no creativity in the selection made of what poems to include or exclude from the compilation.

⁵⁰*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176 (9th Cir. 2018).

⁵¹*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1185 (9th Cir. 2018).

⁵²*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1185 (9th Cir. 2018).

render others non-creative, and (iii) prior uses that render certain selections ‘garden variety.’”⁵³ Stated differently, “when it comes to the selection or arrangement of information, creativity inheres in making non-obvious choices from among more than a few options.”⁵⁴ Thus, in *Feist Publications, Inc. v. Rural Telephone Service Co.*,⁵⁵ Justice O’Connor acknowledged that even a purely factual compilation of data, such as a phone book, could be entitled to copyright protection if it incorporated an original selection or arrangement.⁵⁶ The level of protection accorded a factual compilation, however, is “thin” because the underlying facts are unprotect-

⁵³*Matthew Bender & Co. v. West Publishing Co.*, 158 F.3d 674, 682–83 (2d Cir. 1998), *cert. denied*, 526 U.S. 1154 (1999). This test was restated by Judge Preska to provide that a compilation may lack the requisite creativity where (1) industry conventions or other external factors dictate selection so that any person compiling facts of that type would necessarily select the same categories of information; (2) the author made obvious, garden-variety or routine selections; or (3) the author has a very limited number of options available. *O.P. Solutions Inc. v. Intellectual Property Network Ltd.*, No. 96 Civ. 7952 (LAP), 52 U.S.P.Q.2d 1818, 1823, 1999 WL 1122475 (S.D.N.Y. 1999).

⁵⁴*Mathew Bender & Co.*, 158 F.3d 693 682 (2d Cir. 1998), *cert. denied*, 526 U.S. 1154 (1999); *see also id.* at 689 (summarizing case law by noting that compilations were found protectable in cases where “the compiler selected from among numerous choices, exercising subjective judgment relating to taste and value that were not obvious and that were not dictated by industry convention.”). By contrast:

Selection from among two or three options, or of options that have been selected countless times before and have become typical, is insufficient. Protection of such choices would enable a copyright holder to monopolize widely used expression and upset the balance of copyright law.

Id. at 682.

One way to evaluate whether a compilation is protectable therefore “is to consider what . . . competitors would have to do to avoid an infringement claim.” *Id.* In the context of the *West Publishing Co.* court reporters before it, the Second Circuit concluded that:

West’s claim illustrates the danger of setting too low a threshold for creativity or protecting selection when there are two or three realistic options: West lists only the arguing attorneys and city of practice, while *United States Law Week* lists the arguing and briefing attorneys, their firm affiliations and city and state of practice. If both of these arrangements were protected, publishers of judicial opinions would effectively be prevented from providing any useful arrangement of attorney information for Supreme Court decisions that is not substantially similar to a copyrighted arrangement.

⁵⁵*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991).

⁵⁶*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 348 (1991).

able and what is “original” in the constitutional sense is merely the selection and arrangement.⁵⁷

“All that is needed for a finding of sufficient originality is a ‘distinguishable variation’ that is not merely trivial, even if the copyrighted work is based on prior copyrighted or public domain works.”⁵⁸ Thus, compilations assembled somewhat more creatively than the alphabetically arranged listing of all names, address and telephone numbers found in white page telephone directories have been held protectable.⁵⁹ By contrast, “insubstantial, unoriginal, and uncreative” compila-

⁵⁷*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 347-49 (1991).

⁵⁸*Torah Soft Ltd. v. Drosnin*, 136 F. Supp. 2d 276, 286 (S.D.N.Y. 2001), quoting *Re-Alco Industries, Inc. v. National Center for Health Educ., Inc.*, 812 F. Supp. 387, 393 (S.D.N.Y. 1993) (citation omitted).

⁵⁹See, e.g., *CCC Information Services, Inc. v. Maclean Hunter Market Reports, Inc.*, 44 F.3d 61 (2d Cir. 1994) (selection and arrangement of used automobile valuation criteria), *cert. denied*, 516 U.S. 817 (1995); *Key Publications, Inc. v. Chinatown Today Publishing Enterprises, Inc.*, 945 F.2d 509, 512-14 (2d Cir. 1991) (selection of particular businesses in specialized telephone directory for use by New York’s Chinese-American community); *Kregos v. Associated Press*, 937 F.2d 700, 703-06 (2d Cir. 1991) (predictive pitching form based on selection of nine baseball statistics); *Harper House, Inc. v. Thomas Nelson, Inc.*, 889 F.2d 197, 204-05 (9th Cir. 1989) (selection and arrangement of materials in daily organizer); *BUC Int’l Corp. v. International Yacht Council Ltd.*, 489 F.3d 1129, 1145-51 (11th Cir. 2007) (affirming a jury finding that a factual compilation of yachts listed for sale by yacht brokers was entitled to copyright protection); *Craigslist Inc. v. 3Taps Inc.*, 942 F. Supp. 2d 962, 972 (N.D. Cal. 2013) (holding that plaintiff plausibly stated a claim that classified listings, organized first geographically and then in categories of products or services, were protectable as a compilation); *Nielson Co. v. Truck Ads, LLC*, No. 08 C6466, 2011 WL 3857122, at *9-10 (N.D. Ill. Aug. 21, 2011) (finding that “Designated Market Area” maps that divide up television market areas with collected data about the programs viewed was copyrightable and not barred by the merger doctrine because estimation of viewership, unlike census data, is done through sampling and extrapolation and there are multiple ways to express DMA data); *Dataworks, LLC v. Commlog, LLC*, Civil Action No. 09-CV-00528-WJM-BNB, 2011 WL 2714087, at *5 (D. Colo. July 13, 2011) (finding that the “selection, design, and placement of the calendars, daily logs, repair and maintenance logs, collective repair lists . . .” in a blank log book used to record operational information were sufficiently original and creative to warrant copyright protection as compilations); *Madison River Management Co. v. Business Management Software Corp.*, 387 F. Supp. 2d 521, 534-35 (M.D.N.C. 2005) (holding protectable a database containing telephone customer information where the database imposed a new structure on raw data and included metadata enhancements); *O.P. Solutions, Inc. v. Intellectual Property Network Ltd.*, No. 96 Civ. 7952 (LAP), 52 U.S.P.Q.2d 1818, 1999

tions have been held unprotectable.⁶⁰

A data compilation, such as a phone book, may be entitled to copyright protection, even though purely factual, if uniquely arranged in some artistic or creative manner, rather than alphabetically by last name or in some other logical manner.⁶¹ A third party lawfully would be unable to make an exact duplicate of such an arguably creative compilation, although it would be permitted to copy factual information in the compilation and produce its own compilation (either in standard, alphabetical form, or in its own

WL 1122475 (S.D.N.Y. 1999) (calendaring software for lawyers for PTO filings).

⁶⁰*Matthew Bender & Co. v. West Publishing Co.*, 158 F.3d 674, 683 (2d Cir. 1998) (holding unprotectable the following elements of West Publishing Co.'s case reporters & CD-ROMs: captions, courts and date information; attorney listings; subsequent history; and parallel or alternative citations added by West), *cert. denied*, 526 U.S. 1154 (1999); *see also Victor Lalli Enterprises, Inc. v. Big Red Apple, Inc.*, 936 F.2d 671 (2d Cir. 1991) ("lucky numbers" used for gambling, generated by a formula that was standard in the industry); *Warren Publishing, Inc. v. Microdos Data Corp.*, 115 F.3d 1509 (11th Cir. 1997) (plaintiff "did not exercise any creativity or judgment in 'selecting' cable systems to include in its Factbook, but rather included the entire relevant universe known to it . . ."), *cert. denied*, 522 U.S. 963 (1997); *BellSouth Advertising & Publishing Corp. v. Donnelley Information Publishing, Inc.*, 999 F.2d 1436 (11th Cir. 1993) (*en banc*) (holding that categories for organizing material in a yellow pages telephone directory lacked creativity where many of the selected headings were deemed obvious (such as "attorneys" or "banks") and others resulted from standard industry practices), *cert. denied*, 510 U.S. 1101 (1994); *Torah Soft Ltd. v. Drosnin*, 136 F. Supp. 2d 276 (S.D.N.Y. 2001) (holding unprotectable a database version of the Hebrew Bible where the plaintiff's alterations were "nothing more than non-original changes dictated by the technological requirements of Bible code software and the end-user market"; where the replacement of final consonants of Hebrew letters with non-final consonants "required no skill beyond that of a high school . . . student and displayed no originality" and the substitution of various symbols involved "nothing more than a *de minimis* quantum of creativity" and the plaintiff "failed to demonstrate how an asterisk or a pound symbol is any more distinctive than a plus sign or an ampersand."); citations omitted); *Skinder-Strauss Associates v. Massachusetts Continuing Legal Educ., Inc.*, 914 F. Supp. 665, 676 (D. Mass. 1995) ("in compiling a Massachusetts directory of lawyers and judges, . . . the 'selection' of other directory data, including the attorney name, address, telephone and fax numbers, year of bar admission, and so forth are . . . unoriginal and determined by forces external to the compiler.").

⁶¹Needless to say, it is often the logical arrangement of data that makes a compilation most valuable to users, even though this feature may undermine entitlement to protection for the compilation under U.S. copyright law.

unique arrangement) so long as the amount copied (if the protectable feature of the database is the selection of its components) or the features reproduced or distributed in competition with the database (if it is the arrangement or organization of the data that is protectable) does not rise to the level of substantial similarity or virtual identity.⁶²

For online databases, creativity in the selection of incorporated facts is substantially more important than their arrangement. The arrangement of content in a database often is irrelevant (or merely functional—based on the most efficient way to store the data);⁶³ databases generally may be searched by users through multiple different means.⁶⁴ Factual databases therefore potentially may require greater

⁶²Even an entire database potentially could be copied for internal analysis (as opposed to use in a competing database) if this form of intermediate copying was deemed a fair use because undertaken for a purpose that was permissible. See *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520–28 (9th Cir. 1992) (holding intermediate copying to make a noninfringing videogame interoperable a fair use); *Nautical Solutions Marketing, Inc. v. Boats.com*, Copy. L. Rep. (CCH) P28,815 (M.D. Fla. 2004) (holding that “momentary copying of . . . public Web pages in order to extract yacht listings facts unprotected by copyright law constitutes a fair use . . .”); *Ticketmaster Corp. v. Tickets.com, Inc.*, CV99-7654-HLH (VBKx), 2003 WL 21406289, at *5 (C.D. Cal. Mar. 7, 2003) (“Taking the temporary copy of the electronic information [from the Ticketmaster.com website database] for the limited purpose of extracting unprotected public facts leads to the conclusion that the temporary use of the electronic signals was ‘fair use’ and not actionable.”); see also *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640 (7th Cir. 2003) (citing *Sega* and other cases for the proposition that intermediate copying is a fair use where the only effect of enjoining it would be to give the copyright owner control over noninfringing material produced by a competitor, which was stated in *dicta* as a warning to the plaintiff to not attempt to circumvent the court’s order by reconfiguring its product to make it impossible for customers to extract data without making unauthorized copies); see generally *supra* § 4.10[1] (analyzing intermediate copying).

⁶³Non-creative, efficient software routines are not protectable. See *supra* § 4.07.

⁶⁴Databases typically are arranged in some logical (as opposed to creative) manner, to facilitate easy access by users. The arrangement of a database may be protectable if it is genuinely creative, rather than functional. As a practical matter, however, under the virtual identity test, a plaintiff would only be able to protect a creatively arranged database if the identical arrangement were copied. See *infra* § 5.02[2]. It would not be difficult to rearrange the order of data in a database and create a copy that would be equally useful to a user as the original. Only if the creative selection of material were virtually identical would a rearranged, exact copy of the contents of a database be found infringing.

creativity in selection to offset the lack of creative arrangement.

In addition to the selection and arrangements of underlying facts, their organization—such as the fields used in the structure of a database—may be entitled to limited copyright protection. Database fields may be protectable if their selection and organization is original and creative.⁶⁵ Aspects of the organization, selection or arrangement of a database driven by efficiency considerations, however, will be unprotectable.⁶⁶

Copyright owners have long sought to protect their works by including deliberate errors in order to more easily detect acts of infringement. Copying false or inaccurate facts from a database, however, will not necessarily establish copyright infringement. In *Feist*,⁶⁷ for example, the defendant had copied an entire phone book—100% of plaintiff's work, including all false facts—but the Supreme Court nonetheless held for

Even a creatively arranged factual compilation may be entitled to a lower level of protection when digitized because of the difficulty of translating the arrangement exactly from “hard copy” paper to bits and bytes. The arrangement used in a database may not mirror the arrangement used in a preexisting printed work. Hence, the creative aspect of the compilation (such as West Publishing Co.'s arrangement of court opinions) may be lost entirely when the work is digitized and stored in a database. See *Matthew Bender & Co. v. West Publishing Co.*, 158 F.3d 693, 702 (2d Cir. 1998) (“If one browses through plaintiffs’ CD-ROM discs from beginning to end, using the computer software that reads and sorts it, the sequence of cases owes nothing to West’s arrangement [A] copyrighted arrangement is not infringed by a CD-ROM disc if a machine can perceive the arrangement only after another person uses the machine to rearrange the material into the copyrightholder’s arrangement.”), *cert. denied*, 526 U.S. 1154 (1999).

⁶⁵See, e.g., *Harbor Software, Inc. v. Applied Systems, Inc.*, 925 F. Supp. 1042, 1049 (S.D.N.Y. 1996).

⁶⁶See generally *supra* § 4.07 (extensively analyzing efficiency limitations on copyright protection for software and databases). For example, a compilation “may lack the requisite creativity where: ‘(1) industry conventions or other external factors dictate selection so that any person compiling facts of that type would necessarily select the same categories of information; (2) the author made obvious garden-variety, or routine selections, or (3) the author has a very limited number of options available.’” *Silverstein v. Penguin Putnam, Inc.*, 522 F. Supp. 2d 579, 599 (S.D.N.Y. 2007) (quoting earlier cases); see also *Phantomalert, Inc. v. Google Inc.*, No. 15-CV-03986-JCS, 2015 WL 8648669, at *12 (N.D. Cal. Dec. 14, 2015) (quoting *Silverstein*).

⁶⁷*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 362 (1991).

the defendant because copying unprotected elements does not amount to infringement. In assessing copyright infringement, false or inaccurate facts are treated like actual facts and are unprotectable because they lack sufficient originality.⁶⁸ By contrast, creative facts (or facts derived from a fictional work), unlike actual or false facts, may be protectable where the “creative facts” presented cumulatively amount to a derivative work copied from creative expression.⁶⁹

If a database interface is novel, or allows for novel business uses, it may be entitled to patent protection.⁷⁰ Protection for the arrangement or interface of a database, or how it operates, under either patent or copyright law, would not extend to the underlying data.

5.02[2] Enforcement of Database Copyrights and the Virtual Identity Standard

Where a database is comprised of copyrightable contributions such as company reports and analysis, copying even a tiny percentage of the database may be deemed copyright infringement if the parts taken are independently protectable and those aspects of the infringing product that were copied

⁶⁸See, e.g., *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 344 (1991) (seeded false facts, intended to detect copying); *Nester’s Map & Guide Corp. v. Hagstrom Map Co.*, 796 F. Supp. 729, 733 (E.D.N.Y. 1992); *Skinder-Strauss Associates v. Massachusetts Continuing Legal Educ., Inc.*, 914 F. Supp. 665, 675 (D. Mass. 1995); see also *Phantomalert, Inc. v. Google Inc.*, No. 15-CV-03986-JCS, 2015 WL 8648669, at *7, 10 (N.D. Cal. Dec. 14, 2015) (elaborating that seeded, false and inaccurate facts are unprotectable under *Feist*); *BanxCorp v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 304 (S.D.N.Y. 2013) (explaining that erroneous facts are unprotectable under *Feist*; “very often, data fails to be perfectly representative or entirely complete relative to what it is supposed to measure, but the data nevertheless remains fundamentally factual.”).

⁶⁹See *Castle Rock Entertainment, Inc. v. Carol Publishing Group, Inc.*, 150 F.3d 132 (2d Cir. 1998) (holding defendants liable for copyright infringement for creating the “SAT (Seinfeld Aptitude Test)”, a trivia quiz book which tested readers’ recollection of facts from the fictional television series “Seinfeld”; “unlike the facts in a phone book, which ‘do not owe their origin to an act of authorship,’ . . . each ‘fact’ tested by the SAT is in reality fictitious expression created by Seinfeld’s authors . . . [The] characters and events spring from the imagination of Seinfeld’s authors”).

⁷⁰See *infra* § 8.04[3].

are at least substantially similar to that material.¹ The fact that part or all of a defendant's database or product is *noninfringing* will be irrelevant because infringement focuses on the portion copied, not the extent of material that may be genuine.²

On the other hand, if a database is comprised of material that independently is not entitled to copyright protection—

[Section 5.02[2]]

¹*E.g., Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 328–31 (S.D.N.Y. 2010) (entering a permanent injunction and awarding statutory damages, prejudgment interest and attorneys' fees (but only that portion of fees directly and predominantly concerned with the prosecution of plaintiffs' copyright claim, potentially reduced in light of the disparity in resources between the plaintiffs—major investments firms—and the defendant, and defendant's financial condition) in a case where the defendant was held liable for verbatim copying of 17 “sample” reports prior to the time it changed its practices in 2005), *rev'd on other grounds*, 650 F.3d 876 (2d Cir. 2011) (reversing judgment for plaintiffs on their claim for hot news misappropriation; the defendant did not appeal the copyright judgment); *see generally infra* § 5.04 (discussing the case in greater detail in connection with plaintiffs' common law misappropriation claim).

While reports, articles or other longer works that may be included in a database may contain sufficient original and creative content to be deemed protectable, and raw data or pure facts generally do not, in limited circumstances, as noted in section 5.02[1], “facts” or data may be accorded protection if sufficiently original and creative and not otherwise barred from protection by the merger doctrine. *See, e.g., Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1232–38 (D. Colo. 2009) (denying defendant's motion to dismiss plaintiff's copyright infringement claim premised on the defendant hospital allegedly accessing plaintiff's website and assenting to its click-through license agreement more than 200 times and, in violation of the limited license, commercially reproducing, modifying and/or distributing its healthcare provider award and ranking information from plaintiff's website in press releases and other marketing materials); *supra* § 5.02[1] (discussing the case in the context of copyrightability and the merger doctrine).

In *Robert Wood Johnson*, the hospital's own ranking information and awards presumably comprised a small fraction of the data in plaintiff's database. Where the portion copied is protectable, a database owner may maintain a suit for infringement. As discussed below in the balance of section 5.02[2], where the portion copied is not independently protectable and copyright protection is premised on the selection, arrangement or organization of the database itself, suit may be maintained only where so much of the database has been copied that it is substantially similar or virtually identical to the original.

²*See supra* §§ 4.07, 4.08.

such as unprotectable facts³ or raw data—copying portions of the database is unlikely to be actionable under the Copyright Act. Although a factual database may contain the requisite level of creativity to be deemed protectable as a compilation, a copyright in a database could prove of limited value in protecting its constituent parts which, if unprotectable, may be freely copied unless the extent of copying is so great that the allegedly infringing portion is virtually identical (or at least substantially similar) to the portion that was copied. Where protectable, factual compilations generally are entitled to only “thin” protection⁴ because it is the compilation, not its individual components, that is the protectable work. To prevail in litigation, many (but not all)⁵ courts therefore have held that a plaintiff must show virtual identity (or a heightened showing of similarity), rather than merely substantial similarity.⁶ If an entire database has been copied, and the database is deemed protectable, the

³See *supra* § 5.02[1] (discussing protectable and unprotectable facts).

⁴See *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 349 (1991). As explained by the Supreme Court, “[n]otwithstanding a valid copyright, a subsequent compiler remains free to use the facts contained in another’s publication to aid in preparing a competing work, so long as the competing work does not feature the same selection and arrangement.” *Id.* A copyright similarly may be accorded only thin protection where it builds on a particular style of a work. See, e.g., *Zalewski v. Cicero Builder Dev., Inc.*, 754 F.3d 95, 106-07 (2d Cir. 2014) (characterizing as “very thin” a plaintiff’s copyright in colonial home designs where the plaintiff made no attempt to distinguish those aspects of his designs that were original to him from those dictated by the form in which he worked; “Although he undoubtedly spent many hours on his designs, and although there is certainly something of Plaintiff’s own expression in his work, as long as Plaintiff adhered to a pre-existing style his original contribution was slight—his copyright very thin.”).

⁵See, e.g., *BUC Int’l Corp. v. International Yacht Council Ltd.*, 489 F.3d 1129, 1145–51 (11th Cir. 2007) (affirming the district court’s use of the substantial similarity test, rather than virtual identity, in a case involving a factual compilation, because the case did not involve a claim of nonliteral infringement, but also noting that the defendant neglected to raise the potential applicability of the standard until trial—after it approved jury instructions based on the substantial similarity test—even though the Eleventh Circuit case that had approved of the virtual identity standard had been on the books for many years).

⁶E.g., *Incredible Technologies, Inc. v. Virtual Technologies, Inc.*, 400 F.3d 1007, 1013 (7th Cir. 2005) (affirming the lower court’s holding that documentation for a videogame was either unprotectable or not infringing because the creativity at most was “slight” absent a showing of “identical copying”); *Apple Computer, Inc. v. Microsoft Corp.*, 35 F.3d 1435, 1441–43 (9th Cir. 1994), *cert. denied*, 513 U.S. 1184 (1995); *TransWestern Publish-*

defendant may be held liable for infringement. Rarely, however, does a defendant blatantly copy an entire database that it offers to the public in competition with plaintiff's own work. More commonly, certain unprotectable facts are copied.

ing Co. LP v. Multimedia Marketing Associates, Inc., 133 F.3d 773, 776–77 (10th Cir. 1998) (compilation); *MiTek Holdings, Inc. v. Arce Engineering Co.*, 89 F.3d 1548, 1558–59 n.24 (11th Cir. 1996) (holding that “virtual identicality” must be shown for a plaintiff to prevail on a claim of infringement of a compilation of nonliteral elements of a computer program; substantial similarity must be shown for other aspects of a work); *see also* *Matthew Bender & Co. v. West Publishing Co.*, 158 F.3d 674, 704–05 (2d Cir. 1998) (applying a heightened test for substantial similarity for factual compilations which required a showing of “very similar literal ordering or format” and/or extensive verbatim copying), *cert. denied*, 526 U.S. 1154 (1999).

Even where a court applies the traditional substantial similarity test, it may be difficult for a plaintiff to prevail in a suit for infringement of a database comprised of unprotectable elements entitled to protection based on the selection, arrangement or organization of the work, where less than the entire database is copied. *See, e.g.,* *Ross, Brovins & Oehmke, P.C. v. Lexis Nexis Group*, 463 F.3d 478, 482–83 (6th Cir. 2006) (holding that although the developer's selection of legal forms in a compilation was sufficiently creative to warrant copyright protection, the selections made by the developer and software designer were not similar enough to be actionable where 61% of plaintiff's forms (or 350 out of 576) were used by the defendant; “First, Lexis did not include a sufficiently large percentage of the same forms to permit a finding of copying. Second, nonquantitative aspects of the two compilations support the conclusion that Lexis created a new work rather than a copy of LawMode's.”).

In *Expert Pages v. Buckalew*, No. C–97–2109–VRW, 1997 WL 488011 (N.D. Cal. Aug. 6, 1997), plaintiffs Expert Pages and Advice and Counsel Corp., which operated a website where litigation consultants advertised their services, sued a Virginia man who was alleged to have made a complete, unauthorized copy of plaintiff's website in order to be able to contact each of plaintiffs' advertisers by email to invite them to advertise on a competing site that he had established. The court granted the defendant's motion to dismiss for lack of personal jurisdiction in the interests of justice, based on the court's determination that it would have been unreasonably burdensome for the defendant—a young man—to litigate in California, when compared to the burden imposed on plaintiffs, which were companies owned by a practicing attorney. Although the court did not address the merits of plaintiffs' claim, it appears likely that they alleged verbatim copying of their website since they otherwise would have had difficulty challenging the defendant's act of copying the names and email addresses of individual advertisers. Whether a defendant's efforts to replicate a commercial database by systematically contacting each paying advertiser listed in it (or offering them free inclusion in the competing database) could constitute an unfair trade practice or common law misappropriation would likely depend on the effect of such competition on the original database, among other things.

For example, a database owner may review a competitor's database and extract those components not already in its own work. If a defendant merely copies portions of the database—such as unprotectable facts, public domain material or licensed articles—the defendant's acts of copying may not amount to infringement because of the limited amount copied and the fact that a “thin” copyright in a compilation will only protect the compilation as a whole.

At what point permissible copying of unprotectable facts from a protectable database rises to the level of infringement is difficult to pinpoint in the abstract, but the extent of copying must be substantial. Indeed, because a purely factual database is entitled to such a low level of protection, the extent of copying that must be shown before infringement will be found is much greater than when more creative works are plagiarized.

In *Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*,⁷ the Ninth Circuit held that Experian's list of compiled pairings of names and addresses in its Consumer View Database (CVD) was entitled to “limited protection” based on “Experian's selection process in culling data from multiple sources and selecting the appropriate pairing of addresses with names before entering them in the database”⁸ But the Ninth Circuit characterized the scope of protection for this factual work as “severely limited.”⁹ Applying the virtual identity test, the panel held that a match rate of 80% with the defendant's database was “insufficient to establish a bodily appropriation of Experian's work.”¹⁰

In *Assessment Technologies, LLC v. WIREDATA, Inc.*,¹¹ Judge Posner of the Seventh Circuit held that a database comprised of 456 fields grouped into thirty-four separate

⁷*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176 (9th Cir. 2018); *supra* § 5.02[1].

⁸*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1185 (9th Cir. 2018).

⁹*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1186 (9th Cir. 2018).

¹⁰*Experian Information Solutions, Inc. v. Nationwide Marketing Services Inc.*, 893 F.3d 1176, 1187 (9th Cir. 2018).

¹¹*Assessment Technologies of WI, LLC v. WIREDATA, Inc.*, 350 F.3d 640 (7th Cir. 2003).

tables contained sufficient creativity to be protectable,¹² but nonetheless ruled that the defendant was entitled to freely copy the data stored in the database from municipal governments (even though the plaintiff claimed that its copyright extended to this data) where the data had been collected and inputted by municipal tax assessors, not the plaintiff, and Wisconsin's open records law required that data be provided upon request unless entitled to copyright protection. Judge Posner wrote:

[I]f WIREdata said to itself, "Market Drive is a nifty way of sorting real estate data and we want the municipalities to give us their data in the form in which it is organized in the database, that is, sorted into AT's 456 fields grouped into its 34 tables," and the municipalities obliged, they would be infringing AT's copyright because they are not licensed to make copies of Market Drive for distribution to others; and WIREdata would be a contributory infringer (subject to a qualification concerning the fair-use defense . . .). But WIREdata doesn't want the compilation as structured by Market Drive It only wants the raw data, the data the assessors inputted into Market Drive.¹³

Judge Posner explained that, because the process of extracting the data did not involve making an unauthorized copy or a derivative work, the municipalities were free to do so.¹⁴ He clarified that:

It would be like a Westlaw licensee's copying the text of a

¹²The court concluded that the "modest requirement [that the work involve sufficient originality to distinguish it from material in the public domain] is satisfied . . . because no other real estate assessment program arranges the data collected by the assessor in these 456 fields grouped into these thirty-four categories, and because the structure is not so obvious or inevitable as to lack the minimal originality required." *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 643 (7th Cir. 2003).

¹³*Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640 (7th Cir. 2003).

¹⁴Judge Posner wrote in *dicta* that even if an unauthorized copy were made, it would almost certainly be a fair use intermediate copy. See *Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 644 (7th Cir. 2003), citing *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520–28 (9th Cir. 1992); see generally *supra* § 4.10[1] (analyzing intermediate copying as potentially but not always a fair use).

Judge Posner further warned that if the plaintiff tried to circumvent the decision by reconfiguring its database "in such a way that the municipalities would find it difficult or impossible to furnish the raw data to requesters . . . in any format other than that prescribed by [the plaintiff] . . . it might be guilty of copyright misuse." 350 F.3d at 645; see generally *infra* § 16.04[3] (analyzing copyright misuse). He further sug-

federal judicial opinion that he found in the Westlaw opinion database and giving it to someone else. Westlaw's compilation of federal judicial opinions is copyrighted and copyrightable because it involves discretionary judgments regarding selection and arrangement. But the opinions themselves are in the public domain . . . and so Westlaw cannot prevent its licensees from copying the opinions themselves as distinct from the aspects of the database that are copyrighted.¹⁵

Similarly, in *Hutchins v. Zoll Medical Corp.*,¹⁶ the Federal Circuit, applying First Circuit law, held that plaintiff's compilation copyright did not protect individual words and "fragmentary" phrases when removed from their form of presentation and compilation. The court explained that, "[a]lthough the compilation of public information may be subject to copyright in the form in which it is presented, the copyright does not bar use by others of the information in the compilation." In *Zoll*, the district court had found that the words and phrases on Mr. Hutchins' "Script and Word List" were standard CPR instructions devoid of "creative expression that somehow transcend the functional core of the directions"

In *American Massage Therapy Association v. Maxwell Petersen Associates, Inc.*,¹⁷ a district court in Illinois held that a defendant could not be held liable for copying data from a database of massage therapists, even though the database possessed sufficient creativity to be deemed protectable because, in addition to name, address and telephone number, which were "'entirely typical' of a directory, the listing of the membership category and type of therapist pro-

gested that the plaintiff was "trying to use its copyright to sequester uncopyrightable data, presumably in the hope of extracting a licensing fee from WIREdata." 350 F.3d at 645.

¹⁵*Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 644 (7th Cir. 2003). This analogy may be imperfect because subscription databases typically restrict use by license, which may be permissible to the extent that the *quid pro quo* is access to a database. See *infra* § 5.03[1].

WIREdata is perhaps best understood as a case where, in the absence of privity of contract, a database owner could not restrict access to otherwise unprotectable data.

¹⁶*Hutchins v. Zoll Medical Corp.*, 492 F.3d 1377, 1383–84 (Fed. Cir. 2007).

¹⁷*American Massage Therapy Association v. Maxwell Petersen Associates, Inc.*, 209 F. Supp. 2d 941 (N.D. Ill. 2002).

duces a sufficiently creative selection to make it original.”¹⁸ The court emphasized that copyright protection extends only to those components of a work that are original to the author and thus the defendant’s copying of unprotectable facts did not amount to infringement. It wrote that “[p]laintiff may have been the first to discover and report the names and addresses but this data does not ‘ow[e] its origin’ to plaintiff.”¹⁹ Further, while the selection of data to be included in the database was original, the court found the arrangement and organization—listing names geographically—were not. Moreover, the fact that the plaintiff could have arranged the database in a different form that would have been creative did “not elevate the listing [i.e., the factual data] to the level of creative.”²⁰

In *Snap-on Business Solutions Inc. v. O’Neil & Associates, Inc.*,²¹ the court denied the defendant’s motion to dismiss, holding that Snap-on had presented sufficient evidence to show disputed facts on the issues of protection and infringement in a screen scraping case. In *Snap-on*, O’Neil, a competitor, repeatedly accessed Snap-On’s database to copy data for Mitsubishi, a customer who was trying to transition from Snap-On’s database hosting service to O’Neil, where the issue of whether Mitsubishi was authorized to allow O’Neil to access the database on its behalf was disputed.

The court found that Snap-on had presented evidence that it owned valid copyrights in its Net-Compass software, improvements to Mitsubishi’s data and in the proprietary database used to run the software.

With respect to copying, Snap-On alleged that O’Neil’s scraper program copied protectable elements of Snap-on’s database, including the link structure and navigational element on the left-hand of the site. The court, however, denied summary judgment, finding that, among other things, what information had actually been copied was disputed by the parties.

¹⁸*American Massage Therapy Association v. Maxwell Petersen Associates, Inc.*, 209 F. Supp. 2d 941, 948 (N.D. Ill. 2002).

¹⁹*American Massage Therapy Association v. Maxwell Petersen Associates, Inc.*, 209 F. Supp. 2d 941, 947 (N.D. Ill. 2002), quoting *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340, 361 (1991).

²⁰209 F. Supp. 2d at 949.

²¹*Snap-on Business Solutions Inc. v. O’Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 683–86 (N.D. Ohio 2010).

Snap-on eventually obtained a general jury verdict, although it is not clear whether the verdict was based on Snap-On's claim for copyright infringement or for its other claims for trespass, breach of contract (based on its EULA), or violations of the Computer Fraud and Abuse Act.²² The case is discussed in greater detail in section 5.05 in connection with Snap-On's trespass claim.

Even where a database includes creative elements such as photographs (where the selection, arrangement and organization arguably involves greater creativity than with factual data) the database owner may be powerless to prohibit copying if it does not own the copyrights to the individual components of the database that are copied, and the amount copied is less than the entire work. Indeed, a database owner may be unable to prevail in an infringement action if copying is undertaken for a fair use purpose (rather than to merely offer the same database to the public in competition with the database owner's product, which plainly would be infringing). For example, in *National Football Scouting, Inc. v. Rang*,²³ a district court in Washington found protectable grades assigned to different football players, but held that the use of a small number of these grades in connection with commentary and analysis was a fair use.

Intermediate copying of even an entire database may be deemed a fair use if undertaken for a lawful purpose such as extracting unprotectable data. In *Nautical Solutions Marketing, Inc. v. Boats.com*,²⁴ for example, the plaintiff sought and obtained a declaration that its copying of the defendant's database was a fair use. In that case, the defendant, *Boats.com*, owned and operated *Yachtworld.com*, a website that listed yachts available for sale. Each listing showed pictures with a description provided by the yacht broker who posted it. The descriptions used industry-standard headings

²²See *Snap-On Business Solutions Inc. v. O'Neil & Associates, Inc.*, No. 5:09-CV-1547, 2010 WL 2650875 (N.D. July 2, Ohio 2010) (awarding costs but denying Snap-On's request for an award of attorneys' fees because under Ohio law contractual attorneys' fee provisions are unenforceable as contrary to public policy because they are viewed as encouraging litigation).

²³*National Football Scouting, Inc. v. Rang*, 912 F. Supp. 2d 985, 991-95 (W.D. Wash. 2012) (entering summary judgment for the defendant).

²⁴*Nautical Solutions Marketing, Inc. v. Boats.com*, No. 8:02-CV-760-T-23TGW, Copy. L. Rep. (CCH) P28,815, 2004 WL 783121 (M.D. Fla. Apr. 1, 2004).

such as “electrical,” “accommodations,” “galley,” and “sails and rigging.” Yachtworld.com’s listings had a distinctive look-and-feel: pictures of the yachts always appeared to the left of the description, the basic facts were shown in bullet-points, and a blue wave appeared on the left side of the screen.

Plaintiff Nautical Solutions (“Nautical”) operated a competing website, *Yachtbroker.com*. Nautical generated listings by using a spider program to make temporary copies of Boats.com’s listings. Nautical extracted the descriptions and pictures from the temporary copies it created, discarded those copies, and then used the extracted information to create its own listings. Nautical also offered a “valet service” in which, with the yacht broker’s permission, it copied descriptions and pictures from the broker’s listings on other websites, such as *Yachtworld.com*, and pasted this information into *Yachtbroker.com*. *Yachtbroker.com*’s appearance differed from that of *Yachtworld.com*: the pictures were to the right of the facts, the facts were in a table, and there was no blue wave shown.

Nautical sought a declaratory judgment of non-infringement after *Boats.com* accused Nautical of violating its copyright in the *Yachtworld.com* website. With regards to the valet service, the court noted that *Boats.com* did not hold the copyrights to the individual pictures and descriptions—the brokers who created the listings did.²⁵ *Boats.com* likewise was held not to be entitled to a copyright for the organization of the descriptions because its use of industry-standard

²⁵*Nautical Solutions Marketing, Inc. v. Boats.com*, Copy. L. Rep. (CCH) P28,815, at 2004 WL 783121, at *9 (M.D. Fla. Apr. 1, 2004). The copyright owners of the individual photos included in the Boats.com database could have maintained claims for infringement of their separate copyrights in their individual photos, if they had wanted to do so. As a practical matter, however, these copyright owners would be unlikely to want to sue where the photos were placed online to generate potential sales. Given the purpose for which the photos were taken, it is also unlikely that the individual copyright owners registered their copyrights, which is a precondition to filing suit (although they could do so at any time). See *supra* § 4.08[2]. In addition, because the yacht owners had given permission to Nautical Solutions to copy their Boats.com listings, Nautical would have been able to assert an implied license defense if the individual brokers had then turned around and sued Nautical (although permission from individual brokers would not have supported an implied license to copy the compilation, in which Boats.com, not the individual owners, owned the relevant copyright). See *supra* §§ 4.05[7] (implied license), 5.01 (split copyrights in compilations).

headings lacked creativity. Further, while Boats.com's copyright did extend to the website's distinctive look-and-feel, Nautical's website had its own unique look-and-feel and had not copied Boat.com's.²⁶

The court held that Nautical's copying of *Boats.com* broker listings using a spider program constituted a fair use.²⁷ Unlike the valet service, the spider program copied the entire website, including its protected look-and-feel. The court nonetheless deemed this allowable intermediate copying because the spider program only extracted unprotected facts from the copied site.²⁸ The court found no evidence of harm to the "potential market value for or value of Yachtworld.com," and stressed that the "amount and substantiality of the portion used" was minimal, since Nautical's final product was free of infringing material.²⁹

Fair use also was influential in *NTE, LLC v. Kenny Construction Co.*,³⁰ an unreported opinion in which the court granted summary judgment for the defendant on a copyright infringement claim based on the defendant's having accessed plaintiff's database to extract its own raw data. Citing *Assessment Technologies, LLC v. WIREdata, Inc.*³¹ for the proposition that it could constitute "copyright misuse" to prevent a company from using its own data, and *Sega Enterprises Ltd. v. Accolade, Inc.*³² on fair use intermediate copying, the court explained that in *NTE*, "after extracting the NTE reports, Kenny ended up in possession only of data that it undeniably owns or is in the public domain, which is to say the facts pertaining to the location of Kenny's materials across time. The contested data was input into the NTE system by Kenny in the first place and does not cease to belong to Kenny just because it is manipulated by a copy-

²⁶Copy. L. Rep. (CCH) P28,815, at *9–11.

²⁷Copy. L. Rep. (CCH) P28,815, at *7–8.

²⁸Copy. L. Rep. (CCH) P28,815, at *9.

²⁹Copy. L. Rep. (CCH) P28,815, at *8.

³⁰*NTE, LLC v. Kenny Construction Co.*, No. 14 C 9558, 2016 WL 1623290, at *5 (N.D. Ill. Apr. 25, 2016).

³¹*Assessment Technologies of WI, LLC v. WIREdata, Inc.*, 350 F.3d 640, 646–47 (7th Cir. 2003).

³²*Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1520–28 (9th Cir. 1992).

righted software system.”³³

Since fair use is based on a multipart balancing test,³⁴ not all intermediate copying will be deemed fair. Only intermediate copying undertaken for a lawful purpose will be deemed permissible. Because fair use is not determined by a bright line test, companies or individuals who build business models based on fair use may nonetheless get sued for infringement (as well as any other claims that might be brought).³⁵

5.03 Contractual and Licensing Restrictions

5.03[1] In General

Owners of commercial databases generally seek to protect their rights through end-user license agreements (EULA), database access, use or subscription agreements or similarly-termed contracts or licenses.¹ These agreements typically may include terms prohibiting commercial use, the use of bots or other automated means to access a site or extract data, repeated access to a database, reverse engineering and unauthorized use or access, among other provisions, in addition to disclaiming liability and otherwise generally establishing the terms and conditions of use.² Whether and to what extent these purported use restrictions are effective depends on whether a binding contract has been formed, whether the agreement is part of a broader intellectual property license or merely a stand-alone data contract, whether the agreement is deemed enforceable, and the express or implied rights and restrictions set forth in it. Contract claims generally will not be preempted by the Copyright Act where an additional element (beyond mere copying)—such as the contractual obligation itself—has been alleged.³

Where parties negotiate the terms of a database access

³³*NTE, LLC v. Kenny Construction Co.*, No. 14 C 9558, 2016 WL 1623290, at *5 (N.D. Ill. Apr. 25, 2016).

³⁴*See supra* § 4.10[1].

³⁵*See infra* §§ 5.03 to 5.09, 5.11.

[Section 5.03[1]]

¹The difference between an intellectual property license and a mere contract is addressed in chapters 14 and 16.

²*See infra* chapter 22 (discussing Terms of Use for database and other sites and services).

³*See, e.g., ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447, 1452–53 (7th Cir. 1996) (holding the plaintiff’s breach of contract claim based on a

agreement or where a written or electronic signature is obtained on a contract,⁴ formation issues do not arise. However, where access to or use of a database is conditioned on a EULA or Terms posted on a website and accessed from a computer or mobile phone, whether the terms are deemed to form a binding contract in practice may depend on whether they are presented as a click-to-accept contract or otherwise structured so that express assent is obtained.⁵

To be enforceable, a user must assent to a database license or contract, either expressly or impliedly based on notice and subsequent conduct (in using the database).⁶ As analyzed

shrinkwrap license for a CD containing phone directly listings not preempted); *BanxCorp v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 315-16 (S.D.N.Y. 2013) (holding that plaintiff's claim for breach of contract was not preempted); *BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 611-18 (S.D.N.Y. 2010) (holding that claims for breach of a license agreement and misappropriation based on hot news were not preempted in a case alleging that the defendant, a licensee, misused money market and CD data, but claims for unfair competition and unjust enrichment were preempted); *Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1242-47 (D. Colo. 2009) (holding plaintiff's claim that the defendant hospital breached its click-through license agreement with the plaintiff by commercially reproducing, modifying and/or distributing its healthcare provider award and ranking information from plaintiff's website in press releases and other marketing materials was preempted, but that its breach of contract claim based on unauthorized use of the plaintiff's mark in a way that implied that the owner endorsed the hospital's services was not preempted); *Internet Archive v. Shell*, 505 F. Supp. 2d 755, 763-64 (D. Colo. 2007) (holding that breach of contract and conversion claims arising out of a site owner's objection to her site being copied for inclusion in the Internet Archive's Wayback machine were not preempted); *Huckshold v. HSSL, LLC*, 344 F. Supp. 2d 1203 (E.D. Mo. 2004) (holding trade secret misappropriation and breach of contract claims not preempted where the plaintiff alleged that the defendant owed a duty to protect the confidentiality of plaintiffs' trade secrets and breached its contract by allowing a third party to copy the software in violation of their agreement (and not merely that the defendant itself copied the software), which thus involved an extra element, but finding plaintiff's tortious interference claim preempted where the only element needed to be shown to establish liability was copying); see generally *supra* § 4.18[1] (analyzing copyright preemption).

⁴See *infra* § 15.02 (electronic signatures).

⁵Case law and strategies for maximizing the potential enforceability of a unilateral agreement online are set forth in chapters 21 and 22.

⁶Contract formation for unilateral Internet contracts is addressed extensively in chapters 21 and 22. While cases involving databases are discussed in this section, exclusive consideration of database formation case law could lead to a skewed view of the state of the law of contract

extensively in section 21.03, unilateral online contracts are much more likely to be found binding where express assent is obtained, such as through a click-through contract (even though theoretically, a contract should be equally enforceable where it is formed based on implied assent such as conduct in the face of actual or imputed notice). While posted Terms of Use may be enforced where a defendant acknowledges that it had actual notice of the terms and proceeded to access a database or site thereafter,⁷ implied assent is more difficult to prove in court absent this type of admission. Where Terms merely have been posted online and the defendant disputes knowing that it was bound by those terms, courts may be reluctant to find that a binding contract has been formed.⁸

formation generally. Readers are encouraged to review chapters 21 and 22 in drafting, evaluating or litigating database agreements.

⁷See, e.g., *Meyer v. Uber Technologies, Inc.*, 868 F.3d 66 (2d Cir. 2017) (enforcing an online arbitration agreement where the company provided reasonable notice of the terms and the consumer manifested assent); *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393 (2d Cir. 2004) (holding that the district court was within its discretion in finding that the plaintiff was likely to prevail on the merits for purposes of granting a preliminary injunction where the defendant received actual notice of purported restrictions on access to a database but continued to repeatedly access the database on a daily basis even after receiving notice); *Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007) (holding that the defendant was bound by posted Terms that formed a non-exclusive license to access Ticketmaster's website where the defendant acknowledged that it was on notice that its access to the site was subject to Terms); *Southwest Airlines Co. v. BoardFirst, LLC*, No. 3:06-cv-0891-B, 2007 WL 4823761 (N.D. Tex. Sept. 12, 2007) (holding that the defendant, operator of a site that offered a service to enhance Southwest Airline's passengers' ability to obtain a boarding pass with a high boarding priority level, had knowledge of and therefore was bound by Southwest's website Terms and Conditions of Use which prohibited third parties from accessing user accounts for commercial use, at least as of the time it was sent a cease and desist letter); *Cairo, Inc. v. Crossmedia Services, Inc.*, No. C 04-04825 JW, 2005 WL 756610 (N.D. Cal. Apr. 1, 2005) (following *Register.com* in holding that repeated use of a website with actual knowledge of the posted Terms of Use effectively binds a party to those terms); *Ticketmaster Corp. v. Tickets.com, Inc.*, CV99-7654-HLH (VBKx), 2003 WL 21406289 (C.D. Cal. Mar. 7, 2003) (finding a triable issue of fact precluding summary judgment on the issue of whether the defendant was bound by posted Terms of Use where express assent was not obtained but the defendant had been put on written notice of the conditions governing use of the internal pages of plaintiff's website and thereafter continued to access them); see generally *infra* § 21.03.

⁸See, e.g., *Specht v. Netscape Communications Corp.*, 306 F.3d 17,

Indeed, courts frequently are hostile to efforts to enforce unilateral online agreements in the absence of express assent.⁹ The absence of a binding contract could be especially problematic for database owners and website licensors whose agreements purport to restrict use of factual data (as opposed to reports, articles or other creative content entitled to copyright protection, independent of the database compilation).¹⁰

Courts increasingly use a lexicon of assorted jargon to refer to the various ways in which a contract may be formed online, including *clickwrap* and *browsewrap* agreements, and hybrids characterized by Eastern District of New York Judge Jack Weinstein as so-called *scrollwrap*¹¹ and *sign-in-*

22–24 (2d Cir. 2002) (declining to enforce an arbitration provision and finding assent lacking where users of Netscape’s website were urged to download free software by clicking on a button labeled “Download” but would not even have seen an invitation to review the license agreement available by hyperlink unless they scrolled down to the following page, where the full terms, which warned users that they should not download the software if they did not agree to be bound and included the arbitration provision, were only accessible via that link, and where the defendants alleged that they in fact were unaware that the free software was provided subject to terms); *A.V. v. iParadigms, LLC*, 544 F. Supp. 2d 473, 485 (E.D. Va. 2008) (declining to enforce an indemnification provision contained in defendant’s Usage Policy, which was accessible via a link from every page on the website, where there was no evidence to impute knowledge of the terms to the plaintiffs and where the clickwrap agreement for the site, which unlike the Usage Policy was held enforceable, did not incorporate the Policy by reference and included an integration clause that stated that the clickwrap agreement “constitutes the entire agreement . . . with respect to usage of this Website.”), *aff’d in part and rev’d in part on other grounds*, 562 F.3d 630, 639 (4th Cir. 2009); *see generally infra* § 21.03 (analyzing case law and the circumstances under which courts will enforce unilateral contracts based on implied assent).

⁹*See infra* § 21.03.

¹⁰*See supra* § 5.02.

¹¹A *scrollwrap* agreement, using Judge Weinstein’s terminology, requires a user to scroll through the terms before the user can assent to the contract by clicking on an “I agree” button. *See Berkson v. Gogo LLC*, 97 F. Supp. 3d 359, 386, 398–99 (E.D.N.Y. 2015). Judge Weinstein would put in this category cases typically categorized as *clickwrap* or express assent opinions such as *Feldman v. Google, Inc.*, 513 F. Supp. 2d 229, 236–38 (E.D. Pa. 2007) (enforcing Google’s AdWords clickwrap contract where there was reasonable notice of and mutual assent to the agreement; the contract was immediately visible in a scrollable text box below a prominent admonition in boldface to read the terms and conditions carefully and only assent if the user agreed to the terms, the terms were presented in twelve-

wrap agreements.¹² This jargon may obscure, rather than clarify the question of whether express or implied assent was obtained.

While the ever increasing number of cases evaluating various means for obtaining online assent are analyzed extensively in section 21.03, the bottom line for database owners is that it is always safer to obtain express assent to a database access agreement or EULA—and failing to do so could make it difficult to enforce contractual terms.

Where database use or access is conditioned on acceptance of a unilateral contract, there is also a risk—particularly in more liberal jurisdictions such as California and in consumer

point font and was only seven paragraphs long and was available in a printer-friendly, full-screen version; according to Judge Weinstein, “the plaintiff had the duty to read terms that were presented in a scroll box and required a click to agree and, therefore, the fact that the entire contract was not visible in the scroll box was irrelevant”); *Bar-Ayal v. Time Warner Cable Inc.*, No. 03-CV-9905, 2006 WL 2990032, at *9–10 (S.D.N.Y. Oct. 16, 2006) (finding acceptance where scrolling through thirty-eight screens of text was required—essentially the entire agreement); *Moore v. Microsoft Corp.*, 293 A.D.2d 587, 741 N.Y.S.2d 91, 92 (2d Dep’t 2002) (holding that a contract was formed when “[t]he terms of the [agreement] were prominently displayed on the program user’s computer screen before the software could be installed,” and “the program’s user was required to indicate assent to the [agreement] by clicking on the ‘I agree’ icon before proceeding with the download”); *In re RealNetworks, Inc.*, No. 00-CV-1366, 2000 WL 631341, at *6 (N.D. Ill. May 8, 2000) (approving a license agreement placed in a pop-up window with scroll bar).

¹²See *Berkson v. Gogo LLC*, 97 F. Supp. 3d 359, 392-402 (E.D.N.Y. 2015). A *sign-in-wrap* agreement notifies a user of the existence of terms of use but instead of providing an “I agree” button, advises the user that he or she is agreeing to the terms when registering or signing up for the site or service. See *id.* at 399-400. Judge Weinstein would put in this category *Fteja v. Facebook, Inc.*, 841 F. Supp. 2d 829 (S.D.N.Y. 2012), where express assent was found but the court characterized the agreement as a “hybrid.” Judge Weinstein, analyzing self-described hybrid cases which he characterized as involving so-called sign-in-wrap agreements, explained that these type of agreements have been enforced based on “notice and an effective opportunity to access terms and conditions” in cases where (1) there is a hyperlink to the Terms next to the only button that will allow a user to continue use of the website, (2) the user registered or signed up for a service “with a clickwrap agreement and was presented with hyperlinks” to the Terms; or (3) notice of hyperlinked terms “is present on multiple successive webpages of the site.” *Berkson v. Gogo LLC*, 97 F. Supp. 3d at 401.

As analyzed in section 21.03, these fine distinctions based on past district court cases are not really helpful in evaluating express or implied assent. See *infra* § 21.03.

(rather than commercial) contracts—that the contract or various provisions in the agreement could be challenged as unconscionable in the event of litigation. Readers should closely review sections 21.03, 21.04, 21.05, and 22.05[2][M] for guidance on when unilateral agreements may be held unenforceable as unconscionable.

Where an enforceable contract has been formed, a database owner may bring a breach of contract claim. In some circumstances where there is privity of contract and the defendant has thwarted the owner from benefiting from it, the database owner may bring a claim for breach of the duty of good faith and fair dealing, which under California law requires a showing that (1) the parties entered into a contract, (2) the plaintiff fulfilled its obligations under the contract, (3) any conditions precedent to the defendant's performance occurred, (4) the defendant unfairly interfered with the plaintiff's rights to receive the benefits of the contract, and (5) the plaintiff was harmed by the defendant's conduct.¹³ The implied covenant, however, is limited to assuring compliance with the express terms of a contract, and cannot be extended to create obligations not contemplated by it.¹⁴ Nor can there be a breach of the duty of good faith and fair dealing if the conduct alleged was expressly permitted by the contract.¹⁵

Where a screen scraper or third party aggregator makes available software or other tools to allow users to circumvent restrictions in the contract, a database owner, in limited circumstances, also potentially may be able to bring claims for tortious interference with contract or interference with pro-

¹³See *Herskowitz v. Apple, Inc.*, 940 F. Supp. 2d 1131, 1143 (N.D. Cal. 2013) (reciting the elements from a standard jury instruction).

¹⁴See, e.g., *Herskowitz v. Apple, Inc.*, 940 F. Supp. 2d 1131, 1143 (N.D. Cal. 2013); see generally *infra* § 14.03[2] (discussing the doctrine and its application at greater length).

¹⁵See, e.g., *Song Fi Inc. v. Google, Inc.*, 108 F. Supp. 3d 876, 885 (N.D. Cal. 2015) (granting Google's motion to dismiss claims for breach of YouTube's Terms of Service and breach of the duty of good faith and fair dealing arising out of plaintiffs' removal of a video where the Terms of Service permitted YouTube to remove the video "and eliminate its view count, likes, and comments"; "if defendants were given the right to do what they did by the express provisions of the contract there can be no breach [of the duty of good faith and fair dealing].").

spective economic advantage.¹⁶ Where contract remedies may be unavailable, database owners have sought to assert claims for unjust enrichment,¹⁷ although such claims, to the extent based on copying without an extra element, have been held preempted by the Copyright Act.¹⁸

In lieu of simple contracts, database agreements potentially may be cast as intellectual property licenses—even where the underlying data is unprotectable.¹⁹ Many databases, even if comprised of unprotectable facts or data, may be entitled to copyright protection as compilations if there is sufficient creativity in the selection, arrangement or organization of the data.²⁰ Databases also frequently incorporate software that may be protected by copyright, trade secret and/or patent law and form the basis for an intellectual property license.

Where a database agreement authorizes access to or use of intellectual property and therefore may be characterized as a license, rather than merely a contract, the agreement may be easier to enforce in some instances because courts generally allow rights owners to impose restrictions on licensees that might otherwise would be deemed impermissible in a

¹⁶*See, e.g., Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1059–60 (N.D. Cal. 2010) (entering a default judgment for breach of contract, inducing breach of contract and intentional interference with contractual relations, where the defendants marketed a software product that allowed users to automate access to the Craigslist site, circumvent CAPTCHA restrictions and automatically and repeatedly post identical listings on Craigslist, and harvest email addresses, all in violation of Craigslist's TOU); *infra* § 5.03[5].

¹⁷*See, e.g., Information Handling Services, Inc. v. LRP Publications, Inc.*, No. CIV. A. 00-1859, Copy. L. Rep. (CCH) P28,177, 2000 WL 1468535 (E.D. Pa. Sept. 20, 2000) (denying a motion to dismiss a claim for unjust enrichment in a database copying case); *see generally infra* § 5.03[6].

¹⁸*See, e.g., BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 618 (S.D.N.Y. 2010); *Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 680–81 (N.D. Ohio 2010) (holding plaintiff's unjust enrichment claim preempted where it was based on the allegation that defendants took information). *But see Perfect 10, Inc. v. Google, Inc.*, No. CV04-9484, 2008 WL 4217837, at *9 (C.D. Cal. July 6, 2008) (holding plaintiff's unjust enrichment claim not preempted where the claim was premised on right of publicity and trademark violations); *see generally infra* § 5.03[6] (unjust enrichment); *supra* § 4.18[1] (analyzing copyright preemption).

¹⁹For a discussion of intellectual property licenses, *see infra* chapter 16.

²⁰*See supra* § 5.02.

regular contract (such as prohibitions on competition or reverse engineering²¹) as a condition of gaining access to intellectual property,²² so long as the restrictions are not so severe that they amount to intellectual property misuse²³ or violate antitrust laws.²⁴

Where a database is entitled to copyright protection, a rights owner may have remedies against a licensee under both the Copyright Act (which potentially allows recovery of statutory damages and attorneys' fees²⁵) and for breach of contract—or potentially only for breach of contract. Exceeding the scope of a valid license may be found to constitute infringement.²⁶ Likewise, violating a condition of the license may be deemed copyright infringement, rather than merely a breach of contract.²⁷ Other breaches of a license agreement, however, may be deemed merely contractual, and would not afford independent grounds for a database owner to sue for copyright infringement in federal court or seek

²¹See *infra* § 18.03[6].

²²See, e.g., *Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1113 (C.D. Cal. 2007) (enforcing website Terms of Use for access to a database as a copyright license); see generally *infra* § 21.03 & chapters 14, 16.

²³See *infra* § 16.04.

²⁴See *infra* chapter 34.

²⁵See *supra* §§ 4.14, 4.15. As discussed in section 4.14, statutory damages, if available, allow a plaintiff to recover up to \$150,000 per work infringed where willful infringement may be shown and typically are sought where actual damages would be negligible or would be difficult or expensive to prove.

²⁶See, e.g., *I.A.E., Inc. v. Shaver*, 74 F.3d 768, 775 (7th Cir. 1996); *MAI Systems Corp. v. Peak Computer, Inc.*, 991 F.2d 511, 517 (9th Cir. 1993), *cert. dismissed*, 510 U.S. 1033 (1994); see generally *supra* § 4.08[5]; *infra* § 14.06[2].

²⁷See *MDY Industries, LLC v. Blizzard Entertainment, Inc.*, 629 F.3d 928, 939-41 (9th Cir. 2010) (holding that Terms of Use restrictions on the use of bots—or intelligent agent software—as a form of “cheating” to acquire virtual goods in *World of Warcraft*, was a contractual covenant, rather than a condition of the license, and therefore could not form the basis for a claim for copyright infringement when breached). For a licensee's violation of a contract to constitute copyright infringement, according to the Ninth Circuit, there must be a nexus between the condition and the licensor's exclusive rights of copyright. Otherwise, “any software copyright holder . . . could designate any disfavored conduct during software use as copyright infringement by purporting to condition the license on the player's abstention from the disfavored conduct.” *Id.* at 941; see generally *infra* § 16.02[1] (analyzing this issue in greater detail).

statutory damages or attorneys' fees. Whether and to what extent a contractual restriction may be considered a condition of the license, rather than merely contractual, is analyzed in section 14.06[2].

Even where an agreement is merely a contract for access to data, some courts will enforce use restrictions on the theory that the database owner was not required to grant access to the database in the first place, and that the licensee knowingly gave up certain rights that it otherwise may have had with respect to its use of unprotectable data, in return for obtaining the right to access and use the database.²⁸ Some judges, however, will strain to find copying permissible when undertaken for the purpose of accessing unprotectable data,²⁹ so in practice, even if not necessarily as a matter of black letter law, database owners are better off including restrictions in IP licenses, rather than mere access contracts, and obtaining express assent, where possible.

The major cases involving database contracts and licenses are analyzed below in section 5.03[2]. Readers are cautioned, however, to closely review chapters 21 (especially sections 21.03, 21.04 and 21.05) and 22 on unilateral contracts, which provide a broader perspective on the enforceability of database EULAs and Terms of Use.

Whether an agreement fully protects a database owner, or leaves opportunities for a third party to copy or use the contents of the database, also depends on the particular use restrictions imposed by the agreement. Use restrictions are addressed in section 5.03[2], as well as in *Ticketmaster LLC v. RMG Technologies, Inc.*,³⁰ which is discussed in section 5.03[2].

²⁸See, e.g., *Information Handling Services, Inc. v. LRP Publications, Inc.*, No. CIV. A. 00-1859, Copy. L. Rep. (CCH) P28,177, 2000 WL 1468535 (E.D. Pa. Sept. 20, 2000).

²⁹See *supra* § 5.02.

³⁰*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007).

5.03[2] Database Contract Case Law

The first case to construe a database contract involving Internet use was *ProCD, Inc. v. Zeidenberg*,¹ which was decided in 1996. In that case, the Seventh Circuit upheld the enforceability of a non-negotiated, pre-printed shrinkwrap license, which had been included with a CD-ROM containing a database of unprotectable information compiled from telephone directories.²

Judge Easterbrook assumed, for purposes of the case, that ProCD's database, although more complex and voluminous than a regular telephone directory (which included, for example, full nine-digit zip codes and census industrial codes), nonetheless was unprotectable under *Feist Publications, Inc. v. Rural Telephone Service Co.*³ Notwithstanding this ruling, the Seventh Circuit held that ProCD's purely factual database could be effectively protected by a shrink-wrap license.

The legal authority for this aspect of the court's ruling, however, was not clearly articulated. Judge Easterbrook cited trade secret cases such as *Kewanee Oil Co. v. Bicron*,⁴ and *Aronson v. Quick Point Pencil Co.*⁵ for the proposition that ProCD's shrinkwrap license was enforceable to limit defendants' use of ProCD's otherwise unprotectable database.

[Section 5.03[2]]

¹*ProCD, Inc. v. Zeidenberg*, 86 F.3d 1447 (7th Cir. 1996).

²The CD-ROM also included a protectable software program that allowed users to search the database, but the Seventh Circuit's opinion focused on the defendant's reproduction of the database on a website. The facts of the case are set forth in greater detail in connection with the enforceability of unilateral licenses. *See infra* § 21.02, § 21.03.

³*Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991).

⁴*Kewanee Oil Co. v. Bicron Corp.*, 416 U.S. 470 (1974).

⁵*Aronson v. Quick Point Pencil Co.*, 440 U.S. 257 (1979). *Aronson* involved a license to a novel invention, which provided for different license fees depending on whether the licensor was able to obtain a patent. When a patent could not be obtained, the licensee sought a declaration that the license agreement was unenforceable. In upholding the agreement under the lower license fee (since the licensor was unsuccessful in obtaining a patent), the Supreme Court emphasized that the invention was secret before the licensee commercially exploited it, which allowed the licensee to profit from being first in the market. *Aronson v. Quick Point Pencil Co.*, 440 U.S. 257, 265–66 (1979), *citing Kewanee Oil Co. v. Bicron* 416 U.S. 470 (1974); Restatement of Torts § 757, Comment b. Thus, *Aronson* is a case firmly grounded in trade secret law.

Moreover, in connection with his discussion of the Aronson case, Judge Easterbrook referred to directories collected for inclusion in a hypothetical database as “intellectual property.” Yet in the *Zeidenberg* case itself there is no suggestion that plaintiff’s database constituted a trade secret, or that it was otherwise protectable. In fact, to the contrary, the district court, in analyzing plaintiff’s cause of action for misappropriation, emphasized that plaintiff’s claim was based on common law misappropriation, not misappropriation of trade secrets.⁶

Courts analyzing database licenses since *Zeidenberg* typically have either enforced end user license agreements as valid contracts or licenses or raised concerns about the implications of restricting access to publicly available information. Databases often include software or other intellectual property that must be used to access the data in a database, thus justifying imposing restrictions on the use of otherwise unprotectable data. Most courts seem comfortable with the notion that a licensee may be bound by restrictions on otherwise unprotectable data if its access to the data was provided pursuant to an agreement that granted it rights that the licensor otherwise could freely withhold (especially if the agreement granted access to intellectual property—such as software—or a database protectable based on the selection, arrangement or organization of its contents). On the other hand, attempts to expand the monopoly power granted a copyright owner to unprotectable material could prevent a licensor from enforcing its rights under the copyright misuse doctrine.⁷

Judge Easterbrook’s own opinion in *ProCD, Inc. v. Zeiden-*

A broad reading of *Aronson* (and perhaps the one implicitly intended by Judge Easterbrook) would also support the proposition that a licensee who accepts something of value that a licensor is not required to provide is thereafter bound by the terms of its license agreement, which is consistent with the rationale provided by some courts (in cases cited in this section and in section 18.06) in enforcing contracts that restrict use of otherwise unprotectable data.

⁶*ProCD, Inc. v. Zeidenberg*, 908 F. Supp. 640, 660 (W.D. Wis. 1996), *rev’d*, 86 F.3d 1447 (7th Cir. 1996). Plaintiff’s database was comprised of public phone book entries and therefore was not secret. The software developed to manage the data arguably could have constituted or embodied a trade secret. See *infra* § 10.03.

⁷See, e.g., *Lasercomb America, Inc. v. Reynolds*, 911 F.2d 970, 978 (4th Cir. 1990); *DSC Communications Corp. v. DGI Technologies, Inc.*, 81 F.3d 597, 601 (5th Cir. 1996); *Practice Management Information Corp. v.*

berg underscores the tension between enforcing contract rights and protecting access to unprotected data. For example, under Judge Easterbrook's analysis, Borland, which copied the unprotectable menu command hierarchy of Lotus' Lotus 1-2-3 program and was exonerated by an equally divided Supreme Court in *Lotus Dev. Corp. v. Borland Int'l, Inc.*,⁸ could have been subjected to liability if, instead of proceeding under copyright law, Lotus had sued Borland for violation of a shrinkwrap license that prohibited copying except for personal use. Moreover, in the *Zeidenberg* case itself, although ProCD sought to enjoin defendants' copying of the telephone listings contained on its CD-ROM, ProCD—ironically—would not have been able to compile its CD-ROM listings from over 3,000 telephone books, royalty free, had it not, like *Zeidenberg*, relied on the Supreme Court's ruling in *Feist* that phone book compilations are unprotectable under U.S. copyright law. In holding that access to unprotectable data may be restricted by a shrinkwrap license, the Seventh Circuit seems to have condoned ProCD's copying of unprotectable data from a phone book, while penalizing Zeidenberg for essentially the same conduct. It is doubtful that this was the intended consequence of the Seventh Circuit's otherwise well-reasoned opinion.

In *Hill v. Gateway 2000, Inc.*,⁹ Judge Easterbrook suggested that the *Zeidenberg* decision rested on the UCC and expressly rejected the notion that the sweeping restrictions enforced in *Zeidenberg* were in any way justified by license (and hence underlying intellectual property), rather than merely by contract.

In *Register.com, Inc. v. Verio, Inc.*,¹⁰ the Second Circuit held that Verio was bound by Register.com's posted Terms, even though express assent was neither sought nor obtained, because Verio acknowledged that it was aware that *Register.com* purported to condition use of its site on posted Terms. In that case, *Register.com*, a domain name registrar,

American Medical Ass'n, 133 F.3d 1140 (9th Cir. 1998), *cert. denied*, 522 U.S. 933 (1998). For an analysis of intellectual property misuse doctrines, see *infra* § 16.04.

⁸*Lotus Development Corp. v. Borland Int'l, Inc.*, 49 F.3d 807 (1st Cir. 1995), *aff'd mem.*, 516 U.S. 233 (1996) (4-4 decision); see generally *supra* § 4.07.

⁹*Hill v. Gateway 2000, Inc.*, 105 F.3d 1147 (7th Cir. 1997).

¹⁰*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393 (2d Cir. 2004).

was contractually required to make the contact information of domain name registrants from the WHOIS database available free of charge to the public for any lawful purpose. When the database was queried, however, *Register.com* displayed a purported restriction on use in the results screen. Specifically, users were shown a restrictive legend purporting to prohibit recipients from using the data to transmit “mass unsolicited, commercial advertising or solicitation via email” (or in connection with mail or telephone solicitations).¹¹

Verio used bots (or intelligent agent software) to access the site and copy the contact information of new registrants, who it then solicited via email, telemarketing and direct mail marketing solicitations.

Verio acknowledged that it was aware of the restrictions that *Register.com* purported to impose on users, but argued that it was not bound by them because the legend did not appear on the screen until after Verio had queried the database and received the desired information. Judge Leval, writing for a majority of the panel, however, found Verio bound by the terms, writing that:

It is standard contract doctrine that when a benefit is offered subject to stated conditions, and the offeree makes a decision to take the benefit with knowledge of the terms of the offer, the taking constitutes an acceptance of the terms, which accordingly become binding on the offeree.¹²

In *Information Handling Servs. v. LRP Publications, Inc.*,¹³ the court stated in *dicta* in ruling on a motion to dismiss

¹¹*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 396 (2d Cir. 2004). Register.com initially prohibited solicitation by email, mail or telephone. Its agreement with ICANN, however, prohibited it from restricting access to the data for any lawful purpose except mass unsolicited email. See *infra* § 7.02. Register.com therefore narrowed its policy to just prohibiting email solicitations.

With respect to the earlier policy, the Second Circuit rejected Verio’s argument that it could not be held liable for telephone and mail solicitations given the terms of Register.com’s agreement with ICANN because that agreement provided that there were no third-party beneficiaries. The Second Circuit therefore analyzed Verio’s potential liability on the assumption that Register.com was legally authorized to demand that users of WHOIS data from its system refrain from using it for mass solicitation by mail and telephone, as well as by email.

¹²*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 403 (2d Cir. 2004) (citing multiple sources). The *Verio* case is discussed in greater detail in section 21.03.

¹³*Information Handling Services, Inc. v. LRP Publications, Inc.*, No.

that a shrinkwrap license restricting use of a CD-ROM and its contents was enforceable in preventing the defendant from copying material that was not protected by copyright law. The specific issue in that case was whether plaintiffs' claims against a competing database company were preempted by the Copyright Act. The plaintiffs alleged that the defendant, through an employee, subscribed to plaintiffs' PERSONNET database (a database that included material such as decisions by the Equal Employment Opportunity Commission (EEOC)) by falsely representing herself as an attorney who intended to use the product in her practice. In fact, the employee allegedly intended all along to use her access to the database to engage in wholesale copying, which is what in fact she did. The court dismissed a claim for "misappropriation and unfair competition," finding that it alleged copying without an extra element and was therefore preempted.¹⁴ The court, however, denied defendant's motion with respect to claims for breach of contract, tortious interference with contractual relations, conspiracy, tortious interference with prospective contractual relations, unjust enrichment, fraud and unfair competition/misappropriation of trade secrets.

In addressing plaintiff's breach of contract claim, Judge Fullam wrote that while copyright law does not allow a database owner to prevent third parties from copying unprotectable material simply because it took time and effort to create it, "there is no law that requires [a database owner] to make [its] product publicly available; nor is it permissible to break into [a] house and steal it in order to copy the material it contains." He further observed:

I am not unmindful of the concern that enforcing licenses such as the one involved here, which are not-bargained-for and are offered with the product on a take-it-or-leave-it basis, may be the functional equivalent of expanding, under the rubric of various state laws, copyright protection to otherwise uncopyrightable materials. *See* 1 Melville B. Nimmer & David Nimmer, Nimmer on Copyright, § 3.04[B][3][a]. In the absence of congressional guidance, however, I can see no reason not to enforce the contract under the circumstances presented in this case. It is not unconscionable. Defendant . . . was free to reject it and return the CD-ROM disc to [plaintiffs]. Defendant chose not to do so, and therefore is bound by its terms."

CIV. A. 00-1859, Copy. L. Rep. (CCH) P28,177, 2000 WL 1468535 (E.D. Pa. Sept. 20, 2000).

¹⁴*See infra* § 5.04.

In unreported but widely discussed opinions in *Ticketmaster Corp. v. Tickets.com, Inc.*,¹⁵ a district court case that was the first one to consider the effectiveness of posted Terms of Use, the court initially dismissed Ticketmaster's breach of contract claim against *Tickets.com* where Ticketmaster's TOU were accessible via a link that appeared in "small print" on the bottom of Ticketmaster's home page, but granted leave for Ticketmaster to amend its complaint to allege that *Tickets.com* had knowledge of the terms and impliedly agreed to them.¹⁶ Thereafter, Ticketmaster changed the placement of the link to its notice to a prominent place on its homepage and warned users that proceeding beyond the homepage would be deemed agreement to Terms of Use that prohibited commercial use of the site.¹⁷ Ticketmaster also reiterated the conditions imposed on access to its site in a letter to *Tickets.com*.

In a subsequent decision,¹⁸ the court denied the defendant's motion for summary judgment on Ticketmaster's breach of contract claim, finding that a contract could have been formed when *Tickets.com* proceeded into the interior of the Ticketmaster site after knowing of the conditions imposed by Ticketmaster for doing so. In so ruling, the court emphasized that *Tickets.com* was "fully familiar with the conditions [Ticketmaster] claimed to impose on users," citing in particular Ticketmaster's letter and a response from *Tickets.com* stating that it did not accept the conditions, as well as the new and more prominent notice placed on Ticketmaster's homepage. The court ruled that there was sufficient evidence to defeat summary judgment "if knowledge of the asserted conditions of use was had by [Tickets.com], who nev-

¹⁵*Ticketmaster Corp. v. Tickets.com, Inc.*, CV 99-7654 HLH (BQRx), 2000 WL 525390 (C.D. Cal. Mar. 27, 2000), *aff'd mem.*, Appeal No. 00-56574 (9th Cir. Jan. 2001).

¹⁶Ticketmaster had sought to prevent *Tickets.com*, a competitor, from deep linking to internal pages on its website and from using bots to spider or crawl pages on its site and electronically extract factual information from the Ticketmaster site. See *infra* § 9.06 (discussing the case at greater length in connection with linking).

¹⁷The notice read:

Use of this site is subject to express terms of use, which prohibit commercial use of this site. By continuing past this page, you agree to abide by these terms.

¹⁸*Ticketmaster Corp. v. Tickets.com, Inc.*, CV99-7654-HLH (VBKx), 2003 WL 21406289 (C.D. Cal. Mar. 7, 2003).

ertheless continued to send its spider into the [Ticketmaster] interior Web pages, and if it is legally concluded that doing so can lead to a binding contract.”

The Ticketmaster court lamented the result of its ruling, expressing a preference for “a rule that required an unmistakable assent to the conditions easily provided by requiring clicking on an icon which says ‘I agree’ or the equivalent.” It acknowledged, however, that “the law has not developed in this way” and that “no particular form of words is necessary to indicate assent—the offeror may specify that a certain action in connection with his offer is deemed acceptance, and [the offer will] ripe[n] into a contract when the action is taken.”¹⁹

In a later case also brought by Ticketmaster, *Ticketmaster LLC v. RMG Technologies, Inc.*,²⁰ Ticketmaster was able to enforce its Terms against a user who acknowledged it was aware of the restrictions and the court treated the posted Terms as an intellectual property license. The case is instructive both for contract formation and the terms of the database license that the court enforced.

In *RMG Technologies, Inc.*, Judge Audrey Collins, ruling on a motion for preliminary injunction, ruled that Ticketmaster was likely to prevail on the merits in establishing that a competitor had notice of posted Terms of Use but nonetheless accessed and used the Ticketmaster website in violation of those Terms. The court characterized Ticketmaster’s Terms as creating a non-exclusive copyright license to Ticketmaster’s copyrighted website. In addition, the homepage to the site included the warning that “[u]se of this website is subject to express Terms of Use which prohibit commercial use of this site. By continuing past this page, you agree to abide by these terms.” The underlined phrase “Terms of Use” was a hyperlink to the full Terms of Use. In addition, the same phrase appeared on almost every page of

¹⁹*Ticketmaster Corp. v. Tickets.com, Inc.*, CV99-7654-HLH (VBKx), 2003 WL 21406289, at *2 (C.D. Cal. Mar. 7, 2003) (citing other cases). In addition to Internet contract cases, the court cited the shrinkwrap cases (see *infra* § 21.02), *Carnival Cruise Lines, Inc. v. Shute*, 499 U.S. 585 (1991) (see *infra* §§ 21.02, 21.03, § 53.03[2]), and the fact that “[t]he Carriage of Goods by Sea Act, the Carmack Act, and the Warsaw Convention provide that limitations of liability on the bill of lading, air waybill, or airplane ticket are enforceable if the services are used by the customer.”

²⁰*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007).

the Ticketmaster site. Further, since 2003 users had to affirmatively agree to the Terms as part of the procedure for setting up an account and since mid-2006 had to expressly assent to the Terms any time they purchased tickets from the site. The defendant acknowledged that it had notice of the Terms of Use but argued that it was not bound by them and they were too vague to be enforced, which the court rejected. Because the defendant acknowledged that it was on notice of the Terms, the court found that it had assented to be bound by the Terms by using the website.

Ticketmaster's Terms of Use included a number of provisions expressly intended to thwart competitors from accessing and copying its database. The Terms, among other things prohibited commercial use of the Ticketmaster website, including duplication or downloading of material, and purported to license only personal, non-commercial use.²¹ The Terms also prohibited the use of bots or other automated devices.²² The agreement also included an undertaking by the user "not [to] take any action that imposes an unreasonable or disproportionately large load on our infrastructure."²³ To further limit automated copying, the Terms included an undertaking not to access the site more than once during any three second interval.²⁴ For good measure, the Terms made clear that users "do not have permission to access this

²¹The relevant provisions included the following:

You [the viewer] agree that you are only authorized to visit, view and to retain a copy of pages of this site for your own personal use, and that you shall not duplicate, download, [or] modify . . . the material on this site for any purpose other than to review event and promotions information, for personal use No . . . areas of this site may be used by our visitors for any commercial purposes.

²²The Terms stated:

You agree that you will not use any robot, spider or other automated device, process, or means to access the site You agree that you will not use any device, software or routine that interferes with the proper working of the site nor shall you attempt to interfere with the proper working of the site.

²³*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1107 (C.D. Cal. 2007).

²⁴The Terms provided:

You agree that you will not access, reload or "refresh" transactional event or ticketing pages, or make any other request to transactional servers, more than once during any three-second interval.

site in any way that violates . . . these terms of use.”²⁵ The Terms also provided that users understood and agreed that Ticketmaster could terminate their access to the site or cancel their ticket orders or the actual tickets acquired through the site if Ticketmaster believed that the conduct of a user or anyone who Ticketmaster believed was acting in concert with the user violated or was inconsistent with the Terms of Use or the law or violated the rights of Ticketmaster, a client of Ticketmaster or another user of the site.

In *RMG Technologies, Inc.*, the defendant used bots or another automated means to access the Ticketmaster site and extract information from its database. Although the defendant denied using automated means, Ticketmaster’s expert showed that several webpage requests per second were made to Ticketmaster from the same IP address, amounting to thousands of requests per day that were too numerous to have been generated in a manual, non-automated way. In addition, the defendant advertised a product called Purchase-Master as “do[ing] the work of a dozen people at once.” Accordingly, the court found Ticketmaster likely to prevail based on the defendant’s use of an automated device in violation of the Terms of Use, as well as the provisions restricting access to once every three seconds. Because the court deemed the Terms of Service a license to Ticketmaster’s copyrighted site, the defendant’s access beyond what was permitted by the license also was potentially a copyright violation.

RMG Technologies, Inc., was followed by the court in *Facebook, Inc. v. Power Ventures, Inc.*,²⁶ in which Judge Jeremy Fogel of the Northern District of California denied defendants’ motion to dismiss copyright and DMCA claims arising out of their screen scraping user data from Facebook’s website. In that case, defendants operated a website that allowed users to access their accounts on various different email services and social networks from a single location using screen-scraping tools.

The court ruled that Facebook stated claims based on defendants exceeding the scope of permissible access, as

²⁵*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1108 (C.D. Cal. 2007).

²⁶*Facebook, Inc. v. Power Ventures, Inc.*, 91 U.S.P.Q.2d 1430, 2009 WL 1299698 (N.D. Cal. May 11, 2009).

defined by Facebook's Terms of Use agreement.²⁷ Among other things, Facebook's Terms of Use agreement granted a limited license to users to access the site and service, but only on the terms specifically authorized in the TOU agreement.²⁸ The agreement also prohibited harvesting or collecting email addresses or other contact information by electronic or other means for the purpose of sending unsolicited emails or other unsolicited communications. Facebook's Terms of Use agreement further broadly prohibited the downloading, scraping, or distributing of any content on the website (except that users were permitted to download their own content). It also prohibited "data mining, robots, scraping, or similar data gathering or extraction methods" (with no exception to this prohibition for user access).

Database access restrictions in Craigslist's TOU also were enforced against defendants who accessed the site in excess of TOU restrictions to harvest email addresses and develop and market software and related services to allow users to automate the process of posting listings on Craigslist.²⁹ In that case, the court granted a default judgment on claims for breach of the TOU as well as for inducing breach of contract and intentional interference with contractual relations (based on breaches of the TOU by users of defendants' automated software programs).

In *Health Grades, Inc. v. Robert Wood Johnson University Hospital, Inc.*,³⁰ the court held that plaintiff's claim that the defendant hospital breached its click-through license agreement with the plaintiff by commercially reproducing, modifying and/or distributing its healthcare provider award and ranking information from plaintiff's website in press releases

²⁷See *infra* § 5.08 (discussing Facebook's Lanham Act claim).

²⁸The relevant provision stated that "[a]ny use of the site or the site content other than as specifically authorized herein, without the prior permission of the company, is strictly prohibited and will terminate the license granted herein."

²⁹See *Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039 (N.D. Cal. 2010) (entering a default judgment for copyright infringement, trademark infringement based on use on plaintiff's marks in sponsored links, circumvention (of CAPTCHA) under the DMCA, exceeding authorized access under the CFAA (based on exceeding permitted access under Craigslist's TOU), breach of contract (based on the TOU), fraud and violations of Cal. Penal Code § 502).

³⁰*Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1242–47 (D. Colo. 2009).

and other marketing materials was preempted, but that its breach of contract claim based on unauthorized use of the plaintiff's mark in a way that implied that the owner endorsed the hospital's services was not preempted and could proceed. In that case, the defendant had accessed and clicked assent to plaintiff's limited license more than 200 times.

Similarly, in *Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*,³¹ the court denied in part defendant's summary judgment motion, holding that a reasonable jury could conclude that the defendant had actual or constructive knowledge of the Snap-on EULA where defendant accessed Snap-on's websites, which contained a single page access screen where users were required to input their user names and passwords and then click an "Enter" button to proceed, below which was found the message that "[t]he use of and access to the information on this site is subject to the terms and conditions set forth in our legal statement" and a green box with an arrow that users could click to access the EULA. The EULA restricted access to the site to authorized dealers, customers or other licensees, and there was no dispute that O'Neil, a competitor which entered the site, was none of these, although the parties disputed whether O'Neil was an authorized agent for Mitsubishi, a customer of both O'Neil and Snap-on, and whether Mitsubishi was authorized to access the site.³²

Online and mobile contract formation is addressed more extensively in section 21.03.

5.03[3] The Scope of Contractual Restrictions

Database agreements typically restrict access to and use of a database but may not necessarily prohibit all uses of data. Some broadly restrict commercial use of a database,

³¹*Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 681–83 (N.D. Ohio 2010).

³²Snap-On eventually obtained a general jury verdict at trial, although it is not clear whether the verdict was based on Snap-On's claim for breach of the EULA or other claims for trespass, copyright infringement or violations of the Computer Fraud and Abuse Act. See *Snap-On Business Solutions Inc. v. O'Neil & Associates, Inc.*, No. 5:09–CV–1547, 2010 WL 2650875 (N.D. Ohio July 2, 2010) (awarding costs but denying Snap-On's request for an award of attorneys' fees because under Ohio law contractual attorneys' fee provisions are unenforceable as contrary to public policy because they are viewed as encouraging litigation); see generally *infra* § 5.05[1] (discussing the facts of the case and other rulings in greater detail).

such as the agreements at issue in *Ticketmaster LLC v. RMG Technologies, Inc.*¹ and *Facebook, Inc. v. Power Ventures, Inc.*,² which were discussed above in section 5.03[2]. The access and use restrictions in those agreements, which are discussed in section 5.03[2], should be closely reviewed.

Other database agreements, however, do not broadly restrict commercial use (often times because the agreement by its nature is intended to provide access for business uses). Some agreements include confidentiality provisions with carve outs for information that is in the public domain, which may include factual material in a database. This type of provision is especially common where mutual confidentiality obligations are imposed or in unilateral agreements where drafters may be concerned about avoiding one-sided provisions that appear unconscionable. Other confidentiality provisions may be more akin to trade secret licenses, which tightly restrict information which, while potentially unprotectable under copyright law, may have great value so long as kept secret.³

To determine whether and to what extent the contents of a database may be used, lawyers should, among other terms, review:

- The scope of the grant clause (to determine exactly what aspects of the database are deemed subject to the license and what restrictions, if any, are imposed in the grant clause itself);
- The Term and Termination provisions (which may specify rights following termination or require return or destruction of all data);
- Confidentiality provisions, which may exclude certain categories of information, broadly include even factual data or explicitly include only certain information (thereby impliedly allowing the use of other material in the database, unless restricted by another provision in the agreement); and

[Section 5.03[3]]

¹*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007).

²*Facebook, Inc. v. Power Ventures, Inc.*, 91 U.S.P.Q.2d 1430, 2009 WL 1299698 (N.D. Cal. May 11, 2009).

³See generally *infra* § 10.04[3] (analyzing NDAs in connection with trade secret protection).

- Use restrictions, which may purport to restrict or permit certain uses.

What a contract provides ultimately may present a question of fact precluding summary judgment—and requiring a trial to determine—if the terms are not adequately defined in the agreement itself.⁴

5.03[4] Forms

A sample database EULA is included in the appendix to chapter 21. Sample Terms of Service agreements governing access and use of websites are included in the appendix to chapter 22.

5.03[5] Interference with Contract or Prospective Economic Advantage

Where a third party makes available software or other tools to allow users to circumvent restrictions in a database agreement or website Terms of Use, the database owner potentially may bring claims for tortious interference with contract or interference with prospective economic advantage.¹

A database owner may sue a third party for breach of its database or website access or use agreement, but generally

⁴*See, e.g., Meridian Project Systems, Inc. v. Hardin Const. Co., LLC*, 426 F. Supp. 2d 1101, 1109–10 (E.D. Cal. 2006) (denying summary judgment on the issue of whether the defendant breached a license that prohibited copying “software or documentation” where it was undisputed that the defendant copied the “help” file, but where the court found that text and instructions in the Help file could constitute either, neither or both software and documentation).

[Section 5.03[5]]

¹*See, e.g., Craigslist Inc. v. Kerbel*, No. C-11-3309, 2012 WL 3166798 (N.D. Cal. Aug. 2, 2012) (entering a default judgment for breach of contract and inducing breach of contract, in addition to violations of the DMCA and the Lanham Act, where the defendant sold a service designed to automatically post to craigslist and circumvent its CAPTCHA restrictions, in violation of Craigslist’s TOU); *Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1059–60 (N.D. Cal. 2010) (entering a default judgment for breach of contract, inducing breach of contract and intentional interference with contractual relations, where the defendants marketed a software product that allowed users to automate access to the Craigslist site, circumvent CAPTCHA restrictions and automatically and repeatedly post identical listings on Craigslist, and harvest email addresses, all in violation of Craigslist’s TOU).

only if there is privity of contract. However, if a screen scraper or aggregator is not itself engaged in these practices, and merely makes available the tools for others to use, the database owner potentially may sue the third party directly based on interference claims, provided the restrictions being circumvented are part of an enforceable contract² and the third party has knowledge of its existence.

While the elements of these claims may vary somewhat from state to state, they typically require a showing of a contract, knowledge, interference and damage (or in the case of prospective economic advantage, merely harm to potential business relationships, rather than interference with an existing contractual relationship). For example, to prevail in California on a claim for intentional inducement to breach a contract, a plaintiff must prove: (1) the existence of a valid contract between the plaintiff and a third party; (2) the defendant's knowledge of this contract; (3) intentional acts designed to induce a breach or disruption of the contractual relationship; (4) actual breach or disruption of the relationship; and (5) resulting damage.³ Under California law, a defendant's conduct need not be wrongful apart from the interference with contract⁴ (although, by contrast, a plaintiff must show wrongfulness to state a claim for interference

²See *supra* §§ 5.03[1], 5.03[2]; see generally *infra* §§ 21.03, 21.04 (analyzing the enforceability of unilateral contracts, click-to-accept agreements and posted Terms).

³See *Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1059 (N.D. Cal. 2010), citing *Quelimane Co. v. Stewart Title Guaranty Co.*, 19 Cal. 4th 26, 55, 77 Cal. Rptr. 2d 709, 960 P.2d 513 (1998) and *Metal Lite, Inc. v. Brady Const. Innovations, Inc.*, 558 F. Supp. 2d 1084, 1094 (C.D. Cal. 2007); *Little v. Amber Hotel Co.*, 202 Cal. App. 4th 280, 291 (Cal. App. 2011).

⁴*Quelimane Co. v. Stewart Title Guaranty Co.*, 19 Cal. 4th 26, 77 Cal. Rptr. 2d 709, 726, 960 P.2d 513 (1998). To show the requisite level of intent, it is not necessary that plaintiffs prove that the primary purpose of defendant's conduct was the interference with contract. *Quelimane*, 77 Cal. Rptr. 2d at 727. Rather, it is sufficient to show that the defendant knew that the interference was certain or substantially certain to occur as the result of his action. *Quelimane*, 77 Cal. Rptr. 2d at 727; see also *Davis v. Nadrich*, 174 Cal. App. 4th 1, 10, 94 Cal. Rptr. 3d 414 (2d Dist. 2009) ("The rule applies . . . to an interference that is incidental to the actor's independent purpose and desire but known to him to be a necessary consequence of his action."; quoting *Korea Supply Co. v. Lockheed Martin Corp.*, 29 Cal. 4th 1134, 1155–56, 131 Cal. Rptr. 2d 29 (2003)).

with prospective economic advantage).⁵ California also recognizes an analogous cause of action for tortious interference with contract, which is similar to intentional inducement, but does not require a showing that the contract actually was breached (disruption of the contractual relationship is sufficient).⁶

To state a claim for interference with prospective economic advantage under California law, a plaintiff must show interference with *existing* noncontractual relations which hold the promise of future economic advantage.⁷ To prove interference with prospective economic advantage, a plaintiff must establish: (1) an economic relationship between the plaintiff and some third party, with the probability of future economic benefit to the plaintiff; (2) the defendant's knowledge of the relationship; (3) intentional acts on the part of the defendant designed to disrupt the relationship; (4) actual disruption of the relationship; and (5) economic harm to the plaintiff proximately caused by the acts of the defendant.⁸ Unlike other contract-based torts, under California law a plaintiff also must show that "the defendant engaged in conduct that was wrongful by some legal measure, independent of its impact on the prospective relationship." "A plaintiff must also show that the defendant's conduct was independently unlawful, that is, 'proscribed by some constitutional, statutory, regulatory, common law, or other determinable legal standard.'"⁹

Under Ohio law, tortious interference with a business re-

⁵See, e.g., *Korea Supply Co. v. Lockheed Martin Corp.*, 29 Cal. 4th 1134, 131 Cal. Rptr. 2d 29 (2003).

⁶See *Winchester Mystery House, LLC v. Global Asylum, Inc.*, 210 Cal. App. 4th 579 (2012).

⁷See, e.g., *KEMA, Inc. v. Koperwhats*, 658 F. Supp. 2d 1022, 1034 (N.D. Cal. 2009), citing *Westside Center Associates v. Safeway Stores 23, Inc.*, 42 Cal. App. 4th 507, 524, 528, 49 Cal. Rptr. 2d 793 (5th Dist. 1996) (holding that the plaintiff failed to state a claim based on interference with "with the entire market of all possible but yet unidentified buyers for its property" because the tort "protects the expectation that the relationship will eventually yield the desired benefit, not necessarily the more speculative expectation that a potentially beneficial relationship will eventually arise.").

⁸*Korea Supply Co. v. Lockheed Martin Corp.*, 29 Cal. 4th 1134, 1153 (2003).

⁹*Little v. Amber Hotel Co.*, 202 Cal. App. 4th 280, 292 n.7 (Cal. App. 2011); see also *Winchester Mystery House, LLC v. Global Asylum, Inc.*, 210 Cal. App. 4th 579 (2012); *Korea Supply Co. v. Lockheed Martin Corp.*, 29

lationship may be established by evidence of (1) a business relationship, (2) the tortfeasor's knowledge of the relationship, (3) an intentional interference causing a breach or termination of the relationship, and (4) resulting damages.¹⁰ Similarly, under Washington law, a party claiming tortious interference with a contractual relationship or business expectancy must prove: (1) the existence of a valid contractual relationship or business expectancy; (2) that defendants had knowledge of that relationship; (3) an intentional interference inducing or causing a breach or termination of the relationship or expectancy; (4) that defendants interfered for an improper purpose or used improper means; and (5) resultant damage.¹¹

Under Tennessee law, the test is stated somewhat more strictly as requiring: (1) an existing business relationship with specific third parties or a prospective relationship with an identifiable class of third persons; (2) the defendant's knowledge of that relationship and not a mere awareness of the plaintiff's business dealings with others in general; (3) the defendant's intent to cause the breach or termination of the business relationship; (4) the defendant's improper motive or improper means; and finally, (5) damages resulting from the tortious interference.¹²

In some states, including Florida, only strangers to a contract or business relationship may be held liable for tortious interference.¹³

While knowledge may be inferred based on the nature of the product (such as one targeted directly at a particular database or website), it may also be established expressly by sending the third party a letter or email unambiguously placing it on notice of the restrictions and allowing it a reasonable time to discontinue objectionable practices.

Interference with contract claims may be more difficult to

Cal. 4th 1134, 1159 (2003).

¹⁰*Jedson Engineering, Inc. v. Spirit Const. Services, Inc.*, 720 F. Supp. 2d 904 (S.D. Ohio 2010).

¹¹*Pacific Northwest Shooting Park Ass'n v. City of Sequim*, 158 Wash. 2d 342, 351 (2006).

¹²*Trau-Med of Am., Inc. v. Allstate Ins. Co.*, 71 S.W.3d 691, 701 (Tenn. 2002).

¹³*See, e.g., Alticor v. UMG Recordings, Inc.*, No. 6:14-cv-542-Orl-37DAB, 2015 WL 736346, at * 3 (M.D. Fla. Feb. 20, 2015).

establish, however, where a contract is terminable at will.¹⁴

Interference claims also may be hard to establish where the injury is *de minimis*. In *Fields v. Wise Media, LLC*,¹⁵ for example, Northern District of California Judge William Alsup dismissed plaintiffs' intentional interference with contract claim without leave to amend where they alleged that defendants interfered with their mobile phone contracts by sending them unsolicited text messages that they knew would likely cause them to incur charges. Judge Alsup explained that, without reaching the other elements of the claim, "imposing a twenty cent fee does not, as a matter of law, make performance under the contract more costly or burdensome. A reasonable person would not think that a twenty cent charge plausibly increased the cost of plaintiff Field's performance under his mobile phone contract."¹⁶

Where an interference claim is premised on the contents of information posted on a website, a plaintiff must establish that the contents alleged are actionable facts and not merely protected opinion.¹⁷

Interference and other state law causes of action potentially may be preempted by the Copyright Act unless the plaintiff can plausibly allege an extra element beyond merely copying (such as interference).¹⁸ Claims asserted against websites or other intermediaries for merely republishing third party content or hosting third party material or advertisements likewise may be preempted by the Good

¹⁴See, e.g., *Georgia-Pacific Consumer Products LP v. Myers Supply, Inc.*, 621 F.3d 771 (8th Cir. 2010) (affirming entry of judgment for the defendant under Arkansas law in a tortious interference case based on the absence of evidence that the defendant's conduct was unfair or unreasonable and the "strong presumption that interference with an at-will contract is not improper."); Restatement (Second) of Torts § 768 (1979) (providing that interference with a contract that is terminable at will is less likely to be improper).

¹⁵*Fields v. Wise Media, LLC*, No. C 12-05160 WHA, 2013 WL 5340490 (N.D. Cal. Sept. 24, 2013).

¹⁶*Fields v. Wise Media, LLC*, No. C 12-05160 WHA, 2013 WL 5340490, at *5 (N.D. Cal. Sept. 24, 2013).

¹⁷See *Seaton v. TripAdvisor LLC*, 728 F.3d 592, 603 (6th Cir. 2013) (affirming dismissal of a claim under Tennessee law based on TripAdvisor's inclusion of plaintiff's hotel in its list of "2011 Dirtiest Hotels"). *TripAdvisor* did not involve screenscraping. Rather, it involved unwanted attention on a third party's website.

¹⁸See *supra* § 4.18[1] (copyright preemption in general); *infra* § 5.04 (misappropriation and copyright preemption).

Samaritan Exemption to the Telecommunications Act of 1996 (often referred to as the CDA), 47 U.S.C.A. § 230(c).¹⁹

Where it is not possible for a third party to market a circumvention or screen scraping product without actually accessing an owner's database or website and assenting to its user agreement, the database owner also may be able to sue directly for breach of contract (and potentially other claims based on trespass,²⁰ conversion,²¹ or the Computer Fraud and Abuse Act).²²

5.03[6] Unjust Enrichment

Where contract remedies may be unavailable, database owners have sought to assert claims against screen scrapers for unjust enrichment.¹ While the elements of a claim may vary from state to state, in general a plaintiff must show that (1) defendants were enriched; (2) at plaintiffs' expense; and (3) it is against equity and good conscience to permit defendants to retain what is sought to be recovered.² For example, Virginia law requires a plaintiff to show that (1) it conferred a benefit on the defendant, (2) the defendant knew of the benefit and should reasonably have expected to repay

¹⁹See, e.g., *e360Insight, LLC v. Comcast Corp.*, 546 F. Supp. 2d 605 (N.D. Ill. 2008); *Whitney Information Network, Inc. v. Verio, Inc.*, 79 U.S.P. Q.2d 1606, 2006 WL 66724 (M.D. Fla. Jan. 11, 2006); *Corbis Corp. v. Amazon.com, Inc.*, 351 F. Supp. 2d 1090, 1118 (W.D. Wash. 2004) (no liability where images on *Amazon.com* had been provided by a vendor on its zShops platform); *Novak v. Overture Services, Inc.*, 309 F. Supp. 2d 446, 452–53 (E.D.N.Y. 2004) (dismissing the pro se plaintiff's tortious interference claim based on alleged search result manipulation); *Schneider v. Amazon.com, Inc.*, 108 Wash. App. 454, 31 P.3d 37 (Div. 1 2001) (business expectancy); see generally *infra* § 37.05 (analyzing the scope of preemption). Causes of action based on federal copyright and trademark infringement, as well as certain other IP claims, are excluded from the scope of preemption. See *infra* § 37.05[5][B]. The exemption likewise does not exonerate a party for its own conduct. See *infra* § 37.05.

²⁰See *infra* § 5.05[1].

²¹See *infra* § 5.05[2].

²²See *infra* § 5.06.

[Section 5.03[6]]

¹See, e.g., *Information Handling Services, Inc. v. LRP Publications, Inc.*, No. CIV. A. 00-1859, Copy. L. Rep. (CCH) P28,177, 2000 WL 1468535 (E.D. Pa. Sept. 20, 2000) (denying motion to dismiss a claim for unjust enrichment in a database copying case).

²E.g., *Estate of Goth v. Tremble*, 59 A.D.3d 839, 873 N.Y.S.2d 364, 367 (3d Dep't 2009) (applying New York law).

the plaintiff, and (3) the defendant accepted or retained the benefit without paying for its value.³ Similarly, Florida law required a showing that (1) the plaintiff conferred a benefit on the defendant, (2) the defendant had knowledge of the benefit, (3) the defendant accepted or retained the benefit conferred, and (4) the circumstances are such that it would be inequitable for the defendant to retain the benefit without paying for it.⁴ Unjust enrichment is an equitable remedy that will “effect a ‘contract implied in law’ ” to require a party “who accepts and receives the services of another to make reasonable compensation for those services.”⁵

Not all states recognize a separate cause of action for unjust enrichment, however. For example, a separate claim for unjust enrichment may not be asserted under California law.⁶

To the extent a claim for unjust enrichment merely seeks recovery for unauthorized copying without an extra element, however, the claim will be deemed preempted by the Copyright Act.⁷ Similarly, unjust enrichment claims asserted against websites or other intermediaries for merely repub-

³*Rosetta Stone Ltd. v. Google, Inc.*, 676 F.3d 144, 166 (4th Cir. 2012) (Virginia law).

⁴*Resnick v. AvMed, Inc.*, 693 F.3d 1317, 1328 (11th Cir. 2012) (holding that plaintiffs stated a claim for unjust enrichment in a case arising out of a security breach).

⁵*E.g., Rosetta Stone Ltd. v. Google, Inc.*, 676 F.3d 144, 165 (4th Cir. 2012) (Virginia law; quoting other cases).

⁶*See Hill v. Roll Int’l Corp.*, 195 Cal. App. 4th 1295, 1307, 128 Cal. Rptr. 3d 109 (2011) (holding that “[u]njust enrichment is not a cause of action, just a restitution claim.”); *see also, e.g., Astiana v. Hain Celestial Group, Inc.*, 783 F.3d 753, 762 (9th Cir. 2015) (explaining that in California, there is no standalone cause of action for unjust enrichment, which is synonymous with restitution); *Low v. LinkedIn Corp.*, 900 F. Supp. 2d 1010, 1031 (N.D. Cal. 2012) (dismissing with prejudice plaintiffs’ claim for unjust enrichment because such a claim is not viable under California law); *In re iPhone Application Litig.*, 844 F. Supp. 2d 1040, 1075-76 (N.D. Cal. 2012) (dismissing with prejudice plaintiffs’ claim for unjust enrichment based on *Hill v. Roll Int’l Corp.*); *Fraley v. Facebook, Inc.*, 830 F. Supp. 2d 785, 814-15 (N.D. Cal. 2011) (dismissing a claim for unjust enrichment in light of *Hill v. Roll Int’l Corp.*, “[n]otwithstanding earlier cases suggesting the existence of a separate, stand-alone cause of action for unjust enrichment”); *In re iPhone Application Litig.*, Case No. 11-MD-02250-LHK, 2011 WL 4403963, at *15 (N.D. Cal. Sept. 20, 2011) (dismissing plaintiffs’ claim for unjust enrichment, finding there is no longer any such cognizable claim under California law).

⁷*See, e.g., BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596,

lishing third party content or hosting third party material such as database content may be preempted by the CDA, 47 U.S.C.A. § 230(c).⁸ The CDA, however, does not insulate an interactive computer service provider or user for their own content.

5.04 Common Law Misappropriation and Unfair Competition

Database owners may have claims against parties that scrape material from their websites based on common law misappropriation or unfair competition, to the extent those claims are not preempted by the Copyright Act or, in specific circumstances, the Patent Act, the Uniform Trade Secrets Act or the Communications Decency Act. Those issues are addressed in the following subsections.

5.04[1] Misappropriation (including the “Hot News” Doctrine)

Limited protection for databases may be available under the common law doctrine of misappropriation, to the extent not preempted by the Copyright Act, where an extra element such as breach of fiduciary duty is alleged. Many misappropriation claims brought over database copying arise under the “hot news” doctrine, which makes potentially actionable misappropriation of material that is valuable for its timeliness (such as breaking news stories, stock tips or celebrity news photos), where the copying involves free-riding by a competitor and a threat to the very existence of the product or service supplied by the plaintiff. While a number of courts have recognized and applied the hot news doctrine based on common law misappropriation, as outlined below, a small

618 (S.D.N.Y. 2010); *Snap-on Business Solutions Inc. v. O’Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 680–81 (N.D. Ohio 2010) (holding plaintiff’s unjust enrichment claim preempted where it was based on the allegation that defendants took information). *But see Perfect 10, Inc. v. Google, Inc.*, No. CV04-9484, 2008 WL 4217837, at *9 (C.D. Cal. July 6, 2008) (holding plaintiff’s unjust enrichment claim not preempted where the claim was premised on right of publicity and trademark violations); *see generally supra* § 4.18[1] (analyzing copyright preemption).

⁸*See, e.g., Ascentive, LLC v. Opinion Corp.*, 842 F. Supp. 2d 450 (E.D.N.Y. 2011); *Rosetta Stone Ltd. v. Google Inc.*, 732 F. Supp. 2d 628 (E.D. Va. 2010) (holding plaintiff’s unjust enrichment claim preempted by the CDA), *aff’d in relevant part on other grounds*, 676 F.3d 144, 165–66 (4th Cir. 2012); *see generally infra* § 37.05 (analyzing CDA preemption).

number have declined to so.¹

In *International News Service v. Associated Press*,² a 1918 U.S. Supreme Court decision, the defendant copied AP stories from bulletin boards and early editions of East Coast newspapers and then transmitted and sold paraphrased versions of these stories to newspapers on the West Coast. Although not entitled to copyright protection, the Supreme Court held that breaking news was the “quasi property” of a news-gathering operation and copying this information constituted common law misappropriation.³

The Court, in *International News Service*, emphasized that newsgathering carried with it “the expenditure of labor, skill, and money” and that when another party appropriates breaking news, it “is endeavoring to reap what it has not sown.”⁴

Since the time the Court decided *International News Service*, the Copyright Act was amended to expressly preempt state remedies equivalent to those protected by the Copyright Act.⁵ A common law misappropriation claim will be preempted if it lacks an “extra element” necessary to sup-

[Section 5.04[1]]

¹See, e.g., *Allure Jewelers, Inc. v. Ulu*, No. 1:12CV91, 2012 WL 4322519, at *3 (S.D. Ohio Sept. 20, 2012) (declining to recognize a “hot news” exception under Ohio law for a jeweler’s posting of recent information on the price of precious metals in an eBay advertisement for fine jewelry); *Brainard v. Vassar*, 561 F. Supp. 2d 922, 932 (M.D. Tenn. 2008) (noting that “plaintiffs have cited no case law indicating that the Tennessee courts have adopted New York’s ‘hot-news’ causes of action” but ultimately holding that the hot news doctrine did not apply to the case and that plaintiffs’ claim for common law misappropriation based on appropriation of a song title preempted by the Copyright Act); *Ultra-Precision Mfg., Ltd. v. Ford Motor Co.*, 01-70302, 2002 WL 32878308, at *4 (E.D. Mich. May 31, 2002) (finding no support for a cause of action for commercial misappropriation under Michigan law).

²*International News Service v. Associated Press*, 248 U.S. 215 (1918).

³Federal common law subsequently was abolished by *Erie R. Co. v. Tompkins*, 304 U.S. 64 (1938). As discussed below, the general common law principles of misappropriation in breaking news cases articulated in *International News* remain valid under state common law, to the extent not preempted by the Copyright Act.

⁴*International News Service v. Associated Press*, 248 U.S. 215, 239–40 (1918).

⁵See 17 U.S.C.A. § 301.

port an independent claim.⁶ A claim for common law misappropriation based on copying therefore will be limited to circumstances such as where a confidential relationship may be shown or where there was some other breach of trust or agreement or passing off⁷ or misappropriation under the hot news doctrine.⁸

In *National Basketball Association v. Motorola, Inc.*,⁹ the Second Circuit explained *International News Service* as a “hot news” case, which has continuing validity under New York state law to the extent it is limited to its particular facts (one wire service taking advantage of a time delay to profit from the other company’s collection of current news stories). The Second Circuit held that a claim for common law misappropriation, although not established in that case, may be asserted, and will not be preempted, where (1) a plaintiff generates or gathers information at a cost, (2) the information is time-sensitive, (3) a defendant’s use of the information constitutes free-riding on the plaintiff’s efforts, (4) the defendant is in direct competition with a product or service offered by plaintiff, and (5) the ability of other parties to free-ride on the efforts of the plaintiff or others would so reduce the incentive to produce the product or service that its existence or quality would be substantially threatened.

Although the Second Circuit in *Motorola* reversed the injunction that had been issued by the lower court based on its determination that the plaintiff could not make out a claim for hot news misappropriation, an increasing number of Internet-related suits have been brought over the years

⁶*E.g.*, *Kregos v. Associated Press*, 3 F.3d 656, 666 (2d Cir. 1993); *Summit Mach. Tool Mfg. Corp. v. Victor CNC Systems, Inc.*, 7 F.3d 1434, 1441–42 (9th Cir. 1993); *Quadrille Wallpapers and Fabric, Inc. v. Pucci*, No.1:10-CV-1394, 2011 WL 3794238, at *9 (N.D.N.Y. Aug. 24, 2011) (finding that plaintiff’s unfair competition and misappropriation claims were not preempted because the bases of both claims under state law required an additional element of a breach of confidential relationship); *see generally supra* § 4.18[1] (copyright preemption).

⁷*See, e.g.*, *Cable Vision, Inc. v. KUTV, Inc.*, 335 F.2d 348, 352 (9th Cir. 1964) (*International News* was premised on a theory of “passing off.”), *cert. denied*, 379 U.S. 989 (1965).

⁸*See Financial Information, Inc. v. Moody’s Investors Service, Inc.*, 808 F.2d 204, 209 (2d Cir. 1986) (explaining that hot news misappropriation is “a branch of the unfair competition doctrine, not preempted by the Copyright Act according to the House Report.”).

⁹*National Basketball Ass’n v. Motorola, Inc.*, 105 F.3d 841 (2d Cir. 1997).

(particularly in New York and California) relying on *Motorola* for both its preemption analysis and description of the elements required to state a claim for common law misappropriation under New York law in a case involving “hot news.” In 2011, however, a later Second Circuit panel potentially narrowed its reach—at least as interpreted in some subsequent lower court decisions—and the majority disavowed *Motorola*’s five-part test as *dicta* not necessary to the court’s holding in that case.

In *Barclays Capital, Inc. v. TheFlyOnTheWall.com*,¹⁰ Judge Sack (writing for himself and Judge Pooler, in a case where Judge Raggi filed a concurring opinion) emphasized that the scope of the preemption exception was “narrow,” cautioning that “[t]he broader the exemption, the greater the likelihood that protection of works within the ‘general scope’ of the copyright and the type of works protected by the Act will receive disparate treatment depending on where the alleged tort occurs and which state’s law is found to be applicable.”¹¹ The majority also rejected the “moral dimension” to hot news misappropriation cases, concluding that “unfairness alone is immaterial to a determination whether a cause of action for misappropriation has been preempted by the Copyright Act.”¹² The majority instead applied the traditional three-part test for evaluating copyright preemption, evaluating (1) whether a claim seeks to vindicate “legal or equitable rights that are equivalent” to one of the bundle of exclusive rights already protected by copyright law under 17 U.S.C.A. § 106 (reproduction, distribution, public performance, public display and the right to make derivative works),¹³ (2) whether the work in question falls within the ambit of copyright protection (which may include both protectable and uncopyrightable works) and (3) if an “extra element” is

¹⁰*Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876 (2d Cir. 2011).

¹¹*Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876, 897 (2d Cir. 2011).

¹²*Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876, 896 (2d Cir. 2011). The majority noted that:

The adoption of new technology that injures or destroys present business models is commonplace. Whether fair or not, that cannot, without more, be prevented by application of the misappropriation tort. Indeed, because the Copyright Act itself provides a remedy for wrongful copying, such unfairness may be seen as supporting a finding that the Act preempts the tort.

Id.

¹³*See supra* § 4.04[1].

“required instead of or in addition to the acts of reproduction, performance, distribution or display, in order to constitute a state-created cause of action”¹⁴ Based on this test, the Second Circuit reversed the entry of judgment for the plaintiffs, finding their claims preempted. Judge Raggi concurred, although she would have applied the five-part *Motorola* test rejected by the other judges based on her conclusion that it was not mere *dictum*. In her view, the plaintiffs failed to satisfy the “direct competition” requirement of the *Motorola* test.¹⁵

In *Fox News Network, LLC v. TVEyes, Inc.*,¹⁶ Judge Alvin K. Hellerstein of the Southern District of New York followed *Barclays* in holding that a news network’s hot news misappropriation claim against a service that recorded all television and radio broadcasts for more than 1,400 stations, 24 hours a day, every day, and transformed this material into a searchable database for its paying subscribers, which included the White House, more than 100 members of Congress and ABC Television, was preempted because Fox’s claim did not include an extra element besides copying. In that case, TVEyes offered paying business subscribers an indexing and clipping service that allowed them to search for and find video excerpts for purposes such as evaluating and criticizing broadcast journalism, tracking and correcting misinformation, evaluating commercial advertising, evaluating national security risks, and tracking compliance with financial market regulations. In holding Fox’s claim preempted, Judge Hellerstein explained that “TVEyes is not a valuable service because its subscribers credit it as a reliable news outlet, it is valuable because it reports what the news outlets and commentators are saying and therefore does not ‘scoop’ or free-ride on the news services.”¹⁷

Courts previously had allowed hot news misappropriation

¹⁴*Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876, 892–907 (2d Cir. 2011), *quoting in part Computer Associates Int’l, Inc. v. Altai, Inc.*, 982 F.2d 693, 716 (2d Cir. 1992) (quoting Nimmer on Copyright § 1.01[B], at 1-14-15 (1991)).

¹⁵The *FlyInTheWall* case is discussed in greater detail later in this section.

¹⁶*Fox News Network, LLC v. TVEyes, Inc.*, 43 F. Supp. 3d 379 (S.D.N.Y. 2014).

¹⁷*Fox News Network, LLC v. TVEyes, Inc.*, 43 F. Supp. 3d 379, 399 (S.D.N.Y. 2014). This particular ruling was not challenged on appeal. *See Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 174 n.2 (2d Cir.

claims to proceed based on *Motorola* in a number of Internet cases. In *Pollstar v. Gigmania Ltd.*,¹⁸ for example, a federal district court in California denied the defendant's motion to dismiss, allowing plaintiff's suit for misappropriation (and unfair competition based on palming off) to proceed. In that case, the plaintiff had provided access to a database it created that included current concert information, subject to posted terms and conditions. The court rejected the defendant's arguments that plaintiff's misappropriation and unfair competition claims were preempted by the Copyright Act, holding that the plaintiff had sufficiently pled a "hot news" claim.

Similarly, in *Facebook, Inc. v. ConnectU LLC*,¹⁹ the court found that Facebook's misappropriation claim against a competitor was not preempted where the competitor collected the email addresses of Facebook's registered users, posted them on its website and then sent unsolicited commercial email to those users. The court found that the email addresses were not works of authorship or otherwise protectable under the copyright Act and that the defendant "had not shown that it was alleged to have misappropriated uncopyrightable 'elements' of a work of authorship otherwise within the scope of the Copyright Act."²⁰

In *Weingand v. Harland Financial Solutions, Inc.*,²¹ the court similarly found plaintiff's unjust enrichment claim not preempted because the material at issue was confidential data such as social security numbers, addresses, and bank account information, and thus not copyrightable, and the plaintiff alleged the additional element of a contractual relationship.

In *Chicago Board Options Exchange, Inc. v. International Securities Exchange, LLC*,²² an intermediate appellate court in Illinois affirmed the entry of summary judgment in favor

2018).

¹⁸*Pollstar v. Gigmania, Ltd.*, 170 F. Supp. 2d 974 (E.D. Cal. 2000).

¹⁹*Facebook, Inc. v. ConnectU LLC*, 489 F. Supp. 2d 1087 (N.D. Cal. 2007).

²⁰*Facebook, Inc. v. ConnectU LLC*, 489 F. Supp. 2d 1087, 1092–93 (N.D. Cal. 2007).

²¹*Weingand v. Harland Financial Solutions, Inc.*, No. C-11-3109 EMC, 2012 WL 2327660, at *7 (N.D. Cal. June 19, 2012).

²²*Chicago Board Options Exchange, Inc. v. International Securities Exchange, LLC*, 973 N.E.2d 390 (Ill. App. 2012).

of the plaintiffs, a national securities exchange and index providers, on their claim for common law misappropriation against a competing securities exchange, which had announced its intention to offer index options based on stock indices, without obtaining a license. The court held that plaintiffs' claim was not preempted because it was premised on unlicensed use of plaintiff's ideas, systems, and concepts, which are not entitled to copyright protection—namely unauthorized use of the research, expertise, reputation and goodwill associated with plaintiff's product. The court also held that the defendant's proposed use of the index for its own financial products constituted common law misappropriation under Illinois law.

A court in California likewise held that celebrity news site X17 stated a claim for hot news misappropriation against blogger Perez Hilton for pervasive copying of its paparazzi photographs,²³ although Hilton defeated a motion for a preliminary injunction in a similar suit brought in New York for copying facts from another website where the court found that the information allegedly copied was widely available over the Internet (usually before it had been posted on the plaintiff's own website) and the plaintiff had not shown that she had incurred significant costs compiling it.²⁴ In the California action, the court found that X17 stated a claim where it alleged that: (1) it expended substantial costs and resources to gather, obtain, and create the photographs that Lavandeira (Perez Hilton) disseminated; (2) the photographs were time-sensitive; (3) the parties were direct competitors; (4) Lavandeira was earning revenue by free-riding on the substantial hard work of X17; (5) if the activities continued, they would remove X17's incentive to gather celebrity news photographs and threaten the continued existence of its business; and (6) Lavandeira's activities had substantially harmed X17.²⁵

In so ruling, Judge Feess rejected the defendant's argu-

²³See *X17, Inc. v. Lavandeira*, 563 F. Supp. 2d 1102 (C.D. Cal. 2007).

²⁴See *Silver v. Lavandeira*, No. 08 Civ. 6522(JSR)(DF), 2009 WL 513031 (S.D.N.Y. Feb. 26, 2009) (denying plaintiff's motion for a preliminary injunction because the plaintiff was unlikely to prevail on copyright infringement, DMCA and hot news misappropriation claims).

²⁵See *X17, Inc. v. Lavandeira*, 563 F. Supp. 2d 1102, 1108–09 (C.D. Cal. 2007). The court was ruling on a motion to dismiss. Judge Feess emphasized that “[w]hether or not X17 can prove its case is a matter that the Court does not address in this ruling. The Court concludes only that

ment that a hot news claim was limited to factual information. He held that a claim could be based on photographs, noting that the medium was not important in either *International News Service* or *Motorola*. The court also observed in *dicta* that misappropriation claims were not limited to “hot news” so long as an extra element beyond copying was alleged (so that the claim would not be preempted). Among other things, Judge Feess noted that misappropriation claims could be based on breach of a fiduciary duty, in addition to hot news.

In *Associated Press v. All Headline News Corp.*,²⁶ a court in the Southern District of New York denied the defendant’s motion to dismiss a claim for misappropriation of “hot” or “breaking news,” where the plaintiff alleged each of the *NBA v. Motorola* elements in connection with defendant’s operation of All Headline News Corp. (AHN), an Internet site that does not undertake any original reporting and allegedly hired poorly paid individuals to find news stories on the Internet (including AP stories) and prepare them for republication as AHN stories by either rewriting the articles or copying the stories in full. Plaintiffs alleged that, among other things, defendants copied AP’s breaking news stories and reproduced them as stories that originated with AHN.²⁷

The court also denied defendants’ motion to dismiss AP’s state law unfair competition claim, finding that the plaintiff had stated a claim for passing off by alleging that AHN passed off AP content as its own.²⁸ Neither plaintiffs’ “hot

X17 has adequately pled the . . . elements required to state a claim for California’s misappropriation tort . . .” and the additional elements to come within the sub-set of misappropriation claims based on hot news. *Id.* at 1108.

²⁶*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454 (S.D.N.Y. 2009).

²⁷*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454, 461 (S.D.N.Y. 2009). *All Headline News* was decided prior to the Second Circuit’s decision in *Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876 (2d Cir. 2011), which narrowly construed the scope of “hot news” claims of common law misappropriation that could survive copyright preemption. Nevertheless, as discussed later in this section, the majority cited *All Headline News* approvingly as a case closer to the facts of *International News Service* than the one it was deciding.

²⁸*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454, 464 (S.D.N.Y. 2009). The court also ruled that the Associated Press could state a DMCA claim for removal of copyright management information

news” nor its passing off claims were found preempted.²⁹

In a subsequent case also from the Southern District of New York, *BanxCorp. v. Costco Wholesale Corp.*,³⁰ a hot news claim based on the allegation that 100% of a continuously updated database, including at least some hot news, was misappropriated by a defendant that allegedly exceeded the scope of its license agreement, likewise was held not to be preempted. The court, in ruling on the defendant’s motion to dismiss, held that claims for misappropriation based on hot news and breach of a license agreement, as alleged, were not preempted, but claims for unfair competition and unjust enrichment were preempted.

“Hot news” misappropriation generally has been easier to allege than prove and most Internet cases to date have been resolved through motion practice or settlement. In the one case to proceed to judgment, *Barclays Capital, Inc. v. TheFlyOnTheWall.com*,³¹ the trial court had entered judgment for the plaintiffs following a bench trial based on common law misappropriation of time-sensitive stock recom-

(see *infra* § 5.07[2]), but dismissed certain Lanham Act claims, holding that defendant’s use of the trademarked term “AP” or “Associated Press” in connection with phrases such as “According to an AP report,” to attribute certain facts to the Associated Press, did not constitute trademark infringement, that the defendant’s characterization of itself as a “news service” or news-gathering organization did not constitute false designation of origin, and that defendants did not make false or misleading representations to consumers by editing news stories to omit the creators of original source material when paraphrasing them, but citing them directly by name when quoting them.

²⁹Prior to *Barclay’s*, courts had almost universally held that hot news claims that incorporate the elements set forth in *Motorola* are not preempted by the Copyright Act. In *Lowry’s Reports, Inc. v. Legg Mason, Inc.*, 271 F. Supp. 2d 737, 754–57 (D. Md. 2003), however, the court, citing law review articles, held that the plaintiff’s hot news claim in that case was preempted. The court found that allegations such as “free riding” were equivalent to copyright claims and did not amount to an extra element. The court’s cursory analysis in *Legg* has not been followed by other courts and has been criticized. See *X17, Inc. v. Lavandeira*, 563 F. Supp. 2d 1102, 1106 (C.D. Cal. 2007) (rejecting *Legg* as unpersuasive and inconsistent with the legislative history of the Copyright Act).

³⁰*BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 611–18 (S.D.N.Y. 2010). *BanxCorp.* was decided prior to the Second Circuit’s decision in *Barclays Capital Inc. v. Theflyonthewall.com, Inc.*, 650 F.3d 876 (2d Cir. 2011), which narrowly construed the scope of copyright preemption for “hot news” claims of common law misappropriation.

³¹*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310 (S.D.N.Y. 2010), *rev’d in part*, 650 F.3d 876 (2d Cir. 2011).

mentations (in addition to copyright infringement), but the judgment was reversed on appeal by the Second Circuit, which remanded the case with instructions to dismiss plaintiffs' claim as preempted by the Copyright Act. The defendant in that case did not appeal the court's judgment and entry of an injunction under the Copyright Act.

In *TheFlyOnTheWall*, plaintiffs Barclay's Capital, Merrill Lynch and Morgan Stanley, sued the defendant for unauthorized dissemination of its company analysis and recommendations, which were primarily valuable (to potential investors seeking to buy or sell stocks at a profit) for their timeliness, often overnight before the market opened. Initially, TheFlyOnTheWall.com disseminated plaintiff's copyrighted reports verbatim, but after receiving a cease and desist letter from the plaintiffs, changed its practices to simply disseminate headlines (such as "EQIX: Equinox initiated with a Buy at Bofa/Merrill. Target \$110")—typically around 600 per day—drawn from sixty-five investment firms' research analysts, including the three plaintiff firms. Over time, it diversified its news sources. Whereas recommendations from plaintiffs' firms accounted for 7 percent of its newsfeed in 2005, by 2009 they represented only approximately 2.5 percent. The defendant's sources for the information from plaintiffs also changed over time. Initially, TheFlyOnTheWall relied entirely on employees of plaintiffs' firms, who transmitted the reports to it directly (without authorization), allowing the defendant to disseminate the information to its own subscribers before the opening of the market. As a result of the litigation, the defendant claimed to have stopped looking at plaintiffs' reports directly and instead relied on information received from independent sources, confirming reports from two or three independent sources before publishing them.

In holding the defendant liable for hot news misappropriation, the trial court had rejected the argument that by obtaining plaintiffs' recommendations from sources other than plaintiffs, it was obtaining information that was already "public" and therefore could be freely republished.

District Court Judge Cote had analyzed the *Motorola* factors, finding that they all supported liability. First, with respect to the cost of information, the district court found that plaintiffs collectively employed hundreds of skilled analysts and spent hundreds of millions of dollars each year to produce their equity research reports.

Second, the district court had held that plaintiffs showed that the value of the information generated or collected by them was highly time-sensitive. The value of stock tips, the court ruled, was in disseminating them while they were “fresh.” Judge Cote found that plaintiffs’ clients used the analysts’ opinions “to execute trades in anticipation of stock price movement in order to capture the maximum benefit from the movement.” To reap the greatest benefit from their research reports through commission income, the court wrote that the plaintiffs had to engage “in a costly, frenzied process to try to be the first to inform their clients” of their recommendations.³² District Court Judge Cote also found that the defendant’s own conduct verified the time-sensitive nature of the data by emphasizing this fact to its own subscribers and business partners and by suing one of its own competitors for hot news misappropriation and alleging that the value of its newsfeeds was highly time sensitive.

Third, the trial court held that the defendant’s use of plaintiffs’ information constituted free-riding on the plaintiff’s costly efforts to generate or collect it (or the diversion of value from a business rival’s efforts without payment). The defendant did no research of its own, allowing it to sell plaintiffs’ recommendations at a cut-rate price. Although the defendant provided attribution for the recommendations it relayed to its subscribers and business partners, Judge Cote wrote that this fact merely “underscore[d] its pilfering.”³³ The trial court was unimpressed with defendant’s argument that many others in the industry did the same, writing that “[t]he fact that others also engage in unlawful behavior does not excuse a party’s own illegal conduct.”³⁴

The defendant had argued that because it no longer received plaintiffs’ information directly from plaintiffs, but instead from third parties, it was merely disseminating information that was already in the marketplace. The trial court, however, explained that the legally salient fact was not that lawful subscribers repeat news to their friends or colleagues, which is permissible, it is that plaintiffs system-

³²*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 336 (S.D.N.Y. 2010), *rev’d in part*, 650 F.3d 876 (2d Cir. 2011).

³³*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310 (S.D.N.Y. 2010), *rev’d in part*, 650 F.3d 876 (2d Cir. 2011).

³⁴*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 337 (S.D.N.Y. 2010), *rev’d in part*, 650 F.3d 876 (2d Cir. 2011).

atically gathered plaintiffs' recommendations (even if indirectly from third parties) and then used it to run a profitable business dedicated to systematically gathering and selling plaintiffs' recommendations.³⁵

The trial court had also rejected defendant's argument that its reports included "much more" than merely plaintiffs' recommendations, noting that liability may be imposed even where misappropriation only relates to a small part of a defendant's business (as was the case in *International News Service*).

Fourth, the trial court found that the defendant was a direct competitor of plaintiffs in the area of plaintiffs' "primary business." The defendant also was found to have taken steps to compete even more directly by aligning itself with discount brokerage houses that could execute trades. Judge Cote also noted that the defendant previously had asserted a counterclaim for unfair competition against the plaintiffs.

Fifth, the trial court found that the defendant's ability to freeride on plaintiffs' efforts could so reduce the incentive to produce the reports that their existence or quality were substantially threatened absent injunctive relief. Plaintiffs had shown both significant losses and the risk of continued misappropriation by the defendant.

In addition to entering judgment for plaintiffs on their hot news misappropriation claim, the district court entered judgment in their favor on their copyright infringement claim, issuing a permanent injunction and awarding statutory damages, prejudgment interest and attorneys' fees (but only the portion of fees directly and predominantly concerned with the prosecution of plaintiffs' copyright claim, potentially reduced in light of the disparity in resources between the plaintiffs—major investments firms—and the defendant, and defendant's financial condition).³⁶

Despite the strong opinion, the injunction actually issued by Judge Cote was narrow. Plaintiffs had sought an injunction against reporting a stock recommendation until the later of four hours after it was released or 12:00 PM Eastern Time.

³⁵*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 338 (S.D.N.Y. 2010), *rev'd in part*, 650 F.3d 876 (2d Cir. 2011).

³⁶*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 328–31 (S.D.N.Y. 2010), *rev'd in part*, 650 F.3d 876 (2d Cir. 2011); *see generally supra* §§ 5.02 (copyright protection), 4.14 (copyright damages), 4.15 (attorneys' fees under the Copyright Act).

The court, however, only enjoined the defendant from disseminating information from research reports released while the market is closed until the later of thirty minutes after the market closed or 10:00 AM Eastern Time.

Judge Cote also invited the defendant to seek a reevaluation of the order in a year's time if, during that time, plaintiffs did not also take action against defendant's competitors. The trial court observed that since Fly "first built its business around the misappropriation of" plaintiffs' reports and recommendations, the practice of posting this information had "become a widespread phenomenon."³⁷ Accordingly, Judge Cote wrote that "[i]t would be unjust to restrain Fly from publishing" plaintiffs' recommendations if plaintiffs "were to acquiesce in the unauthorized publication . . . by others"³⁸ The trial court held that the defendant could "apply to modify or vacate the injunction" if it could demonstrate that plaintiffs did not "take reasonable steps to restrain the systematic, unauthorized misappropriation of their Recommendations, for instance, through the initiation of litigation against any parties with whom negotiation proves unsuccessful."³⁹

In reversing the judgment for the defendant on plaintiffs' claim for common law misappropriation, the majority of the Second Circuit panel that considered the case (as noted earlier) rejected the five-part *Motorola* test as based on *dicta*, but purported to apply *Motorola*'s actual holding. Noting that it was not determinative for preemption purposes that the facts contained in plaintiffs' recommendations were not entitled to copyright protection, the majority found that the reports and recommendations fell within the "general scope" of the Copyright Act. The majority took issue with the use of the term "free riding" in hot news jurisprudence, noting that in *International News Service*, as underscored in *Motorola*, free riding involved taking material from a plaintiff and selling it as the defendant's own. Applying this narrower understanding of free riding to plaintiffs' recommendations—the conclusions contained in reports on whether to buy or

³⁷*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 347 (S.D.N.Y. 2010), *rev'd in part*, 650 F.3d 876 (2d Cir. 2011).

³⁸*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 347 (S.D.N.Y. 2010), *rev'd in part*, 650 F.3d 876 (2d Cir. 2011).

³⁹*Barclays Capital Inc. v. Theflyonthewall.com*, 700 F. Supp. 2d 310, 347–48 (S.D.N.Y. 2010), *rev'd in part*, 650 F.3d 876 (2d Cir. 2011).

sell a stock and what the target value should be—the majority concluded that TheFlyOnTheWall was very different from the defendant in *International News Service*. The majority underscored that the plaintiffs’ recommendations were “*create[d]*/using their expertise and experience, rather than *acquire[d]* through efforts akin to reporting.”⁴⁰ In the majority’s view, “[t]he Firms are making the news; Fly, despite the Firms’ understandable desire to protect their business model, is breaking it.”⁴¹ In addition, the majority emphasized that TheFlyOnTheWall was not selling the recommendations as its own, which further distinguished the case from *International News Service*.

The majority also found significant the fact that the Supreme Court in *International News Service* referred to the defendant’s tortious behavior as “amount[ing] to an unauthorized interference with the normal operation of complainant’s business *precisely at the point where the profit is to be reaped*, in order to *divert a material portion of the profit* from those who have earned it to those who have not”⁴² By contrast, although the majority conceded that the plaintiffs would likely earn more revenue if their recommendations were not disseminated by third parties and that there was some evidence that TheFlyOnTheWall had linked some of its own subscribers to competing discount brokerage services, the majority did not view its publication of the recommendations as “an unauthorized interference with the normal operations of [plaintiffs’] legitimate business *precisely at the point where the profit is to be reaped*”⁴³

Judge Reena Raggi, concurring, would have applied the *Motorola* test, finding it binding (and not *dictum*), but concurred based on her conclusion that direct competition could not be shown. Like the majority, she premised her opinion on the fact that the defendant produced an aggregate product reporting recommendations from many different firms, among other financial news, attributing each rec-

⁴⁰*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 903 (S.D.N.Y. 2011)(emphasis in original).

⁴¹*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 902 (S.D.N.Y. 2011).

⁴²*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 904 (S.D.N.Y. 2011), quoting *International News Service v. Associated Press*, 248 U.S. 215, 240 (1918) (emphasis added by the Second Circuit).

⁴³*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 904–05 (S.D.N.Y. 2011).

ommendation to its source. The majority, she wrote, drew “a bright line distinguishing between the Firms, who generate news, and Fly and other news aggregators, who ‘break’ the news, with the former falling outside of hot-news protection.”⁴⁴ By contrast, she wrote that she was “not prepared to foreclose the possibility of a ‘hot-news’ claim by a party who disseminates news it happens to create.”⁴⁵ Instead, she concluded that plaintiffs could not state a non-preempted claim because plaintiffs and defendant were not direct competitors. *Motorola*, she explained, involved the “plaintiff’s failure to show free riding on and a sufficient threat to its services,” but also underscored that “only products in the ‘keenest’ of competition satisfy the direct competition requirement for a non-preempted claim.”⁴⁶

Although the majority disclaimed that it did “not mean to be parsing the language of *INS* as though it were a statement of law the applicability of which determines the outcome of this appeal”⁴⁷ the majority, in fact, applied *International News Service* very narrowly to its literal facts, ascribing as *dicta* broader notions of “free riding” that were seemingly approved-of by the Second Circuit panel in *Motorola*. The net effect, at least in the Second Circuit, is to scale back the circumstances under which a claim of hot news misappropriation may be brought to circumstances where a competitor takes valuable data and seeks to sell it as its own under circumstances akin to *International News Service*.⁴⁸

Even if otherwise viable, misappropriation claims may be

⁴⁴*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 913 (S.D.N.Y. 2011) (Raggi, J., concurring).

⁴⁵*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 913 (S.D.N.Y. 2011) (Raggi, J., concurring).

⁴⁶*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 913 (S.D.N.Y. 2011) (Raggi, J., concurring), quoting *International News Service v. Associated Press*, 248 U.S. 215, 221 (1918).

⁴⁷*Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 905 (S.D.N.Y. 2011).

⁴⁸Despite its characterization of much of the *Motorola* decision as *dicta*, the majority speculated that “[i]f a Firm were to collect and disseminate to some portion of the public facts about securities recommendations in the brokerage industry (including, perhaps, such facts it generated itself—its own Recommendations), and were Fly to copy the facts contained in the Firm’s hypothetical service, it might be liable to the Firm on a ‘hot-news’ misappropriation theory.” *Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 905–06 (S.D.N.Y. 2011).

preempted by the Uniform Trade Secrets Act (where enacted), which provides that the Act “displaces conflicting tort, restitutionary, and other law of this State pertaining to civil liability for misappropriation of a trade secret.”⁴⁹ Section 7 has been construed in some (but not all jurisdictions) to preempt claims premised on the wrongful taking and use of confidential business and proprietary information, regardless of whether the information constitutes a trade secret.⁵⁰ Claims against interactive computer service providers for misappropriation by third parties also may be preempted by the Communications Decency Act.⁵¹

5.04[2] Unfair Competition

Most states have enacted statutes or recognize common law claims for unfair competition. The heading “unfair competition” may include claims for common law misappropriation.

The Second Circuit majority also cited approvingly *Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454 (S.D.N.Y. 2009), which was discussed earlier in this section, as a case “presenting facts more closely analogous to INS . . .” *Barclays Capital Inc. v. Theflyonthewall.com*, 650 F.3d 876, 906 (S.D.N.Y. 2011). On the other hand, the fact that in *All Headline News* the parties had argued over choice of law, assuming that a hot news claim under New York law would not be viable if Florida law were to have been applied, was cited by the majority as a reason why the doctrine should be narrowly construed to promote greater uniformity, rather than the “sort of patchwork protection that the drafters of Copyright Act preemption provisions sought to minimize . . .” *Id.* at 897–98.

⁴⁹UTSA § 7; *infra* § 10.17 (addressing UTSA preemption). A copy of the UTSA is reprinted in the appendix to chapter 10.

⁵⁰*See, e.g., Heller v. Cepia, LLC*, No. C 11–01146 JSW, 2012 WL 13572, at *7 (N.D. Cal., Jan. 4, 2012) (dismissing claims for common law misappropriation, conversion, unjust enrichment, and trespass to chattels, because these claims, “premised on the wrongful taking and use of confidential business and proprietary information, regardless of whether such information constitutes trade secrets, are superseded by the CUTSA.”); *Glasstech, Inc. v. TGL Tempering Sys., Inc.*, 50 F. Supp. 2d 722, 730 (N.D. Ohio 1999) (holding common law claims for misuse and misappropriation, unfair competition, and unjust enrichment preempted by Ohio’s Uniform Trade Secrets Act); *see generally infra* § 10.17 (discussing conflicting lines of cases on whether a claim is preempted even if based on information that may not be protectable as a trade secret).

⁵¹47 U.S.C.A. § 230(c); *Stevo Design, Inc. v. SBR Mktg. Ltd.*, 919 F. Supp. 2d 1112, 1127 (D. Nev. 2013) (dismissing plaintiffs’ complaint for common law misappropriation under Florida law with leave to amend); *infra* § 10.17; *see generally infra* § 37.05 (analyzing the CDA in substantially greater detail).

tion,¹ trademark infringement or dilution, passing off or equivalent state law corollaries to the remedies available under the federal Lanham Act,² and potentially even broader claims based on any action that may be deemed unfair if undertaken by a business or potential competitor. Under California's infamous unfair competition statute, Business & Professions Code § 17200, for example, a plaintiff may sue for "any unlawful, unfair or fraudulent business act or practice and unfair, deceptive, untrue or misleading advertising . . ." and any act prohibited by California's false advertising statute.³ Any violation of a state or federal law (including those that do not afford private causes of action) or even a state or federal policy potentially may be actionable under this very broad statute, so long as the plaintiff may show that it has "suffered injury in fact and has lost money or property as a result of such unfair competition."⁴ Indeed, "a practice may be deemed unfair even if not specifically proscribed by some other law."⁵

[Section 5.04[2]]

¹See *Fox News Network, LLC v. TVEyes, Inc.*, 43 F. Supp. 3d 379, 399–400 (S.D.N.Y. 2014) (holding that a New York state law misappropriation claim, grounded in either deception or appropriation of the exclusive property of a plaintiff, may be maintained in certain circumstances "based on the equitable doctrine that recognizes that 'a person shall not be allowed to enrich himself unjustly at the expense of another.'"; quoting *Georgia Malone & Co. v. Rieder*, 19 N.Y.3d 511, 516, 950 N.Y.S.2d 333, 973 N.E.2d 743 (2012)); see also *Financial Information, Inc. v. Moody's Investors Service, Inc.*, 808 F.2d 204, 209 (2d Cir. 1986) (explaining that hot news misappropriation is "a branch of the unfair competition doctrine, not preempted by the Copyright Act according to the House Report."); *supra* § 5.04[1] (analyzing hot news misappropriation). In *Fox News*, the court held that plaintiff's unfair competition claim, like its claim for hot news misappropriation, was preempted by the Copyright Act. See *Fox News Network, LLC v. TVEyes, Inc.*, 43 F. Supp. 3d 379, 399–400 (S.D.N.Y. 2014) (holding that bad faith or a bad intent did not constitute an extra element); see generally *supra* § 5.04[1] (discussing Fox's hot news claim).

²See *infra* § 5.08.

³Cal. Bus. & Prof. Code § 17200. The provisions of California's false advertising statute that are also expressly made actionable under section 17200 are codified at Cal. Bus. & Prof. Code §§ 17500 to 17580; see generally *infra* § 6.12[6] (analyzing section 17200 in greater detail).

⁴Cal. Bus. & Prof. Code § 17200. "An injury in fact is '[a]n actual or imminent invasion of a legally protected interest, in contrast to an invasion that is conjectural or hypothetical.'" *Hall v. Time Inc.*, 158 Cal. App. 4th 847, 853, 70 Cal. Rptr. 3d 466, 470 (4th Dist. 2008).

⁵*Hall v. Time Inc.*, 158 Cal. App. 4th 847, 853, 70 Cal. Rptr. 3d 466,

Database owners may assert unfair competition claims against those who copy their databases or provide the tools for third parties to do so, provided the claim alleges more than mere copying, which otherwise would be preempted by the Copyright Act⁶ (or, less frequently, the Patent Act, which is separately addressed in section 5.04[3]). In *Marobie-FL, Inc. v. National Association of Fire Equipment Distributors*,⁷ for example, a federal court in Chicago entered summary judgment on plaintiff's claim for common law unfair competition against the owner of a website and its hosting company based on preemption, where the plaintiff had failed to allege likelihood of confusion or facts to support a finding of an "extra element" that "changes 'the nature of the action so it is qualitatively different from a copyright infringement claim'."⁸ The court noted in *dicta*, however, that while unfair competition claims premised on "passing off" generally are not barred by the Copyright Act, state law claims based on "reverse passing off" typically are preempted.⁹

State law unfair competition claims brought against interactive computer services or users based on content originating with others may be preempted by the Com-

470 (4th Dist. 2008); see also, e.g., *Kwikset Corp. v. Superior Ct.*, 51 Cal. 4th 310, 318, 120 Cal. Rptr. 3d 741 (2011) (finding that injury from frustration of patriotic desire to buy fully American-made products where the defendant falsely advertised that products were "Made in U.S.A." was sufficient to satisfy the standing requirement to state a claim under section 17204); see generally *infra* § 6.12[6] (analyzing this statute in greater detail).

⁶17 U.S.C.A. § 301; *Information Handling Services, Inc. v. LRP Publications, Inc.*, No. CIV. A. 00-1859, Copy. L. Rep. (CCH) P28,177, 2000 WL 1468535 (E.D. Pa. Sept. 20, 2000) (holding plaintiff's unfair competition claim to be preempted in a case alleging database copying). But see *BanxCorp v. Costco Wholesale Corp.*, 723 F. Supp. 2d 596, 611–18 (S.D.N.Y. 2010) (holding that claims for breach of a license agreement and misappropriation based on hot news were not preempted in a case alleging that the defendant, a licensee, misused money market and CD data, but claims for unfair competition and unjust enrichment were preempted); see generally *supra* § 4.18[1] (copyright preemption).

⁷*Marobie-FL, Inc. v. National Ass'n of Fire Equipment Distributors*, 983 F. Supp. 1167 (N.D. Ill. 1997).

⁸983 F. Supp. at 1167, quoting *Computer Associates Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 716 (2d Cir. 1992) (quoting an earlier case).

⁹983 F. Supp. at 1167, citing *FASA Corp. v. Playmates Toys, Inc.*, 869 F. Supp. 1334, 1361–64 (N.D. Ill. 1994); see generally *supra* § 4.18[1].

munications Decency Act (CDA).¹⁰ CDA preemption is analyzed extensively in section 37.05.

Unfair competition claims also potentially may be preempted by the Patent Act. “If a plaintiff bases its tort action on conduct that is protected or governed by federal patent law, then the plaintiff may not invoke the state law remedy, which must be preempted for conflict with federal patent law.”¹¹ The Patent Act preempts state law claims that “offer patent-like protection to intellectual property inconsistent with the federal scheme.”¹² To determine whether state law torts are in conflict with federal patent law and accordingly preempted, a court must assess a defendant’s allegedly tortious conduct:

If a plaintiff bases its tort action on conduct that is protected or governed by federal patent law, then the plaintiff may not invoke the state law remedy, which must be preempted for conflict with federal patent law. Conversely, if the conduct is not so protected or governed, then the remedy is not preempted. This approach, which considers whether a state

¹⁰47 U.S.C.A. § 230(c); *Small Justice LLC v. Xcentric Ventures LLC*, 873 F.3d 313, 322-23 (1st Cir. 2017) (affirming dismissal of certain aspects of plaintiff’s unfair competition based on the CDA); *Caraccioli v. Facebook, Inc.*, 700 F. App’x 588 (9th Cir. 2017) (affirming dismissal of plaintiff’s claims under California’s Unfair Competition Law and various tort theories “because the basis for each of these claims is Facebook’s role as a ‘republisher’ of material posted by a third party, and the claims are, therefore, barred by the Communications Decency Act.”); *Roca Labs, Inc. v. Consumer Opinion Corp.*, 140 F. Supp. 2d 1311, 1319-22 (M.D. Fla. 2015) (granting summary judgment for the defendant on plaintiff’s claim for allegedly violating the Florida Deceptive and Unfair Trade Practices Act (FDUTPA)); *Doe No. 1 v. Backpage.com, LLC*, 104 F. Supp. 3d 149, 162-64 (D. Mass. 2015) (dismissing plaintiff’s unfair competition claim as preempted by the CDA), *aff’d on other grounds*, 817 F.3d 12, 24-25 & n.8 (1st Cir. 2016) (expressing no opinion on the district court’s holding); *see generally infra* § 37.05[5][B] (analyzing CDA preemption in greater detail and citing other cases holding unfair competition claims preempted).

¹¹*Hunter Douglas, Inc. v. Harmonic Design, Inc.*, 153 F.3d 1318, 1335 (Fed. Cir. 1998), *cert. denied*, 525 U.S. 1143 (1999); *see also Sears, Roebuck & Co. v. Stiffel Co.*, 376 U.S. 225, 231 (1964) (holding that state law claims for unfair competition cannot be applied to “give protection of a kind that clashes with the objectives of the federal patent laws”); *Carson Optical, Inc. v. Prym Consumer USA, Inc.*, 11 F. Supp. 3d 317, 328-35 (E.D.N.Y. 2014) (holding that plaintiffs’ state law claims that defendants engaged in unfair competition by copying and reproducing plaintiff’s products were preempted by the Patent Act). *Hunter Douglas* was overruled in part on other grounds.

¹²*Dow Chemical Co. v. Exxon Corp.*, 139 F.3d 1470, 1475 (Fed. Cir. 1998), *cert. denied*, 525 U.S. 1138 (1999).

law tort, “as-applied,” conflicts with federal patent law, is consistent with that employed by the Supreme Court in cases involving preemption of state unfair competition law.¹³

Patent law will not preempt state law claims that “include additional elements not found in the federal patent law cause of action and . . . [that] are not an impermissible attempt to offer patent-like protection to a subject matter addressed by federal law.”¹⁴

In *Associated Press v. All Headline News Corp.*,¹⁵ the court denied defendants’ motion to dismiss AP’s state law unfair competition claim based on passing off, in a case where the plaintiff alleged that defendants copied AP breaking news reports and reprinted its news stories on their All Headline News (“AHN”) website, either as AP reports or AHN content. In allowing plaintiff’s state law unfair competition claim to proceed, the court held that the plaintiff had stated a claim by alleging that AHN passed off AP content as its own.¹⁶

Depending on the type of unfair competition claim asserted, it may be possible for a database owner to assert an equivalent claim under the Lanham Act.¹⁷

5.04[3] Patent Preemption of State Law Claims

Unlike the Copyright Act,¹ federal patent law will be deemed to preempt state law claims only in very narrow circumstances where state law presents an obstacle to the execution and accomplishment of patent laws or offers patent-like protection to intellectual property that is inconsistent with federal law.²

Whether a state law claim is preempted by the Patent Act

¹³*Hunter Douglas, Inc. v. Harmonic Design, Inc.*, 153 F.3d 1318, 1336 (Fed. Cir. 1998), *cert. denied*, 525 U.S. 1143 (1999).

¹⁴*Rodime PLC v. Seagate Tech., Inc.*, 174 F.3d 1294, 1306 (Fed. Cir. 1999), *cert. denied*, 528 U.S. 1115 (2000).

¹⁵*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454 (S.D.N.Y. 2009).

¹⁶608 F. Supp. 2d at 464.

¹⁷*See infra* § 5.08.

[Section 5.04[3]]

¹*See supra* §§ 5.04[1], 5.04[2] (copyright preemption in database cases); *see generally supra* § 4.18[1] (analyzing copyright preemption in greater detail).

²*See Hunter Douglas, Inc. v. Harmonic Design, Inc.*, 153 F.3d 1318, 1331-37 (Fed. Cir. 1998) (patent law does not provide for explicit preemp-

is a question governed by Federal Circuit law.³ Federal patent law preempts a state law claim that “offer[s] patent-like protection to intellectual property inconsistent with the federal scheme.”⁴ A claim will survive preemption if the plaintiff “plead[s] conduct in violation of [state law] that is separate and independent from its patent law claim.”⁵ To determine whether a state law tort claim is preempted by federal patent law, a court must “assess a defendant’s allegedly tortious conduct. If a plaintiff bases its tort action on conduct that is protected or governed by federal patent law, then the plaintiff may not invoke the state law remedy, which must be preempted for conflict with federal patent law.”⁶ Accordingly, the Federal Circuit instructs courts to consider “whether a state law tort, ‘as-applied,’ conflicts with federal patent law. . . .”⁷ State law claims are preempted if they fail to “include additional elements not found in the

tion but may create conflict preemption), *cert. denied*, 525 U.S. 1143 (1999); *Dow Chemical Co. v. Exxon Corp.*, 139 F.3d 1470, 1475 (Fed. Cir. 1998) (holding a state law claim not preempted), *cert. denied*, 525 U.S. 1138 (1999); *800 Adept, Inc. v. Murex Securities, Ltd.*, 539 F.3d 1354, 1369 (Fed. Cir. 2008) (“State tort claims against a patent holder, including tortious interference claims, based on enforcing a patent in the marketplace, are ‘preempted’ by federal patent laws, unless the claimant can show that the patent holder acted in ‘bad faith’ in the publication or enforcement of its patent.”); *see also Sears, Roebuck & Co. v. Stiffel Co.*, 376 U.S. 225, 231 (1964) (holding that a state law claim for unfair competition cannot be applied to “give protection of a kind that clashes with the objectives of the federal patent laws”).

³*Ultra-Precision Mfg. Ltd. v. Ford Motor Co.*, 411 F.3d 1369, 1376 (Fed. Cir. 2005).

⁴*Dow Chemical Co. v. Exxon Corp.*, 139 F.3d 1470, 1475 (Fed. Cir. 1998); *see also Carson Optical, Inc. v. Prym Consumer USA, Inc.*, 11 F. Supp. 3d 317, 328 (E.D.N.Y. 2014) (citing *Dow Chemical* in dismissing plaintiff’s unfair competition claim under New York law as preempted by the Patent Act).

⁵*Veto Pro Pac, LLC v. Custom Leathercraft Mfg. Co.*, Civil Action No. 3:08-cv-00302 (VLB), 2009 WL 276369, at *2 (D. Conn. Feb. 5, 2009) (holding that “the third count of the complaint, unjust enrichment, is completely preempted as it simply incorporates the two counts of patent infringement by reference and asserts that these also constitute unjust enrichment on the part of the defendants”).

⁶*Hunter Douglas, Inc. v. Harmonic Design, Inc.*, 153 F.3d 1318, 1336 (Fed. Cir. 1998), *overruled in part on other grounds, Midwest Ind. Inc. v. Karavan Trailers, Inc.*, 175 F.3d 1356, 1360–61 (Fed. Cir. 1999) (en banc).

⁷*Hunter Douglas, Inc. v. Harmonic Design, Inc.*, 153 F.3d 1318, 1336 (Fed. Cir. 1998), *overruled in part on other grounds, Midwest Ind. Inc. v. Karavan Trailers, Inc.*, 175 F.3d 1356, 1360–61 (Fed. Cir. 1999) (en banc).

federal patent law cause of action,” or if they are “an impermissible attempt to offer patent-like protection to subject matter addressed by federal law.”⁸

Given the scope of patent preemption, it is unlikely to arise in most database disputes.

5.05 Trespass and Conversion

5.05[1] Trespass to Chattels

Common law trespass potentially provides a basis for a database owner to exclude third parties from accessing its site or service, but in California a plaintiff must show harm in the form of diminishment of server capacity, which may be difficult today given that large commercial websites typically maintain ample capacity and that entities seeking to scrape data from a site often time their access to off-peak hours. Other states may impose less exacting damage requirements, although the damage shown generally must be to the chattel itself and not merely an injury to the business. Some states, however, will not even recognize a trespass claim involving intangibles. Even where recognized, a claim potentially may be preempted by the Copyright Act or precluded by the Uniform Trade Secrets Act.

The Restatement of Torts 2d provides that one who commits a trespass to a chattel is subject to liability to the possessor of the chattel if, but only if,

- (a) he dispossesses¹ the other of the chattel, or
- (b) the chattel is impaired as to its condition, quality, or value, or
- (c) the possessor is deprived of the use of the chattel for a substantial time, or

⁸*Rodime PLC v. Seagate Tech., Inc.*, 174 F.3d 1294, 1306 (Fed. Cir. 1999).

[Section 5.05[1]]

¹“A dispossession may be committed by intentionally (a) taking a chattel from the possession of another without the other’s consent, or . . . (c) barring the possessor’s access to a chattel.” Restatement (Second) of Torts § 221 (1965); *see also Ground Zero Museum Workshop v. Wilson*, 813 F. Supp. 2d 678, 697-98 (D. Md. 2011) (holding that the plaintiff could proceed to trial on its trespass to chattels claim based on dispossession where the defendant, a former web developer employee, took down plaintiff’s website and replaced it with an earlier version from 2007, holding that the issue of the defendant’s intent presented a jury question precluding summary judgment).

- (d) bodily harm is caused to the possessor, or harm is caused to some person or thing in which the possessor has a legally protected interest.²

In the 1990s, Internet service providers successfully used common law trespass to chattels to prevent spammers from directing unsolicited electronic mail messages to their servers and subscribers.³ In *CompuServe Inc. v. Cyber Promotions, Inc.*,⁴ for example, a federal court in Ohio preliminarily enjoined a bulk commercial emailer liable for trespass to chattels, where the defendant had directed spam emails to plaintiff's subscribers. The court found that defendant's intrusions into CompuServe's computer system resulted in CompuServe's customers receiving unwanted bulk email messages, causing many of them to terminate their accounts. The court held that this harm to plaintiff's business reputation and goodwill was actionable.⁵

In *eBay, Inc. v. Bidder's Edge, Inc.*,⁶ a federal court in San Jose extended this precedent to enjoin a competitor of eBay from repeatedly accessing and copying eBay's database through the use of bots (or intelligent agent software) where the court found that if injunctive relief was not granted "it would likely encourage other auction aggregators to crawl the eBay site, potentially to the point of denying effective access to eBay's customers . . . [T]here appears to be little doubt that the load on eBay's computer system would qualify as a substantial impairment of condition or value."⁷

By accessing eBay's site repeatedly without authorization—in violation of eBay policies—Judge Whyte concluded that Bidder's Edge committed a trespass, which could be enjoined to protect eBay from the lost server capacity caused

²Restatement (Second) of Torts § 218 (1965). The "tort recovery requires not only [a] wrongful act plus causation reaching to the plaintiff, but proof of some harm for which damages can reasonably be assessed." *Doe v. Chao*, 540 U.S. 614, 621 (2004).

³See *infra* § 29.04.

⁴*CompuServe Inc. v. Cyber Promotions, Inc.*, 962 F. Supp. 1015 (S.D. Ohio 1997).

⁵*CompuServe Inc. v. Cyber Promotions, Inc.*, 962 F. Supp. 1015, 1023 (S.D. Ohio 1997).

⁶*eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

⁷*eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058, 1071–72 (N.D. Cal. 2000).

by Bidder's Edge's repeated intrusions.⁸ He conceded, however, that there was "some uncertainty as to the precise level of possessory interference required to constitute intermeddling."⁹

Subsequently, in *Intel Corp. v. Hamidi*,¹⁰ the California Supreme Court held that a claim for trespass to chattels under California law may not be based on an electronic communication that neither damages the recipient's computer system nor impairs its functioning. In California, trespass requires a showing of intermeddling harmful to a materially valuable interest, rather than mere interference that does not amount to dispossession. Whereas eBay seemed to suggest that even the absence of actual damage could be actionable if a trespass occurred, the California Supreme Court disagreed with that broad a reading of the case, emphasizing that injunctive relief was justified in eBay based on the likely consequences of failing to enjoin future trespasses. Under California law, the court explained, intermeddling is actionable only if the "condition, quality, or value" of a chattel is impaired or "the possessor is deprived of the use of the chattel for a . . . time . . . so substantial that it is possible to estimate the loss caused thereby. A mere momentary or theoretical deprivation of use is not sufficient unless there is a dispossession."¹¹ Relief for trespass to chattels is appropriate, the California Supreme Court explained, where the unauthorized access to a computer server "actually did, or threatened to, interfere with the intended functioning of the system, as by significantly reducing its available memory and processing power."¹²

Following *Hamidi*, the court in *In re iPhone Application*

⁸Bidder's Edge's bots accessed eBay's site approximately 100,000 per day, accounting for as much as 1.53% of the total requests received by eBay and as much as 1.10% of the total data transferred by it over the web.

⁹The case settled while an appeal was pending before the Ninth Circuit. Bidder's Edge agreed to abide by the terms of the injunction entered by the district court and paid eBay an undisclosed amount of money. See Troy Wolverton, "eBay, Bidder's Edge end legal dispute," *cnet*, Mar. 1, 2001.

¹⁰*Intel Corp. v. Hamidi*, 30 Cal. 4th 1342, 1 Cal. Rptr. 3d 32 (2003). Hamidi involved bulk email transmissions sent to Intel by a disgruntled former employee. Unlike Bidder's Edge, it did not involve database access or copying.

¹¹30 Cal. 4th at 1357.

¹²30 Cal. 4th at 1356. The court reached this conclusion following a

*Litigation*¹³ dismissed plaintiffs' trespass claims with prejudice in a data privacy putative class action suit where two sets of plaintiffs alleged that (1) the creation of location history files and app software components "consumed portions of the cache and/or gigabytes of memory on their devices" and (2) apps had taken up valuable bandwidth and storage space on mobile devices and the defendants' conduct subsequently shortened the battery life of the device. In holding that plaintiffs had not met the standard set in *Hamidi*, Judge Lucy Koh of the Northern District of California explained that "[w]hile these allegations conceivably constitute a harm, they do not plausibly establish a significant reduction in service constituting an interference with the intended functioning of the system, which is necessary to establish a cause of action for trespass."¹⁴

In another data privacy case also brought under California law, *In re Google Android Consumer Privacy Litigation*,¹⁵ the court dismissed plaintiff's trespass to chattels claim because the alleged loss of CPU processing and battery capacity and Internet connectivity did not constitute a harm sufficient to establish a cause of action for trespass.

Likewise, in *Mount v. PulsePoint, Inc.*,¹⁶ a putative data privacy class action suit brought in federal court in New

discussion of *eBay* and *Ticketmaster Corp. v. Tickets.com, Inc.*, CV 99-7654 HLH (BQRx), 2000 WL 525390 (C.D. Cal. Mar. 27, 2000), *aff'd mem.*, Appeal No. 00-56574 (9th Cir. Jan. 2001). In *Tickets.com*, Judge Harry Hupp had distinguished *eBay, Inc.* because in the case before him he found insufficient evidence of "physical harm to the chattel . . . or some obstruction of its basic function." However, as underscored by the text of an initial tentative ruling that was subsequently withdrawn and replaced, Judge Hupp's analysis of *Ticketmaster's* trespass claim (as well as the contract claim discussed in section 5.03[2]) was heavily influenced by his concern that the data copied by Tickets.com was largely uncopyrightable, and that any protected content copied by Tickets.com without authorization amounted to a fair use intermediate copying (by analogy to reverse engineering). Judge Hupp wrote that "[t]he primary star in the copyright sky . . . is that purely factual information may not be copyrighted . . . Thus, unfair as it may seem . . . , the basic facts [Tickets.com] gathers and publishes cannot be protected from copying."

¹³*In re iPhone Application Litig.*, 844 F. Supp. 2d 1040 (N.D. Cal. 2012).

¹⁴*In re iPhone Application Litig.*, 844 F. Supp. 2d 1040, 1069 (N.D. Cal. 2012).

¹⁵*In re Google Android Consumer Privacy Litig.*, No. 11-MD-02264, 2013 WL 1283236, at *13 (N.D. Cal. Mar. 26, 2013).

¹⁶*Mount v. PulsePoint, Inc.*, 13 Civ. 6592 (NRB), 2016 WL 5080131, at

York under New York law over the defendant's alleged use of tracking cookies, the court dismissed plaintiff's trespass claim. Judge Naomi Reice Buchwald of the Southern District of New York wrote that "[t]o establish trespass to chattels, plaintiffs must show that PulsePoint intentionally, and without justification or consent, physically interfered with the use and enjoyment of personal property in their possession, and that they were harmed thereby."¹⁷ Citing *Intel Corp. v. Hamidi*¹⁸ and other cases that she characterized as applying the Restatement (Second) of Torts standard, Judge Buchwald explained that:

Possessors of chattel, unlike possessors of land, are not protected from "harmless intermeddlings." Restatement (Second) of Torts § 218 cmt. e (1965). There must be a resulting harm to "the possessor's materially valuable interest in the physical condition, quality, or value of the chattel," or else the possessor must be "deprived of the use of the chattel for a substantial time" or have some other legally protected interest in the property affected. *Id.*; see *Kuprewicz*, 3 Misc. 3d at 281, 771 N.Y.S.2d at 807-08 (adopting Restatement standard). For this reason, as applied to the online context, trespass "does not encompass . . . an electronic communication that neither damages the recipient computer system nor impairs its functioning." *Intel Corp. v. Hamidi*, 30 Cal. 4th 1342, 1347, 71 P.3d 296, 300 (2003); see *id.* at 1356, 71 P.3d at 306 ("In the decisions so far reviewed, the defendant's use of the plaintiff's computer system was held sufficient to support an action for trespass when it actually did, or threatened to, interfere with the intended functioning of the system, as by significantly reducing its available memory and processing power.").¹⁹

In the case before her she held that plaintiffs had not alleged the necessary harm to sustain their trespass claim by alleging, at most, "some unspecified increase in the use of device storage or processing capacity, without alleging that this uptick was significant or caused any discernible effect

*9-10 (S.D.N.Y. Aug. 17, 2016), *aff'd on other grounds*, 684 F. App'x 32 (2d Cir. 2017).

¹⁷*Mount v. PulsePoint, Inc.*, 13 Civ. 6592 (NRB), 2016 WL 5080131, at *9 (S.D.N.Y. Aug. 17, 2016) (citing *Sch. of Visual Arts v. Kuprewicz*, 3 Misc. 3d 278, 281, 771 N.Y.S.2d 804, 807 (Sup. Ct. N.Y. Cty. 2003); *Chevron Corp. v. Donziger*, 871 F. Supp. 2d 229, 258 (S.D.N.Y. 2012)), *aff'd on other grounds*, 684 F. App'x 32 (2d Cir. 2017).

¹⁸*Intel Corp. v. Hamidi*, 30 Cal. 4th 1342, 1 Cal. Rptr. 3d 32 (2003).

¹⁹*Mount v. PulsePoint, Inc.*, 13 Civ. 6592 (NRB), 2016 WL 5080131, at *9 (S.D.N.Y. Aug. 17, 2016), *aff'd on other grounds*, 684 F. App'x 32 (2d Cir. 2017).

on the operation of their devices.”²⁰ Judge Buchwald also rejected the argument that deprivation of the use of Safari’s third-party cookie blocker constituted sufficient harm, finding no authority for the proposition that “one feature of a particular software application” may be viewed as a chattel.²¹

By contrast, in *Sotelo v. DirectRevenue, LLC*,²² a federal court in Illinois held that the plaintiff stated a claim for trespass under Illinois law where it alleged that the defendant’s spyware, which was either directly downloaded from the plaintiff or bundled with software obtained from a third party, “interfered with and damaged his personal property, namely his computer and his Internet connection, by over-burdening their resources and diminishing their functioning.” In that case, the plaintiff had alleged that defendant’s spyware “bombarded” users’ computers with

²⁰*Mount v. PulsePoint, Inc.*, 13 Civ. 6592 (NRB), 2016 WL 5080131, at *9 (S.D.N.Y. Aug. 17, 2016), *aff’d on other grounds*, 684 F. App’x 32 (2d Cir. 2017). In so ruling, Judge Buchwald distinguished the Second Circuit’s ruling in *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393 (2d Cir. 2004), which is discussed later in this section, and a New York state trial court case, *Sch. of Visual Arts v. Kuprewicz*, 3 Misc. 3d 278, 771 N.Y.S.2d 804 (Sup. Ct. N.Y. Cty. 2003), as a case involving significant harm:

In *Register.com, Inc. v. Verio, Inc.*, the Court relied in part on the district court’s finding that the defendant’s use of search robots “consumed a significant portion of the capacity of [plaintiffs] computer systems,” 356 F.3d 393, 404-05 (2d Cir. 2004), and in *Kuprewicz*, the defendant had allegedly sent “large volumes” of unwanted e-mails which “depleted hard disk space, drained processing power, and adversely affected other system resources on [plaintiffs] computer system,” 3 Misc. 3d at 281-82, 771 N.Y.S.2d at 808 (internal quotation marks omitted). Those cases, unlike this one, involved allegations or findings of activity that either had or threatened to have a significant effect on the capacity of computer systems.

Mount v. PulsePoint, Inc., 13 Civ. 6592 (NRB), 2016 WL 5080131, at *10 (S.D.N.Y. Aug. 17, 2016), *aff’d on other grounds*, 684 F. App’x 32 (2d Cir. 2017).

²¹*Mount v. PulsePoint, Inc.*, 13 Civ. 6592 (NRB), 2016 WL 5080131, at *10 (S.D.N.Y. Aug. 17, 2016), *aff’d on other grounds*, 684 F. App’x 32 (2d Cir. 2017). She explained:

Many harmless electronic intrusions could potentially be recast as deprivations of a particular feature of an application meant to keep the electronic communication out. For example, the circumvention of a spam filter by junk e-mail could be characterized as depriving the user of his or her spam filter even if the junk e-mail had no effect whatever on the functionality of the user’s e-mail service. We think such a holding would upset the principle that no action for trespass lies for harmless intermeddlings with chattel.

Id.

²²*Sotelo v. DirectRevenue, LLC*, 384 F. Supp. 2d 1219, 1229-33 (N.D. Ill. 2005).

pop-up advertisements that obscured the web page a user was viewing and “destroy[ed] other software on a computer.” Plaintiff also alleged that the spyware and resource-consuming advertisements sent to a computer by the spyware caused computers to slow down, use up the bandwidth of the user’s Internet connection, incur increased Internet-use charges, deplete a computer’s memory, utilize pixels and screen-space on monitors, require more energy because slowed computers must be kept on for longer, and reduce a user’s productivity while increasing their frustration.²³

In *Craigslist Inc. v. 3Taps Inc.*,²⁴ a federal court in California likewise denied a defendant’s motion to dismiss where plaintiff plausibly alleged that defendant’s use of its website “could divert sufficient computing and communications resources to impair the website’s and servers’ functionality.”²⁵ In that case, Craigslist also had alleged that one of the defendants “boasts that it mass copies tens of millions of postings from craigslist in ‘real time.’”²⁶ The court conceded, however, that these allegations would need to be proven for the plaintiff to actually establish liability.

In *Grace v. Apple Inc.*,²⁷ Judge Koh, who had authored the *iPhone* decision, denied Apple’s motion to dismiss plaintiffs’ common law trespass claim under California law where the plaintiffs alleged injury from Apple allegedly permanently disabling FaceTime on its iOS6 and earlier operating systems which allegedly substantially harmed the functioning of their iPhones and significantly impaired the devices’ condition, quality and value. Judge Koh distinguished her own earlier opinion in *In re iPhone Application Litigation*²⁸ as a case that merely alleged conduct that took up device memory and reduced battery life. Judge Koh also rejected

²³*Sotelo v. DirectRevenue, LLC*, 384 F. Supp. 2d 1219, 1230 (N.D. Ill. 2005).

²⁴*Craigslist Inc. v. 3Taps Inc.*, 942 F. Supp. 2d 962, 980-81 (N.D. Cal. 2013).

²⁵*Craigslist Inc. v. 3Taps Inc.*, 942 F. Supp. 2d 962, 981 (N.D. Cal. 2013).

²⁶*Craigslist Inc. v. 3Taps Inc.*, 942 F. Supp. 2d 962, 981 (N.D. Cal. 2013).

²⁷*Grace v. Apple Inc.*, Case No. 17-CV-00551, 2017 WL 3232464 (N.D. Cal. July 28, 2017).

²⁸*In re iPhone Application Litig.*, 844 F. Supp. 2d 1040 (N.D. Cal. 2012).

the argument that plaintiffs had incurred no injury because they could have upgraded to iOS7 for free because plaintiffs alleged that iOS7 on their older iPhone 4 devices caused slowness, system crashes, erratic behavior and loss of critical features. In so ruling, Judge Koh adopted Judge Alsup's formulation that injury in the context of electronic trespass is adequately alleged where the plaintiff pleads that the purported trespass (1) caused physical damage to personal property, (2) impaired the condition, quality or value of the personal property, or (3) deprived plaintiff of the use of personal property for a substantial time.²⁹

Similarly, in *In re Lenovo Adware Litigation*,³⁰ Judge Ronald Whyte of the same court held that the plaintiffs in a putative class action suit stated a claim for trespass based on the operation of Superfish software loaded on Lenovo computers, where plaintiffs alleged that the software, which was constantly running in the background, interfered substantially with their use of their computers, by decreasing battery life by as much as 55% and slowing down internet upload and download speeds.

By contrast, in *Fields v. Wise Media, LLC*,³¹ Judge William Alsup dismissed plaintiffs' trespass claim without leave to amend where plaintiffs had alleged that defendants interfered with their mobile phones by sending unsolicited text messages where plaintiffs could allege neither physical harm nor impairment to their phones as a result of the messages. Plaintiffs, Judge Alsup wrote, alleged "a financial injury that did not result from the physical damage or interference with their phones."³²

Similarly, in *In re Apple & ATTM Antitrust Litigation*,³³ the court held that the loss of use of a personal iPhone for a few days, before plaintiffs received free replacements, was

²⁹*Grace v. Apple Inc.*, Case No. 17-CV-00551, 2017 WL 3232464, at *11 (N.D. Cal. July 28, 2017), quoting *Fields v. Wise Media, LLC*, No. C 12-05160 WHA, 2013 WL 5340490, at *4 (N.D. Cal. Sept. 24, 2013).

³⁰*In re Lenovo Adware Litig.*, Case No. 15-md-02624-RMW, 2016 WL 6277245, at *8-9 (N.D. Cal. Oct. 27, 2016).

³¹*Fields v. Wise Media, LLC*, No. C 12-05160 WHA, 2013 WL 5340490, at *4-5 (N.D. Cal. Sept. 24, 2013).

³²*Fields v. Wise Media, LLC*, No. C 12-05160 WHA, 2013 WL 5340490, at *5 (N.D. Cal. Sept. 24, 2013).

³³*In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965 (N.D. Cal. July 8, 2010).

not a sufficient injury to establish Article III standing to maintain a claim for trespass to chattels based on harm allegedly caused by an update to the iOS operating system.³⁴ In so ruling, the court contrasted the “loss of commercial email servers in a large corporation for a ‘substantial’ or ‘measurable’ time . . . ,” which had been found sufficient harm to support a claim in *Intel v. Hamidi*.³⁵ The court also held that plaintiffs failed to introduce evidence to show loss based on third party software applications which allegedly had become inaccessible, where there was no evidence that the plaintiffs paid for the apps or that they had been lost due to the iOS software upgrade, as opposed to other reasons suggested in plaintiffs’ deposition testimony, such as user deletion or the interaction with another software application.³⁶

In the alternative, the court held that even if plaintiffs could establish standing, there was no evidence of intentional interference, which is a required element of a claim for trespass to chattels.³⁷ The court explained that there was no evidence that Apple intended to harm plaintiffs’ devices. In addition, because plaintiffs voluntarily downloaded the iOS upgrade, “[v]oluntary installation runs counter to the notion that the alleged act was a trespass”³⁸

In *Register.com, Inc. v. Verio, Inc.*,³⁹ the Second Circuit affirmed entry of a preliminary injunction under New York law based on trespass to chattels based on evidence that the plaintiff’s computer systems were valuable resources of finite capacity, unauthorized use of the systems depleted the capacity available to end-users, and unauthorized use created risks of congestion and overload that could have disrupted

³⁴*In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965, at *6 (N.D. Cal. July 8, 2010) (entering summary judgment for defendants).

³⁵*See Intel Corp. v. Hamidi*, 30 Cal. 4th 1342, 1352-53, 1 Cal. Rptr. 3d 32 (2003).

³⁶*In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965, at *7 (N.D. Cal. July 8, 2010).

³⁷*In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965, at *7 (N.D. Cal. July 8, 2010).

³⁸*In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965, at *7 (N.D. Cal. July 8, 2010).

³⁹*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 438-39 (2d Cir. 2004).

plaintiffs' operations.⁴⁰

The Second Circuit neither cited to *Hamidi* nor imposed as exacting a standard for injury under New York law. Rather than suggesting that a significant reduction in server capacity had to be shown or threatened, the Second Circuit affirmed the district court's reliance on *eBay* for the proposition that "any interference with an owner's use of a portion of its property causes injury to the owner."⁴¹ The Second Circuit explained that a trespass to chattel occurs under New York law when a party intentionally damages or interferes with the use of property belonging to another, where interference may be accomplished by "dispossessing another of the chattel" (which does not require a showing of actual damage) or "using or intermeddling with a chattel in the possession of another" which requires a showing of actual damage.⁴² In *Register.com*, the Second Circuit accepted the district court's findings that Verio's unauthorized use of software robots posed a risk to the integrity of Register.com's systems due to potential congestion and overload problems that were shown to pose risks that were "real and potentially disruptive of its operations"⁴³

In *Snap-on Business Solutions Inc. v. O'Neil & Associates*,

⁴⁰The court emphasized that "the chattels in question are Register.com's computer systems, and the alleged trespass is Verio's intentional, unauthorized consumption of the capacity of those systems to handle, process and respond to queries." *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 437 n.55 (2d Cir. 2004).

⁴¹*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 438 (2d Cir. 2004) (emphasis added).

⁴²*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 437 (2d Cir. 2004).

⁴³*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 438 (2d Cir. 2004). Although not discussed in the Second Circuit's opinion, the evidence of harm presented to the district court showed that as much as 2.3% of Register.com's system resources were diminished by Verio's use of bots. In addition, Verio conceded that its practices occupied some of Register.com's system capacity. Indeed, evidence showed that "Verio was aware that its robotic queries could slow the response times of the registrars' databases and even overload them" and that it contemplated using IP aliasing to make it more difficult for Register.com to identify (and presumably block) its attempts to access Register.com's servers. Register.com's position was also bolstered by Verio's contention that no limit need ever be placed on the number of companies permitted to harvest data from Register.com's computers, which Judge Jones found unreasonable. See *Register.com, Inc. v. Verio, Inc.*, 126 F. Supp. 2d 238 (S.D.N.Y. 2000), *aff'd*, 356 F.3d 393 (2d Cir. 2004).

Inc.,⁴⁴ a court in Ohio held that genuine issues of fact precluded summary judgment on a database owner's claim for trespass against a competitor that accessed its database to copy proprietary information ostensibly belonging to a client (with the client's permission), where its access caused the database owner's servers to crash and slow down run times, leading to customer complaints.

Snap-On was an electronic parts catalog provider to clients in the automotive and heavy equipment industries, including Mitsubishi Caterpillar Forklift. Snap-On customers such as Mitsubishi typically provided raw data, such as parts catalogs, to Snap-On, which in turn created a searchable database with links to data and images. At some time after June 2005, Mitsubishi requested a copy of its data in electronic format. Snap-On offered to sell Mitsubishi a copy with minimal enhancements or, for substantially more money, provide a copy with full enhancements (hot spots, links and photographs). Mitsubishi, which believed that it had already paid for the enhanced data pursuant to the terms of the license and website development agreements it entered into with Snap-On, refused. Thereafter, Mitsubishi began talks with O'Neil, a competitor of Snap-On, eventually agreeing by letter agreement dated October 2008 to provide an electronic parts catalogue and parts content management services to Mitsubishi, presumably in place of Snap-On. Mitsubishi, however, did not have an electronic copy of its data to provide O'Neil and concluded it would be too expensive to have O'Neil build its own database from Mitsubishi's paper records, as Snap-On had done. Accordingly, O'Neil suggested that it could use a scraper tool to retrieve the electronic data from Snap-On's servers.

With Mitsubishi's approval and permission, O'Neil used an automated tool that it had developed to access Snap-On's database, copy information stored on the system, and save it to O'Neil's database, where O'Neil could analyze and manipulate it. O'Neil's alleged objective was to extract raw data which it could then use on its own system. Because the scraper tool mimicked a user accessing Snap-On's password-protected website, Mitsubishi gave O'Neil approximately thirty existing logon credentials to avoid detection. O'Neil ran the scraper tool for three months beginning in February

⁴⁴*Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 678–80 (N.D. Ohio 2010).

2009. According to Snap-On, its website crashed in April and May 2009 because of “enormous spikes” in website traffic caused by O’Neil’s scraping sessions. In May 2009, Snap-On began blocking O’Neil’s IP addresses. After obtaining indemnification from Mitsubishi, O’Neil resumed scraping Snap-On’s servers using different IP addresses and randomizing the times when it accessed Snap-On’s servers (presumably to avoid detection and make it more difficult for Snap-On to block O’Neil). In July 2009, Snap-On filed suit.

In denying O’Neil’s motion for summary judgment, the court held that to state a claim for trespass under Ohio law, a plaintiff must show that it has a possessory interest in a chattel and that the defendant (1) dispossessed the plaintiff of the chattel; (2) impaired the chattel’s condition, quality or value; (3) deprived the plaintiff of the chattel’s use for a substantial time; or (4) caused bodily harm to the plaintiff or to some person or thing from which plaintiff had a legally protected interest.⁴⁵ The court found that Snap-On had presented evidence that O’Neil’s scraper program damaged Snap-On’s servers (impairing the servers’ condition, quality or value or depriving Snap-On of their use for a substantial time). It also held that Snap-On’s claim was not preempted by the Copyright Act. Although the parties did not dispute that Snap-On had not provided permission to O’Neil to access its servers, the court held that whether Mitsubishi was authorized to grant access was a disputed fact.

Snap-On eventually obtained a general jury verdict, although it is not clear whether the verdict was based on Snap-On’s claims for trespass, breach of contract (based on its EULA), copyright infringement or violations of the Computer Fraud and Abuse Act.⁴⁶

Snap-On provides a cautionary tale on what not to do when a contract dispute arises over ownership to content in a database. In that case, Mitsubishi had signed license and web development agreements that were favorable to Snap-On

⁴⁵708 F. Supp. 2d 669, 678–80, citing *CompuServe Inc. v. Cyber Promotions, Inc.*, 962 F. Supp. 1015, 1021–22 (S.D. Ohio 1997); see generally *infra* § 29.04[4] (analyzing *Cyber Promotions*).

⁴⁶See *Snap-On Business Solutions Inc. v. O’Neil & Associates, Inc.*, No. 5:09–CV–1547, 2010 WL 2650875 (N.D. Ohio July 2, 2010) (awarding costs but denying Snap-On’s request for an award of attorneys’ fees because under Ohio law contractual attorneys’ fee provisions are unenforceable as contrary to public policy because they are viewed as encouraging litigation).

and did not clearly allow it a copy of the electronic version of its data, making it difficult for Mitsubishi to ever change database hosts without incurring substantial costs.⁴⁷ Rather than negotiating a solution or seeking a declaratory judgment of its rights back in 2005 when the dispute over ownership to its data first arose, it retained a competitor in 2008 and spent significant time and money exercising self-help that eventually resulted in a judgment for Snap-On against O'Neil in 2010, for which Mitsubishi had agreed to provide indemnification to O'Neil. A case by Mitsubishi to obtain a copy of a database comprised of its own data would have been perceived differently by a judge or jury than a suit by the database company against a competitor that repeatedly screen scraped a database and undertook significant measures to avoid detection.

In *Jedson Engineering, Inc. v. Spirit Construction Services, Inc.*,⁴⁸ which was also decided under Ohio law, the court granted the plaintiff's motion for summary judgment on its claim of trespass to chattels where the defendant used a password to access a website maintained by the plaintiff for a particular construction project. In rebuffing the defendant's challenge that damages could not be shown, the court accepted the plaintiff's argument that by virtue of the defendant's trespass the plaintiff "suffered harm in terms of the diminished value of servers as a safe, secure location for project files."⁴⁹

Judge Barrett also rejected defendant's argument that the plaintiff could not maintain a claim for trespass where it did not have "possession" of its website, which was hosted by a third party. To state a cause of action for trespass to chattels under Ohio law, the court held that it was not necessary to establish that it was the owner of the property, merely that it had a superior right to possession.⁵⁰

The court also held that plaintiff's trespass claim was not

⁴⁷In the absence of an agreement, a database developer will own the rights to any software or other original creative expression, even if the underlying data is owned by the customer. See *infra* § 11.02[2]. Website development agreements and their provisions are separately analyzed in chapter 19.

⁴⁸*Jedson Engineering, Inc. v. Spirit Const. Services, Inc.*, 720 F. Supp. 2d 904 (S.D. Ohio 2010).

⁴⁹*Jedson Engineering, Inc. v. Spirit Const. Services, Inc.*, 720 F. Supp. 2d 904, 926 (S.D. Ohio 2010).

⁵⁰*Jedson Engineering, Inc. v. Spirit Const. Services, Inc.*, 720 F. Supp.

preempted by the Copyright Act.

In *A.V. v. iParadigms, LLC*,⁵¹ the court granted summary judgment on a plagiarism website's counterclaim for trespass against a student who falsely submitted a paper to a school where he was not enrolled. The site owner alleged that it had expended significant time and resources investigating and rectifying the student's unauthorized use of the system. The court, however, held that this evidence supported a claim for consequential damages but did not evidence impairment to the condition, quality or value of the chattel (in this case, the plagiarism website) or that the site owner incurred actual damages as a result of loss of use of the chattel, which was what was required to be shown under Virginia law.

In *Inventory Locator Service, LLC v. Partsbase, Inc.*,⁵² the court likewise dismissed the plaintiff's trespass claim based on defendant's alleged unlawful access to plaintiff's database where the plaintiff did not explicitly allege interference with its physical server, as opposed to unauthorized access to its database, and where Florida trespass law required that trespass to chattels involve movable personal property, not intangible property such as a database.⁵³

Similarly, in *Universal Tube & Rollform Equipment Corp. v. YouTube, Inc.*,⁵⁴ the court granted YouTube's motion to dismiss a trespass claim based on disruptions to plaintiff's *utube.com* website caused by intended visitors to YouTube mistakenly calling up plaintiff's *utube.com* website. First, the court held that trespass to chattels must be based on interference with a plaintiff's computer system, rather than its website or domain name. Under Ohio law, the court, wrote, a "chattel" is limited to property that is "visible,

2d 904, 926 (S.D. Ohio 2010).

⁵¹*A.V. v. iParadigms, LLC*, 544 F. Supp. 2d 473, 485 (E.D. Va. 2008), *aff'd in part and rev'd in part on other grounds*, 562 F.3d 630, 639 (4th Cir. 2009).

⁵²*Inventory Locator Service, LLC v. Partsbase, Inc.*, No. 02-2695 MA/V, 2005 WL 2179185 (W.D. Tenn. Sept. 6, 2005) (applying Florida law).

⁵³By contrast, the court held that the plaintiff had stated a claim for conversion where the plaintiff alleged that the defendant hacked into the database to obtain customer passwords, accessing the entire customer list, and making changes to the database that sabotaged plaintiff's customer relations, where Florida law recognized an action for conversion based on a wrongful taking over of intangible interests in a business.

⁵⁴*Universal Tube & Rollform Equipment Corp. v. YouTube, Inc.*, 504 F. Supp. 2d 260 (N.D. Ohio 2007).

tangible and moveable.”⁵⁵ Second, the plaintiff’s claim failed because YouTube did not make contact with the computers hosting plaintiff’s website. “[T]hose making contact with Universal’s website were thousands of mistaken visitors, but not YouTube itself.”⁵⁶ In other words, there is no claim for secondary liability for trespass to chattels.

While a database owner may seek to deter screen scraping and establish its potential entitlement to sue for trespass by including appropriate language in its Terms of Use providing that access is unauthorized,⁵⁷ *Register.com* underscores that notice that access is prohibited may simply be provided by a cease and desist letter or other means. Unauthorized access also may be communicated through the Robot Exclusion Standard discussed in *eBay*⁵⁸ or through other header information.

In *Mortensen v. Bresnan Communication, LLC*,⁵⁹ a court in Montana denied an ISP’s motion to dismiss Computer Fraud and Abuse Act⁶⁰ and trespass claims where the plaintiff alleged that the ISP had modified user computer settings, even as the court dismissed plaintiff’s ECPA and invasion of privacy claims based on the finding that the ISP provided notice to consumers in its Privacy Notice and Subscriber

⁵⁵*Universal Tube & Rollform Equipment Corp. v. YouTube, Inc.*, 504 F. Supp. 2d 260, 269 (N.D. Ohio 2007).

⁵⁶*Universal Tube & Rollform Equipment Corp. v. YouTube, Inc.*, 504 F. Supp. 2d 260, 269 (N.D. Ohio 2007).

⁵⁷See *infra* § 22.05[2][P] (discussing anti-trespass provisions that may be employed).

⁵⁸The Robot Exclusion Standard is a protocol that allows sites to use “robot exclusion headers” (which are messages that may be read and detected by computers that comply with the Standard) and “robots.txt.” files to define the extent to which robotic activity will be permitted on a site.

Courts have held that failing to provide notice of objection in a robots.txt file could support a defense of implied license to a claim of copyright infringement, at least for the limited purpose of allowing a search engine to cache content (although an implied license potentially could be revoked). See *Parker v. Yahoo!, Inc.*, No. 07-2757, 2008 WL 4410095, at *4 (E.D. Pa. Sept. 25, 2008); *Field v. Google Inc.*, 412 F. Supp. 2d 1106, 1115-16 (D. Nev. 2006); see generally *supra* § 4.05[7] (analyzing implied licenses under copyright law and discussing these cases).

⁵⁹*Mortensen v. Bresnan Communication, LLC*, No. CV 10-13-BLG-RFC, 2010 WL 5140454 (D. Mont. Nov. 15, 2010), *vacated on other grounds*, 722 F.3d 1151 (9th Cir. 2013).

⁶⁰18 U.S.C.A. § 1030; *infra* § 5.06.

Agreement that their electronic transmissions might be monitored and would in fact be transferred to third parties, and also provided specific notice via a link on its website of its use of the NebuAd Appliance to transfer data to NebuAd (and of subscribers' right to opt out of the data transfer (via a link in that notice). The court concluded that altering privacy settings and security controls was outside the scope of the access permitted by the ISP's Privacy Notice and Subscriber Agreement to constitute trespass under Montana law. The court, however, relied on pre-*Hamidi* California law and therefore did not consider whether plaintiff had alleged impairment to its computer, as opposed to merely unauthorized access. The *Mortensen* court's ruling subsequently was vacated on other grounds, based on the district court's earlier denial of the ISP's motion to compel arbitration.⁶¹

Where a claim of trespass may be asserted, it generally will not be preempted by the Copyright Act.⁶² If it is based on copying information without an extra element, however, it will be preempted.⁶³

Where a trespass claim is premised on the acquisition of data, it also may be preempted by the Uniform Trade Secrets Act, depending on the applicable state law. Section 7 of the UTSA provides that the Act "displaces conflicting tort, restitutionary, and other law of this State pertaining to civil liability for misappropriation of a trade secret."⁶⁴ Section 7 has been construed to preempt trespass claims premised on

⁶¹See *Mortensen v. Bresnan Communications, LLC*, 722 F.3d 1151 (9th Cir. 2013).

⁶²See, e.g., *Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 678–80 (N.D. Ohio 2010) (holding plaintiff's trespass claim not preempted where plaintiff alleged trespass to its physical computer servers, not merely interference with possessory rights); *eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058, 1072 (N.D. Cal. 2000) ("The right to exclude others from using physical personal property is not equivalent to any rights protected by copyright and therefore constitutes an extra element that makes trespass qualitatively different from a copyright infringement claim.").

⁶³See, e.g., *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey*, 497 F. Supp. 2d 627, 650 (E.D. Pa. 2007) (holding plaintiff's trespass claim preempted and granting summary judgment in favor of the defendants where plaintiff's claim was based on defendants making allegedly unauthorized copies of archived website screen shots and website content stored on the Wayback Machine (www.archive.org), not plaintiff's servers); see generally *supra* § 4.18[1] (analyzing copyright preemption).

⁶⁴UTSA § 7; *infra* § 10.17 (addressing UTSA preemption). A copy of

the wrongful taking and use of confidential business and proprietary information, even in cases where the information at issue may not constitute a trade secret (at least in some jurisdictions).⁶⁵

Whether a claim is viable ultimately may turn on the extent of harm incurred to servers or tangible property (and not simply business information) and the extent of harm required under applicable state law for a potential claim to be deemed actionable, assuming applicable state law allows for a cause of action for trespass to intangibles.

In addition to common law trespass, some states have enacted specific computer trespass statutes, which are addressed in section 5.06 in conjunction with an analysis of the federal Computer Fraud and Abuse Act. The CFAA provides a federal remedy for computer trespass, where the specific elements of the statute may be satisfied.

5.05[2] Conversion

Conversion typically was not a viable claim for protecting the contents of a database because unlike a claim for trespass to chattels, which may be maintained where there is unauthorized access (plus damage), conversion usually requires a showing of dispossession or at least substantial interference.¹ Under California law, for example, conversion generally is defined as “the wrongful exercise of dominion

the UTSA is reprinted in the appendix to chapter 10.

⁶⁵See, e.g., *Synopsys, Inc. v. Ubiquiti Networks, Inc.*, 313 F. Supp. 3d 1056, 1080 (N.D. Cal. 2018) (dismissing plaintiff’s California common law trespass claim on alternative grounds; “To the extent defendants base their trespass claims on the accessing of their systems by the anti-piracy software, they must *but have not* alleged facts showing that access *impaired* the intended functioning of defendants’ systems. And to the extent that defendants base their trespass claim on the anti-piracy software’s securing of and use of defendants’ data, that common law claim would be preempted by CUTSA.”), citing *Heller v. Cepia, LLC*, No. C 11–01146 JSW, 2012 WL 13572, at *7 (N.D. Cal., Jan. 4, 2012) (common law claims, including trespass, “premised on the wrongful taking and use of confidential business and proprietary information, regardless of whether such information constitutes trade secrets, are superseded by the CUTSA.”); see generally *infra* § 10.17 (discussing conflicting lines of cases on whether a claim is preempted even if it is based on information that may not be protectable as a trade secret).

[Section 5.05[2]]

¹See *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 437–38 (2d Cir. 2004) (“Traditionally, courts have drawn a distinction between interfer-

over the personal property of another.”² Similarly, under Utah law, conversion is “an act of wilful interference with a chattel, done without lawful justification by which the person entitled thereto is deprived of its use and possession.”³ Likewise, under North Carolina law, conversion requires a showing of ownership by the plaintiff and wrongful possession or conversion by the defendant.⁴ When digital information is accessed without authorization, it is usually copied

ence by dispossession, . . . which does not require a showing of actual damages, . . . and interference by unauthorized use or intermeddling, . . . which requires a showing of actual damages”; citations omitted) (New York law); *eBay, Inc. v. Bidder’s Edge, Inc.*, 100 F. Supp. 2d 1058, 1067 (N.D. Cal. 2000) (distinguishing trespass from conversion). But see *CompuServe Inc. v. Cyber Promotions, Inc.*, 962 F. Supp. 1015, 1022 (S.D. Ohio 1997) (suggesting in *dicta* that less than complete dispossession may be sufficient under Ohio law; illegal use or misuse or wrongful detention of the property would be sufficient to show conversion); Restatement (Second) of Torts §§ 217 to 220.

On similar grounds, courts have generally declined to find that software can be converted. See, e.g., *Rich Media Club, LLC v. Mentchoukov*, No. 2:11-CV-1202 TS, 2012 WL 1119505, at *4 (D. Utah Apr. 3, 2012) (denying a conversion of intellectual property claim in part because the plaintiff was not “deprived of the use of any of its source code, software, or access codes”); *Ho v. Taflove*, 696 F. Supp. 2d 950, 957 (N.D. Ill. 2010) (granting summary judgment for the defendant on plaintiff’s claim for conversion because taking a copy did not prevent the owner “from conducting, controlling, accessing, using, or publishing” their material), *aff’d on other grounds*, 648 F.3d 489, 502 (7th Cir. 2011) (holding that plaintiff’s conversion claim was preempted by the Copyright Act); *Tegg Corp. v. Beckstrom Elec. Co.*, 650 F. Supp. 2d 413, 432 (W.D. Pa. 2008) (“because it is intangible property, software is generally not subject to a conversion claim.”).

²See *CRS Recovery, Inc. v. Laxton*, 600 F.3d 1138, 1145 (9th Cir. 2010). Conversion generally requires a showing of (1) plaintiff’s possessory right or interest in the property and (2) defendant’s exercise of dominion over the property or interference with it “in derogation of plaintiff’s rights.” *Colavito v. New York Organ Donor Network, Inc.*, 8 N.Y.3d 43, 49–50, 827 N.Y.S.2d 96, 860 N.E.2d 713 (2006); see also *Fremont Indem. Co. v. Fremont General Corp.*, 148 Cal. App. 4th 97, 119, 55 Cal. Rptr. 3d 621 (2d Dist. 2007) (“The basic elements of the tort [of conversion] are (1) the plaintiff’s ownership or right to possession of personal property; (2) the defendant’s disposition of the property in a manner that is inconsistent with the plaintiff’s property rights; and (3) resulting damages.”).

³*Fibro Trust, Inc. v. Brahman Fin., Inc.*, 974 P.2d 288, 295–96 (Utah 1999).

⁴See *Spirax Sarco, Inc. v. SSI Engineering, Inc.*, 122 F. Supp. 3d 408, 445 (E.D.N.C. 2015) (dismissing plaintiff’s conversion claim in a case based on defendants’ allegedly improper access to, copying and deletion of plaintiff’s electronic records and trade secrets).

exactly without dispossessing the owner of the data.

By contrast, where data is actually taken or damaged, a claim for conversion may arise. For example, in *Inventory Locator Service, LLC v. Partsbase, Inc.*,⁵ the court held that the plaintiff had stated a claim for conversion where the plaintiff alleged that the defendant hacked into its database to obtain customer passwords, accessing its entire customer list, and made changes to the database that sabotaged plaintiff's customer relations, where Florida law recognized an action for conversion based on a wrongful taking-over of intangible interests in a business.

Courts are gradually becoming more amenable to conversion claims involving digital information to the extent that intangible interests have been taken (assuming such a conversion claim is not preempted). In *In re Easysaver Rewards Litigation*,⁶ for example, a court allowed a claim to go forward where plaintiffs alleged conversion based on the alleged misappropriation of their private financial information, which was then used by defendants to make allegedly unauthorized debits from their financial accounts.

Likewise, in *Teva Pharmaceuticals USA, Inc. v. Sandhu*,⁷ the court held that plaintiff could state a claim for conversion of trade secrets taken from the plaintiff's computer network by an employee, but could not state a claim against the two competitors she was working with. To state a claim for conversion of trade secrets under Pennsylvania law, the court held that a plaintiff must allege that: (1) it owns a trade secret; (2) the trade secret was communicated to the defendant within a confidential relationship; and (3) the defendant used the trade secret to the plaintiff's detriment.⁸ The court emphasized that intangible intellectual property could be converted under Pennsylvania law. In holding that plaintiff stated a claim against defendant Sandhu, the court explained that plaintiff alleged that Sandhu knowingly provided its trade secrets and other confidential materials to

⁵*Inventory Locator Service, LLC v. Partsbase, Inc.*, No. 02-2695 MA/V, 2005 WL 2179185 (W.D. Tenn. Sept. 6, 2005) (applying Florida law).

⁶*In re Easysaver Rewards Litig.*, 737 F. Supp. 2d 1159 (S.D. Cal. 2010).

⁷*Teva Pharmaceuticals USA, Inc. v. Sandhu*, 291 F. Supp. 3d 659, 680 (E.D. Pa. 2018).

⁸*Teva Pharmaceuticals USA, Inc. v. Sandhu*, 291 F. Supp. 3d 659, 680 (E.D. Pa. 2018).

Desai and Apotex, who used the trade secrets to compete with Teva, to its detriment, which was sufficient to state a claim. By contrast, the court held that Teva could not state a claim against Desai and Apotex because neither was in a confidential relationship with Teva.

Courts also may recognize conversion claims arising out of dispossession of a domain name,⁹ at least in jurisdictions that recognize domain names as intangible property.¹⁰ Personal information, however, generally has been found not to be property that can be converted.¹¹

Among other defenses to conversion, a defendant may assert abandonment, which generally requires a clear, unequivocal and decisive act demonstrating a waiver of the plaintiff's property rights.¹²

A conversion claim also may not lie when premised on breach of a contract,¹³ such as a TOU agreement or a Privacy Policy.

⁹See *CRS Recovery, Inc. v. Laxton*, 600 F.3d 1138 (9th Cir. 2010) (suit by a domain name registrant who allowed the mat.net registration to lapse, which then enabled the new registrant, Li Qiang, to create an email address at mat.net that matched the email address of the administrative contact for rl.com, a different registration, which Li then transferred to his own name and sold to someone in India who in turn sold it to the defendant).

¹⁰See *infra* § 7.23 (analyzing domain names as property).

¹¹See, e.g., *Low v. LinkedIn Corp.*, 900 F. Supp. 2d 1010, 1030–31 (N.D. Cal. 2012) (dismissing with prejudice plaintiffs' claim for conversion because personal information does not constitute property under California law, plaintiffs could not establish damages and some of the information allegedly "converted," such as a LinkedIn user ID number, was generated by LinkedIn, and therefore not property over which a plaintiff could claim exclusivity); *In re iPhone Application Litig.*, 844 F. Supp. 2d 1040, 1074–75 (N.D. Cal. 2012) (dismissing with prejudice plaintiffs' conversion claim because personal information does not constitute property under California law, plaintiffs failed to establish that "the broad category of information referred to as 'personal information' is an interest capable of precise definition" and the court could not conceive how "the broad category of information referred to as 'personal information' . . . is capable of exclusive possession or control.").

¹²See, e.g., *CRS Recovery, Inc. v. Laxton*, 600 F.3d 1138, 1146 (9th Cir. 2010).

¹³See, e.g., *AD Rendon Commc'n, Inc. v. Lumina Americas, Inc.*, No. 04-CV-8832 (KMK), 2007 WL 2962591, at *4 (S.D.N.Y. Oct. 10, 2007) ("[E]ven if a plaintiff meets all of the elements of a conversion claim, the claim will still be dismissed if it is duplicative of a breach of contract claim."), citing *Wechsler v. Hunt Health Systems, Ltd.*, 330 F. Supp. 2d

Although unlikely to arise in a database or screen scraping case, California law also recognizes a defense where an acquirer of allegedly converted property was an innocent purchaser for value.¹⁴

Where a claim for conversion is based on the acquisition of data, it may be preempted by the Uniform Trade Secrets Act in states that have enacted section 7 of the UTSA. That section provides that the Act “displaces conflicting tort, restitutionary, and other law of this State pertaining to civil liability for misappropriation of a trade secret.”¹⁵ Section 7 has been construed in some, but not all, jurisdictions to preempt conversion claims premised on the wrongful taking and use of confidential business and proprietary information, regardless of whether the information constitutes a trade secret.¹⁶

Where a claim for conversion may be stated, it will usually not be preempted by the Copyright Act because a claim for conversion presupposes dispossession or substantial interfer-

383, 431 (S.D.N.Y. 2004) and *Richbell Information Services, Inc. v. Jupiter Partners, L.P.*, 309 A.D.2d 288, 765 N.Y.S.2d 575, 590 (1st Dep’t 2003); *AJW Partners LLC v. Itronics Inc.*, 68 A.D.3d 567, 568, 892 N.Y.S.2d 46 (1st Dep’t 2009) (holding that conversion claim was properly dismissed as duplicative of the breach of contract claim because it based on same alleged violation of the parties’ agreement).

¹⁴See *CRS Recovery, Inc. v. Laxton*, 600 F.3d 1138, 1145 (9th Cir. 2010). California law distinguishes between a purchaser whose vendor obtained title by fraud (which renders title merely voidable) and a purchaser whose vendor obtained title by theft (which is void). An innocent purchase for value without notice (actual or constructive) that his vendor has secured the goods by fraudulent purchase is not liable for conversion. Where property is stolen, it is not possible to acquire title under California law and the purchaser may be held liable for conversion. *CRS Recovery, Inc. v. Laxton*, 600 F.3d 1138, 1146 (9th Cir. 2010) (citing California state cases).

¹⁵UTSA § 7; *infra* § 10.17 (addressing UTSA preemption). A copy of the UTSA is reprinted in the appendix to chapter 10.

¹⁶See, e.g., *Synopsys, Inc. v. Ubiquiti Networks, Inc.*, 313 F. Supp. 3d 1056, 1080 (N.D. Cal. 2018) (dismissing plaintiff’s California common law conversion claim; “As to conversion, ‘if the only property identified in the complaint is confidential or proprietary information, and the only basis for any property right is trade secrets law, then a conversion claim predicated on the theft of that property is preempted’ by CUTSA.”), quoting *Avago Technologies U.S. Inc. v. Nanoprecision Products, Inc.*, Case No. 16-cv-03737-JCS, 2017 WL 412524, at *7 (N.D. Cal. Jan. 31, 2017); see generally *infra* § 10.17 (discussing conflicting lines of cases on whether a claim is preempted when based on information that may not be protectable as a trade secret).

ence, which would be an extra element beyond mere copying.¹⁷ Some courts, however, skip over the thornier issue of dispossession where it is clear that the claim is based solely on copying and is preempted.¹⁸

A conversion claim asserted against an interactive computer service (or user) for the misconduct of a different person or entity also may be preempted by the CDA.¹⁹

In most instances, conversion will not provide protection where the contents of a database are merely copied.

5.06 Computer Fraud and Abuse Act

The Computer Fraud and Abuse Act (CFAA),¹ a criminal statute, provides a trespass-like civil remedy under federal law when a third party accesses a database, website, or other

¹⁷See, e.g., *Opperman v. Path, Inc.*, 84 F. Supp. 3d 962, 971-72, 989 (N.D. Cal. 2015) (holding plaintiff's conversion claim to not be preempted where plaintiffs alleged that the defendants had preloaded their devices with apps that allowed them to access plaintiffs' electronic address books and disseminate information from these files to third parties without plaintiffs' knowledge or authorization, because, in addition to reproduction, plaintiff alleged unauthorized access, transmission, misuse and misappropriation of the data); *Internet Archive v. Shell*, 505 F. Supp. 2d 755, 763-64, (D. Colo. 2007) (holding that breach of contract and conversion claims arising out of a site owner's objection to her site being copied for inclusion in the Internet Archive's Wayback machine were not preempted). But see *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey*, 497 F. Supp. 2d 627, 650 (E.D. Pa. 2007) (holding plaintiff's conversion claim preempted and granting summary judgment in favor of the defendants where plaintiff's claim was based on defendants making allegedly unauthorized copies of archived website screen shots and website content stored on the Wayback Machine (www.archive.org)); see generally *supra* § 4.18[1] (analyzing copyright preemption).

¹⁸See, e.g., *Ho v. Taflove*, 648 F.3d 489, 502 (7th Cir. 2011) (holding that plaintiff's conversion claim was preempted by the Copyright Act); *Phantomalert, Inc. v. Google Inc.*, No. 15-CV-03986-JCS, 2016 WL 879758, at *12-14 (N.D. Cal. Mar. 8, 2016) (dismissing as preempted plaintiff's conversion claim, arising out of alleged copying of certain elements of plaintiff's program in defendants' Waze app); see generally *supra* § 4.18[1] (analyzing copyright preemption).

¹⁹See 47 U.S.C.A. § 230(c)(1); see also, e.g., *Franklin v. X Gear 101, LLC*, 17 Civ. 6452 (GBD) (GWG), 2018 WL 3528731, at *19 (S.D.N.Y. July 23, 2018) (dismissing claims for unjust enrichment and conversion against Instagram and GoDaddy as barred by the CDA); see generally *infra* § 37.05 (analyzing the scope of CDA preemption).

[Section 5.06]

¹18 U.S.C.A. § 1030; see generally *infra* § 44.08.

protected computer, without permission or exceeds authorized access. The CFAA “is primarily a criminal anti-hacking statute.”² It prohibits a number of different specific acts of misconduct involving access to protected computers (and mobile phones or other computerized devices³). “The statute . . . provides two ways of committing the crime of improperly accessing a protected computer: (1) obtaining access without authorization; and (2) obtaining access with authorization but then using that access improperly.”⁴

Using a scraper software program to systematically extract a company’s prices⁵ or other data⁶ (such as email addresses or customer listings)⁷ from a database or to otherwise gain unauthorized access to a website⁸ have been held actionable under the Act. The CFAA also may be violated by former or departing employees by, for example, continuing to access a company database after employment is terminated (constituting unauthorized access)⁹ or sabotaging the employer’s network shortly before leaving the company (while access

²*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1079 (7th Cir. 2016).

³The definition of *computer* pursuant to 18 U.S.C.A. § 1020(e)(1) is “exceedingly broad” and encompasses a mobile phone. *U.S. v. Kramer*, 631 F.3d 900, 902-03 (8th Cir. 2011); *see also U.S. v. Nosal*, 844 F.3d 1024, 1050 n.2 (9th Cir. 2016) (citing *Kramer* for this point in *dicta*).

⁴*Musacchio v. United States*, 136 S. Ct. 709, 713 (2016). As explained by the Fifth Circuit, “courts have interpreted ‘access without authorization’ as targeting outsiders who access victim systems, while ‘exceeds authorized access’ is applied to ‘insiders, such as employees of a victim company. . . . [The CFAA punishes] those who have no permission to access a system and those who have some permission to access but exceed it’” *U.S. v. Thomas*, 877 F.3d 591, 596 (5th Cir. 2017) (citations omitted).

⁵*See, e.g., EF Cultural Travel BV v. Explorica, Inc.*, 274 F.3d 577, 581–83 (1st Cir. 2001).

⁶*See Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1113 (C.D. Cal. 2007) (event and ticket sales information).

⁷*See Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1057 (N.D. Cal. 2010) (entering a default judgment under the CFAA).

⁸*See, e.g., CoStar Realty Information, Inc. v. Field*, 612 F. Supp. 2d 660 (D. Md. 2009) (access to a web-based database beyond what was authorized by the site’s user agreement); *I.M.S. Inquiry Management Systems, Ltd. v. Berkshire Information Systems, Inc.*, 307 F. Supp. 2d 521 (S.D.N.Y. 2004); *Shurgard Storage Centers, Inc. v. Safeguard Self Storage, Inc.*, 119 F. Supp. 2d 1121 (W.D. Wash. 2000).

⁹*See, e.g., Sell It Social, LLC v. Strauss*, 15 Civ. 970 (PKC), 2018 WL 2357261, at *3-4 (S.D.N.Y. Mar. 8, 2018) (denying summary judgment

was still authorized for performing routine work functions, but was exceeded by actions taken to disable remote access and damage the network).¹⁰

In contrast to a claim for common law trespass (at least in California), diminishment of server capacity need not specifically be shown.¹¹ However, to bring a claim under the CFAA a plaintiff generally must be able to show a minimum loss of \$5,000 from the defendant's conduct.¹² This dollar threshold has been an obstacle in some Internet cases where \$5,000 in actual losses cannot be shown.¹³ There is authority for the

where there was a dispute over when, and if, Strauss had been terminated and, in turn, whether he knowingly or intentionally accessed his employer's database without authorization, where there was at least some evidence that the defendant knew he had been fired before he accessed the database and that "Strauss retained login credentials solely because SIS neglected to remove him as an administrator on Listrak and not because Strauss continued to have permission to access the database.").

¹⁰See *U.S. v. Thomas*, 877 F.3d 591, 598-99 (5th Cir. 2017) (affirming the conviction of an IT employee who had authority to stop backups or delete files but did not have authority to put in place a series of harmful acts to disable remote access to the network and cause other harm after he left the company).

¹¹See *supra* § 5.05[1].

¹²See *infra* § 44.08 (specifically enumerating all of the alternative grounds for showing loss sufficient to maintain a civil CFAA claim and identifying potentially conflicting lines of authority).

¹³See, e.g., *Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 440 (2d Cir. 2004) (finding the plaintiff not likely to prevail on its CFAA claim arising out of the defendant exceeding authorized access to plaintiff's database); *In re Google Inc. Cookie Placement Consumer Privacy Litig.*, 806 F.3d 125, 148-49 (3d Cir. 2015) (affirming dismissal of defendants' motions to dismiss plaintiffs' CFAA claim for failure to allege the threshold loss of \$5,000 in a putative data privacy class action suit where plaintiffs could not allege any viable lost marketing opportunity for their data), *cert. denied*, 137 S. Ct. 36 (2016); *Bose v. Interclick, Inc.*, No. 10 Civ. 9183, 2011 WL 4343517 (S.D.N.Y. Aug. 17, 2011) (dismissing with prejudice a CFAA claim alleging general impairment to the value of plaintiff's computer in a putative behavioral advertising class action suit); *Lyons v. Coxcom, Inc.*, No. 08-CV-02047-H, 2009 WL 347285 (S.D. Cal. Feb. 6, 2009) (dismissing a CFAA claim where inadequate damage was alleged); *Pearl Investments, LLC v. Standard I/O, Inc.*, 257 F. Supp. 2d 326 (D. Me. 2003) (inability to quantify alleged loss to computer network); *Spec Simple, Inc. v. Designer Pages Online LLC*, 56 Misc. 3d 700, 54 N.Y.S.3d 837, 842-46 (N.Y. Cty. 2017) (dismissing a CFAA claim by the operator of a virtual library (online database) for paid subscribers in architectural, interior design, engineering, and facility management professions, brought against a competitor and the operator's former client, which had an ownership interest in the competitor, claiming that the client illicitly provided the operator's propri-

proposition that the \$5,000 threshold may be met by harm overall to a computer system and need not be suffered by just one computer during one particular intrusion.¹⁴ It also may include the costs associated with responding to the unauthorized intrusion.¹⁵ Case law addressing the \$5,000 threshold—and related requirements under the statute for showing damage or loss—is analyzed more extensively in section 44.08[1]. Where the \$5,000 threshold cannot be met, a claim may be asserted in some cases under equivalent state anti-trespass or computer crime laws, which are discussed briefly at the end of this section 5.06.

At least one court has held that the CFAA has extraterritorial effect and may provide a civil claim when a U.S. company is accused of scraping data from a foreign website.¹⁶

In *Ticketmaster LLC v. RMG Technologies, Inc.*,¹⁷ which was decided on motion for preliminary injunction, the court found Ticketmaster likely to prevail on the merits of its claim that the defendant violated the CFAA by accessing its website in violation of Ticketmaster's Terms of Use and, for commercial purposes, accessing its database thousands of times a day.

Similarly, in *Southwest Airlines Co. v. Farechase, Inc.*,¹⁸ the court in the Northern District of Texas denied the

etary information to the competitor, where the only damages alleged were for unfair competition and therefore plaintiff could not meet the \$5,000 threshold); *see generally infra* § 44.08.

¹⁴*See Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1113 (C.D. Cal. 2007); *see generally infra* § 44.08.

¹⁵18 U.S.C.A. § 1030(g), 1030(c)(4)(A)(i)(I), 1030(e)(11); *see, e.g.*, *A.V. v. iParadigms, LLC*, 562 F.3d 630, 645 (4th Cir. 2009) (reversing the district court's entry of summary judgment for the counterclaim defendant (A.V.) and remanding the case for further consideration, where the district court erroneously excluded from consideration the costs of investigation undertaken by iParadigms to determine how A.V. had gained access to its site); *see generally infra* § 44.08.

¹⁶*See Ryanair DAC v. Expedia Inc.*, Case No. C17-1789RSL, 2018 WL 3727599 (W.D. Wash. Aug. 6, 2018) (holding that Ryanair could maintain a CFAA suit against a U.S. website accused of scraping data from its website in Ireland, in violation of its Terms of Use and denying defendant's motion to dismiss for forum non conveniens or based on comity), *citing RJR Nabisco, Inc. v. European Cmty.*, 136 S. Ct. 2090, 2100 (2016) and *WesternGeco LLC v. ION Geophysical Corp.*, 138 S. Ct. 2129, 2136 (2018).

¹⁷*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007).

¹⁸*Southwest Airlines Co. v. Farechase, Inc.*, 318 F. Supp. 2d 435 (N.D.

defendant's motion to dismiss plaintiff's CFAA claim where Southwest alleged that the defendant accessed fare and scheduling information that Southwest Airlines published on its website, *southwest.com*, where Southwest directly informed the defendant that its access was unauthorized. Regardless of whether the Terms of Use formed a binding contract, the plaintiff alleged that the defendant was directly and repeatedly warned that Southwest prohibited "any deep-link, page-scrape, robot, spider or other automatic device, program, algorithm or methodology which does the same things."¹⁹

In *eBay, Inc. v. Digital Point Solutions, Inc.*,²⁰ the court denied defendants' motion to dismiss eBay's CFAA claim, arising out of the defendants' cookie-stuffing scheme. Defendants were accused of using software to direct users' browsers surreptitiously to the eBay site (without their knowledge), where a cookie would be deposited on their hard drive and plaintiffs in turn would earn commissions from advertising revenue. In denying the defendants' motion, the court found that their access to eBay site was unauthorized because they exceeded the scope of eBay's user agreement.

In *Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*,²¹ the court denied defendants' motion for summary judgment in a screen-scraping case where defendant O'Neil, a competitor of the plaintiff Snap-On, repeatedly accessed Snap-On's database (causing it to run slowly and on two occasions, crash) to copy data for Mitsubishi, a customer who was trying to transition from Snap-On's database hosting service to O'Neil, where the issue of whether Mitsubishi was authorized to allow O'Neil to access the database on its behalf was disputed. Snap-On subsequently obtained a general jury verdict, although it is not clear whether the verdict was based on Snap-On's claims for trespass, breach of contract (based on its EULA), copyright infringement or

Tex. 2004).

¹⁹*Southwest Airlines Co. v. Farechase, Inc.*, 318 F. Supp. 2d 435, 439–40 (N.D. Tex. 2004).

²⁰*eBay Inc. v. Digital Point Solutions, Inc.*, 608 F. Supp. 2d 1156 (N.D. Cal. 2009).

²¹*Snap-on Business Solutions Inc. v. O'Neil & Associates, Inc.*, 708 F. Supp. 2d 669, 676–78 (N.D. Ohio 2010).

violations of the Computer Fraud and Abuse Act.²² The case is discussed in greater detail in section 5.05 in connection with Snap-On's trespass claim.

In *Mortensen v. Bresnan Communication, LLC*,²³ a court in Montana denied an ISP's motion to dismiss CFAA and trespass claims where the plaintiff alleged the defendant had modified user computer settings, even as the court dismissed plaintiff's ECPA and invasion of privacy claims based on the finding that the ISP provided notice to consumers in its Privacy Notice and Subscriber Agreement that their electronic transmissions might be monitored and would in fact be transferred to third parties, and also provided specific notice via a link on its website of its use of the NebuAd Appliance to transfer data to NebuAd (and of subscribers' right to opt out of the data transfer (via a link included in that notice)). The court held that the defendant had been authorized by its Privacy Notice and Subscription Agreement to access plaintiff's computer, based on plaintiff's use of the service after having received notice that his use was subject to terms, but that authorized access had been allegedly exceeded by altering or tampering plaintiff's computer settings, which was not disclosed in plaintiff's Privacy Notice or Subscription Agreement.²⁴ The *Mortensen* court's ruling subsequently was vacated on other grounds, based on the district court's earlier denial of the ISP's motion to compel arbitration.²⁵

By contrast, merely accessing a publicly accessible

²²See *Snap-On Business Solutions Inc. v. O'Neil & Associates, Inc.*, No. 5:09-CV-1547, 2010 WL 2650875 (N.D. Ohio July 2, 2010) (awarding costs but denying Snap-On's request for an award of attorneys' fees because under Ohio law contractual attorneys' fee provisions are unenforceable as contrary to public policy because they are viewed as encouraging litigation).

²³*Mortensen v. Bresnan Communication, LLC*, No. CV 10-13-BLG-RFC, 2010 WL 5140454 (D. Mont. Nov. 15, 2010), *vacated on other grounds*, 722 F.3d 1151 (9th Cir. 2013).

²⁴The court gave only cursory consideration to whether the plaintiff could show \$5,000 in damages, assuming that the mere allegation of damage by a putative class that had not yet been certified would be sufficient. This aspect of the court's holding is inconsistent with the weight of authority. See *infra* §§ 26.15 (privacy class action suits), 44.08 (analyzing the Computer Fraud and Abuse Act in greater detail).

²⁵See *Mortensen v. Bresnan Communications, LLC*, 722 F.3d 1151, 1157-61 (9th Cir. 2013).

webpage does not constitute a CFAA violation.²⁶ Likewise, in limited circumstances, a journalist, researcher, or other person may claim a First Amendment right to scrape data from private websites (using bots and fictitious user profiles) and publish the results of their research, as a defense to a criminal CFAA prosecution, even where doing so violates a site's Terms of Service agreement.²⁷

In *Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*,²⁸ the Seventh Circuit affirmed the lower court's entry of summary judgment, holding that a data analytics company's use of a web-harvester to obtain county land records in bulk was not actionable under the CFAA. In the words of Judge Joel Martin Flaum, writing on behalf of himself and Judges Daniel A. Manion and Ilana Rovner, *Fidlar*, by the lawsuit "attempt[ed] to convert its failure to prohibit LPS's action by

²⁶See, e.g., *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailley*, 497 F. Supp. 2d 627, 646–49 (E.D. Pa. 2007) (granting summary judgment for the defendants in a case where plaintiffs sued the law firm that previously had represented an opposing party in a trademark infringement suit, alleging that defendants obtained archived copies of its website without authorization from the Wayback Machine, www.Archive.org, where the copies were only accessible because of a computer malfunction that caused the Archive.org site to ignore the Robots.txt files on plaintiff's site that would otherwise have resulted in the archived pages being made publicly inaccessible; "No evidence has been presented showing that the Harding firm exceeded authorized access. The facts do not show that the Harding firm did anything other than use the Wayback Machine in the manner it was intended to be used.").

²⁷See *Sandvig v. Sessions*, 315 F. Supp. 3d 1, 15–30 (D.D.C. 2018) (denying in part the government's motion to dismiss; holding that researchers planning to engage in audit testing of internet real estate, hiring, and other websites through the use of bots and fictitious profiles, for research purposes, plausibly alleged a First Amendment interest in doing so, plausibly alleged that they have standing to sue, and plausibly alleged that the CFAA's access provision violates the Free Speech and Free Press Clauses of the First Amendment as applied to them). In *Sandvig*, the court opined that:

Scraping or otherwise recording data from a site that is accessible to the public is merely a particular use of information that plaintiffs are entitled to see. The same goes for speaking about, or publishing documents using, publicly available data on the targeted websites. The use of bots or sock puppets is a more context-specific activity, but it is not covered in this case. Employing a bot to crawl a website or apply for jobs may run afoul of a website's ToS, but it does not constitute an access violation when the human who creates the bot is otherwise allowed to read and interact with that site.

Id. at 26–27.

²⁸*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075 (7th Cir. 2016).

contract into an allegation of criminal conduct.”²⁹

In that case, the appellate panel affirmed the lower court holding that no reasonable jury could find that LPS acted with intent to defraud³⁰ where the plaintiff alleged that the defendant had undertaken a fraudulent scheme to avoid paying printing fees by accessing plaintiff’s records through its own web harvester. In so ruling, the appellate panel emphasized that LPS had authority to “access the county records as a general matter” but the question presented was “whether the *way* in which it did so violated the statute.”³¹

The appellate panel found no basis to support plaintiff’s alleged scheme where LPS used its web-harvester even in those counties that did not charge a print fee, suggesting its goal was to accelerate data acquisition, not avoid fees. In addition, LPS continued to pay for unlimited subscriptions in all 82 counties, even though it was not logging any time by using its web-harvester. If LPS wanted to defraud the counties, the court reasoned, it could have selected a limited subscription for less money. Further, LPS did not conceal its use of the web-harvester, which was inconsistent with an intent to defraud. In addition, the evidence showed that

²⁹*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1084 (7th Cir. 2016).

³⁰18 U.S.C.A. § 1030(a)(4) punishes anyone who “knowingly and with intent to defraud, accesses a protected computer without authorization, or exceeds authorized access, and by means of such conduct furthers the intended fraud and obtains anything of value” *Intent to defraud* is not defined in the statute, but according to the Seventh Circuit means “that the defendant acted willfully and with specific intent to deceive or cheat, usually for the purpose of getting financial gain for himself or causing loss to another.” *Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1079 (7th Cir. 2016) (quoting earlier cases). Because direct evidence of intent is often unavailable, intent to defraud may be established by circumstantial evidence “and by inferences drawn from examining the scheme itself, which demonstrate that the scheme was reasonably calculated to deceive persons of ordinary prudence and comprehension.” *Id.* (quoting an earlier case; affirming the lower court’s entry of summary judgment in a civil case finding no intent to deceive where the defendant used a web-harvester to copy county land records in bulk, and was not expressly prohibited from doing so by contract).

The legislative history of section 1030(a)(4) reflects a Congressional intent to “reach cases of computer theft. . . . The intent to defraud element is meant to distinguish computer theft from mere trespass.” *Id.*, citing S. Rep. No. 99-432, at 9, reprinted in 1986 U.S.C.C.A.N. 2479, 2486-87.

³¹*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1080 (7th Cir. 2016) (emphasis in original).

Fidlar was aware of two other companies that used their own tools to access county records, which supported LPS's assertion that its intent was speed and efficiency, not to avoid fees. Moreover, an internal Fidlar email stated that "Fidlar *could* make screen-scraping or web-harvesting illegal with a 'simple disclaimer that states the information can't be scraped from the image'" but didn't do so, suggesting that Fidlar itself "did not believe that web-harvesting was impermissible."³² Finally, and significantly, the court noted that the agreements between LPS and the counties did not prohibit LPS from using a web-harvester or require LPS to access records exclusively through the plaintiff's program.

The Seventh Circuit also affirmed the lower court judgment that LPS did not violate section 1030(a)(5)(A), which punishes anyone who knowingly causes the transmission of a program, information, code or command, and as a result of such conduct, intentionally causes damage without authorization to a protected computer. Fidlar's claim, by contrast, the court wrote, was "trespassory in nature. LPS accessed the middle tier servers without following Fidlar's 'rules' (i.e., logging its activity or using the Laredo client)."³³ *Damage*, however, is defined as "any impairment to the integrity or availability of data, a program, a system or information . . . ,"³⁴ which the court explained contemplated destructive behavior, such as using a virus or destroying data, not merely circumventing the plaintiff's method for tracking user activity without altering any data or disrupting Fidlar's system in any way.³⁵

Finally the court affirmed summary judgment for LPS on Fidlar's claim under the Illinois Computer Crime Prevention Law,³⁶ which requires a showing that a defendant knew or had reason to know that its insertion or attempt to insert a program into a computer would cause loss. Because LPS believed that "it was entitled to download records without incurring a fee, it follows that LPS did not know or have rea-

³²*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1082 (7th Cir. 2016).

³³*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1085 (7th Cir. 2016).

³⁴18 U.S.C.A. § 1030(e)(8).

³⁵*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1084 (7th Cir. 2016).

³⁶720 ILCS § 5/17-51(a)(4)(C).

son to know that it was causing a loss.”³⁷

While a number of courts have held that accessing a website for purposes prohibited by website Terms of Use constitutes unauthorized access (or exceeds authorized access) under the CFAA in civil cases where a conscious violation has been shown,³⁸ in the Second, Fourth and Ninth Circuits a CFAA claim for *exceeding authorized access* may not be based on a defendant’s violating a contract or policy that imposes use, rather than access restrictions.³⁹ For example, a contract provision that allows access for particular uses—such as personal but not commercial use—would not be actionable in the Second, Fourth or Ninth Circuits,⁴⁰

³⁷*Fidlar Technologies v. LPS Real Estate Data Solutions, Inc.*, 810 F.3d 1075, 1085-86 (7th Cir. 2016).

³⁸See *infra* § 44.08 (enumerating cases).

³⁹See *U.S. v. Valle*, 807 F.3d 508, 524-28 (2d Cir. 2015); *U.S. v. Nosal*, 676 F.3d 854 (9th Cir. 2012) (*en banc*); *WEC Carolina Energy Solutions, LLC v. Miller*, 687 F.3d 199 (4th Cir. 2012), *cert. dismissed*, 568 U.S. 1079 (2013); see also *Teva Pharmaceuticals USA, Inc. v. Sandhu*, 291 F. Supp. 3d 659, 669-71 (E.D. Pa. 2018) (agreeing with the Fourth and Ninth Circuit’s narrow approach and holding that plaintiff failed to state a claim against an employee who was permitted to access plaintiff’s computer in the course of her employment to access information in its database, including the information allegedly shared with the other defendants); *Hedgeye Risk Management, LLC v. Heldman*, 271 F. Supp. 3d 181, 193-95 (D.D.C. 2017) (collecting cases and agreeing with the Second, Fourth and Ninth Circuits in holding that the prohibition on *exceeding authorized access* only applies to unauthorized access to information, not to unauthorized use of properly accessed material); *Tank Connection, LLC v. Haight*, 161 F. Supp. 3d 957, 969-70 (D. Kan. 2016) (granting summary judgment for a former employee who was accused of improperly accessing files on his employer’s network, where the employer mistakenly had not blocked access to the files as intended; “Case law makes clear that the relevant question is whether he was authorized to access the area or the information, not whether he did so with an improper purpose in mind.”); *Cloud-path Networks, Inc. v. SecureW2 B.V.*, 157 F. Supp. 3d 961, 983 (D. Colo. 2016) (agreeing “with Second, Fourth, and Ninth Circuits’ shared conclusion: ‘exceeds authorized access’ in the CFAA does not impose criminal liability on individuals who are authorized to access company data but do so for disloyal purposes; it applies only to individuals who are allowed to access a company computer and use that access to obtain data they are not allowed to see for any purpose.”); see generally *infra* § 44.08 (analyzing these cases in greater detail).

⁴⁰See *U.S. v. Valle*, 807 F.3d 508, 524-28 (2d Cir. 2015); *U.S. v. Nosal*, 676 F.3d 854 (9th Cir. 2012) (*en banc*); *WEC Carolina Energy Solutions, LLC v. Miller*, 687 F.3d 199 (4th Cir. 2012), *cert. dismissed*, 568 U.S. 1079 (2013); see generally *infra* § 44.08[1] (analyzing *Nosal* case law and the

but could be in other circuits.⁴¹ As explained by the Second Circuit, a person exceeds authorized access under the narrow view applied in that circuit “only when he obtains or alters information that he does not have authorization to access for any purpose which is located on a computer that he is otherwise authorized to access.”⁴² In courts applying the narrow view of what may constitute *exceeding authorized access*, an individual could not exceed authorized access, within

current circuit split).

In *Oracle America, Inc. v. Service Key, LLC*, No. C 12-00790, 2012 WL 6019580 (N.D. Cal. Dec. 3, 2012), for example, the court dismissed Oracle’s CFAA claim against a third party that accessed its website to provide software support services to third parties. Oracle had argued that the defendant was not allowed to access its website and therefore acted without authorization within the meaning of the statute, but the court ruled that Oracle’s claim was barred by *Nosal* because Oracle’s complaint alleged that the defendant was authorized to access its website, but not for the ostensibly improper purpose of using its authorized access to provide support services to third parties. *See id.* at *5.

⁴¹*See, e.g., EF Cultural Travel BV v. Explorica, Inc.*, 274 F.3d 577, 583–84 (1st Cir. 2001) (holding that an employee likely exceeded his authorized access when he used that access to disclose information in violation of a confidentiality agreement into which the employee voluntarily entered); *U.S. v. John*, 597 F.3d 263, 271 (5th Cir. 2010) (holding that an employee of Citigroup exceeded her authorized access when she accessed confidential customer information in violation of her employer’s computer use restrictions and used that information to commit fraud, writing that a violation occurs “at least when the user knows or reasonably should know that he or she is not authorized to access a computer and information obtainable from that access in furtherance of or to perpetrate a crime”); *International Airport Centers, LLC v. Citrin*, 440 F.3d 418, 420–21 (7th Cir. 2006) (reversing dismissal of a claim against an employee who accessed plaintiff’s network and caused transmission of a program that caused damage to a protected computer where the court held that an employee who had decided to quit and violate his employment agreement by destroying data breached his duty of loyalty to his employer and therefore terminated the agency relationship, making his conduct unauthorized (or exceeding authorized access)); *U.S. v. Teague*, 646 F.3d 1119, 1121–22 (8th Cir. 2011) (upholding the conviction under section 1030(a)(2) and 1030(c)(2)(A) of an employee of a government contractor who used his privileged access to a government database to obtain President Obama’s private student loan records); *U.S. v. Rodriguez*, 628 F.3d 1258, 1263 (11th Cir. 2011) (holding that a Social Security Administration employee exceeded authorized access by obtaining information about former girlfriends and potential paramours to send flowers to their houses, where the Administration told the defendant that he was not authorized to obtain personal information for nonbusiness reasons); *see generally infra* § 44.08[1] (analyzing case law).

⁴²*U.S. v. Valle*, 807 F.3d 508, 511 (2d Cir. 2015).

the meaning of the CFAA, by accessing a computer, “with an improper purpose . . . to obtain or alter information that he is otherwise authorized to access”⁴³ By contrast, accessing files on a network beyond those which an employee was authorized to view meets the statute’s definition of “exceeds authorized access” under the narrow view.⁴⁴

On similar grounds, a Ninth Circuit panel held that Terms of Use prohibitions on the use of bots, scrapers, and other automated means to access a site may not form the basis for claims under either California’s Computer Data Access and Fraud Act⁴⁵ or Nevada’s Computer Crimes Law,⁴⁶ because “taking data using a *method* prohibited by the applicable terms of use, when the taking itself generally is permitted, does not violate . . .” either state statute.⁴⁷

Even in courts that apply a narrow view of the scope of *exceeding authorized access* under the CFAA, it may be possible for a database owner to use the CFAA to prevent screen scraping by simply revoking access. For example, while the Ninth Circuit ruled *en banc* in *U.S. v. Nosal*,⁴⁸ that the phrase *exceeds authorized access* does not extend to viola-

⁴³*U.S. v. Valle*, 807 F.3d 508, 511 (2d Cir. 2015); *see also, e.g., Satmodo, LLC v. Whenever Communications, LLC*, Case No. 17-cv-0192-AJB NLS, at *5 (S.D. Cal. Apr. 14, 2017) (dismissing plaintiff’s CFAA claim in a click fraud case, where the plaintiff alleged improper access by (a) violating the terms and conditions of the search engine’s advertising contracts, and (b) accessing plaintiff’s website after the plaintiff blocked various IP addresses).

⁴⁴*Space Systems/Loral, LLC v. Orbital ATK, Inc.*, 306 F. Supp. 3d 845, 852 (E.D. Va. 2018).

⁴⁵Cal. Penal Code § 502(c).

⁴⁶Nev. Rev. Stat. §§ 205.511(1), 205.4765.

⁴⁷*Oracle USA, Inc. v. Rimini Street, Inc.*, 879 F.3d 948, 961-62 (9th Cir. 2018) (reversing judgment for Oracle on claims under California and Nevada law) (emphasis in original). In *Rimini Street*, the defendant used an automated tool, in violation of Terms of Use restrictions, to download in bulk customer support files that were available for individual download. The court explained, “Oracle obviously disapproved of the method—automated downloading—by which Rimini took Oracle’s proprietary information. But the key to the state statutes is whether Rimini was authorized in the first instance to take and use the information that it downloaded. . . . Because it indisputably had such authorization, at least at the time it took the data in the first instance, Rimini did not violate the state statutes.” *Id.* at 962.

⁴⁸*U.S. v. Nosal*, 676 F.3d 854, 862-63 (9th Cir. 2012) (*en banc*). This opinion was written by Chief Judge Alex Kozinski. Judge Barry G. Silverman filed a dissenting opinion, in which Judge Tallman concurred.

tions of use restrictions, such as those found in employment policies and website Terms of Use agreements, four years later the Ninth Circuit affirmed the defendant's conviction on the same facts for accessing the same computer system *without authorization*.⁴⁹ In that case, David Nosal, an employee at Korn/Ferry, left to start his own competing executive search firm. Although Korn/Ferry explicitly revoked Nosal's computer access credentials, Nosal continued accessing Korn/Ferry computers and information by using the password of his former executive assistant, who remained authorized to access Korn/Ferry's computers.⁵⁰ Although in its first, *en banc* opinion, in 2012, the Ninth Circuit held that Nosal could not be prosecuted for exceeding authorized access, the appellate court subsequently upheld his conviction for accessing Korn/Ferry computers without authorization, in its later opinion in 2016. The court explained that *without authorization* is "an unambiguous, non-technical term that, given its plain and ordinary meaning, means accessing a protected computer without permission."⁵¹ As applied to Nosal, the court explained that the definition "has a simple corollary: once authorization to access a computer has been affirmatively revoked, the user cannot sidestep the statute by going through the back door and accessing the computer through a third party."⁵² Hence, the court held that Nosal didn't simply exceed authorized access; "the 'without authorization' prohibition of the CFAA extends to a former employee whose computer access credentials have been rescinded but who, disregarding the revocation, accesses the computer by other means."⁵³

⁴⁹*U.S. v. Nosal*, 844 F.3d 1024 (9th Cir. 2016). The opinion in *Nosal* was written by Judge Margaret M. McKeown, on behalf of herself and Chief Judge Sidney R. Thomas. Judge Stephen Reinhardt filed a dissent.

⁵⁰*U.S. v. Nosal*, 844 F.3d 1024, 1034-41 (9th Cir. 2016).

⁵¹*U.S. v. Nosal*, 844 F.3d 1024, 1028 (9th Cir. 2016).

⁵²*U.S. v. Nosal*, 844 F.3d 1024, 1028 (9th Cir. 2016).

⁵³*U.S. v. Nosal*, 844 F.3d 1024, 1029-30 (9th Cir. 2016). The Fourth Circuit had earlier concluded, consistent with this view, that "an employee is authorized to access a computer when his employer approves or sanctions his admission to that computer. Thus, he accesses a computer 'without authorization' when he gains admission to a computer without approval." *WEC Carolina Energy Solutions, LLC v. Miller*, 687 F.3d 199, 204 (4th Cir. 2012), *cert. dismissed*, 568 U.S. 1079 (2013); *see also Pulte Homes, Inc. v. Laborers' Int'l Union of North America*, 648 F.3d 295, 303-04 (6th Cir. 2011) (" 'authorization' is '[t]he conferment of legality; . . . sanc-

Likewise, in *Facebook, Inc. v. Power Ventures, Inc.*,⁵⁴ the Ninth Circuit held that a company that accessed Facebook's servers to scrape data with the permission of their joint customers, but in violation of Facebook's Terms of Use agreement, nonetheless could be held civilly liable for accessing Facebook's computers *without authorization*, where Facebook sent Power Ventures a cease and desist letter advising Power Ventures that it was not permitted to do so. In that case, Power Ventures, a rival social network that allowed its users to aggregate and manage their social network accounts from different services including Facebook, ran a promotion offering its users the opportunity to win \$100 by signing up 100 new Power.com friends. If a Power.com user clicked on an icon that included the words "Yes, I do!" then Power Ventures would create an event, photo or status update on the user's Facebook page. The court conceded that this process "arguably gave Power permission to use Facebook's computers to disseminate messages."⁵⁵ However, because Facebook sent Power Ventures a cease and desist letter informing Power Ventures that it was violating Facebook's Terms of Use and demanding that Power Ventures "stop soliciting Facebook users' information, using Facebook content, or otherwise interacting with Facebook through automated scripts[,] the Ninth Circuit held that Facebook had expressly rescinded permission to access its site."⁵⁶

Although Facebook's cease and desist letter referenced a Terms of Use violation, and the Ninth Circuit conceded that

tion.' Commonly understood, then, a defendant who accesses a computer 'without authorization' does so without sanction or permission . . . [and] *has no rights, limited or otherwise*, to access the computer in question.'"; emphasis in original, quoting 1 Oxford English Dictionary 798 (2d ed. 1989) and *LVRC Holdings LLC v. Brekka*, 581 F.3d 1127, 1133 (9th Cir. 2009)).

⁵⁴*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058 (9th Cir. 2016).

⁵⁵*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1067 (9th Cir. 2016). The court elaborated that:

Power reasonably could have thought that consent from Facebook users to share the promotion was permission for Power to access Facebook's computers. In clicking the "Yes, I do!" button, Power users took action akin to allowing a friend to use a computer or to log on to an e-mail account. Because Power had at least arguable permission to access Facebook's computers, it did not initially access Facebook's computers "without authorization" within the meaning of the CFAA.

Id.

⁵⁶*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1067 (9th Cir. 2016).

under *Nosal*, exceeding authorized access as stated in a Terms of Use agreement wouldn't be actionable, at least in the Ninth and Fourth Circuits, the panel explained in a footnote that the reference to Facebook's Terms of Use was not dispositive because the cease and desist letter also "warned Power that it may have violated federal and state law and plainly put Power on notice that it was no longer authorized to access Facebook's computers."⁵⁷

The court also emphasized that after it received the cease and desist letter, Power Ventures knew it no longer had authorization, but continued to access Facebook's computers anyway. After sending the letter, Facebook blocked access to Facebook's website from Power Ventures' IP addresses, which "further demonstrated that Facebook has rescinded permission for Power to access Facebook's computers."⁵⁸ Power Ventures internal emails showed that it was aware that Facebook was blocking its IP addresses but continued to access the Facebook site. In addition, the panel found significant the fact that "a Power executive sent an e-mail agreeing that Power engaged in four 'prohibited activities' [using a person's Facebook account without Facebook's authorization, using automated scripts to collect information from the site, incorporating Facebook's site in another database, and using Facebook's site for commercial purposes]; acknowledging that Power may have 'intentionally and without authorization interfered with [Facebook's] possessory interest in the computer system,' while arguing that the 'unauthorized use' did not cause damage to Facebook;

⁵⁷*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1067 n.3 (9th Cir. 2016). The court also observed that "[b]ecause, initially, Power users gave Power permission to use Facebook's computers to disseminate messages, we need not decide whether websites such as Facebook are presumptively open to all comers, unless and until permission is revoked expressly." *Id.* n.2, citing Orin S. Kerr, *Norms of Computer Trespass*, 116 Colum. L. Rev. 1143, 1163 (2016) (asserting that "websites are the cyber-equivalent of an open public square in the physical world").

⁵⁸*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1068 (9th Cir. 2016). The court cautioned, however, that "[s]imply bypassing an IP address, without more, would not constitute unauthorized use." *Id.* n.5. The court explained that:

Because a blocked user does not receive notice that he has been blocked, he may never realize that the block was imposed and that authorization was revoked. Or, even if he does discover the block, he could conclude that it was triggered by misconduct by someone else who shares the same IP address, such as the user's roommate or co-worker.

Id.

and noting additional federal and state statutes that Power ‘may also be accused of violating,’ ” . . .⁵⁹

The appellate panel explained that the consent that Power Ventures received from Facebook users to access their accounts “was not sufficient to grant continuing authorization to access Facebook’s computers after Facebook’s express revocation of permission.”⁶⁰ Accordingly, the court held that effective on the date it received Facebook’s cease and desist letter, Power Ventures’ access to Facebook’s computers was *without authorization* within the meaning of the CFAA.⁶¹ On similar grounds, the appellate panel also affirmed the entry of judgment in favor of Facebook on its claim under the California Comprehensive Computer Data Access and Fraud Act (Cal. Penal Code § 502).⁶²

⁵⁹*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1067-68 (9th Cir. 2016).

⁶⁰*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1068 (9th Cir. 2016). The court offered the following analogy:

Suppose that a person wants to borrow a friend’s jewelry that is held in a safe deposit box at a bank. The friend gives permission for the person to access the safe deposit box and lends him a key. Upon receiving the key, though, the person decides to visit the bank while carrying a shotgun. The bank ejects the person from its premises and bans his reentry. The gun-toting jewelry borrower could not then reenter the bank, claiming that access to the safe deposit box gave him authority to stride about the bank’s property while armed. In other words, to access the safe deposit box, the person needs permission both from his friend (who controls access to the safe) and from the bank (which controls access to its premises). Similarly, for Power to continue its campaign using Facebook’s computers, it needed authorization both from individual Facebook users (who controlled their data and personal pages) and from Facebook (which stored this data on its physical servers). Permission from the users alone was not sufficient to constitute authorization after Facebook issued the cease and desist letter.

Id.

⁶¹A similar outcome was reached in *Tech Systems, Inc. v. Pyles*, 630 F. App’x 184, 186 (4th Cir. 2015) (affirming that once she was no longer employed, “Pyles accessed her corporate email account and company-issued Blackberry without authorization.”); *see also U.S. v. Shahulhameed*, 629 F. App’x 685 (6th Cir. 2005) (affirming the defendant’s conviction where, a few hours after he was fired, a cyberattack was launched from his Toyota-assigned laptop; rejecting the defendant’s argument that his access to his former employer’s network at the time was authorized because his account was not disabled until 8 hours later because “Toyota’s failure to disable his account does not mean his access was authorized . . .”).

⁶²Cal. Penal Code § 502; *Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1069 (9th Cir. 2016) (affirming liability under section 502 where the defendant continued to access Facebook’s servers after having received

On remand, the district court in *Power Ventures* awarded \$79,640.50 in damages under the Computer Fraud and Abuse Act for losses incurred on or after the date that Facebook revoked consent by sending the cease and desist letter and entered a permanent injunction pursuant to the Computer Fraud and Abuse Act and Cal. Penal Code § 502.⁶³ In entering injunctive relief, the district court noted that

a cease and desist letter instructing it to stop doing so). A claim under section 502 is similar to a claim under the federal Computer Fraud & Abuse Act, 18 U.S.C.A. § 1030, except that “the California statute does not require *unauthorized* access. It merely requires *knowing* access.” *U.S. v. Christensen*, 828 F.3d 763, 789 (9th Cir. 2016) (emphasis in original). *Access*, according to the Ninth Circuit, “includes logging into a database with a valid password and subsequently taking, copying, or using the information in the database improperly.” *Id.*

Although at the margin there may be a difference in a given case between no authorization (or exceeding authorization) and knowing access, section 502 and CFAA claims often are decided in tandem. *See, e.g., In re Apple & ATTM Antitrust Litig.*, No. C 07-05152 JW, 2010 WL 3521965, at *7 (N.D. Cal. July 8, 2010) (granting summary judgment for defendant Apple on plaintiffs’ CFAA, section 502 and common law trespass to chattels claims where the plaintiffs alleged injury arising from their installation of iOS 1.1.1., which allegedly caused damage to their iPhones and made certain third party apps inaccessible, because plaintiffs voluntarily installed the software upgrade and “[v]oluntary installation runs counter to the notion that the alleged act was a trespass and to [the] CFAA’s requirement that the alleged act was ‘without authorization’ as well as the CPC’s requirement that the act was ‘without permission.’” (citing 18 U.S.C.A. § 1030(a)(5)(A)(1) (which requires a showing that a defendant “intentionally caus[ed] damages without authorization) and Cal. Penal Code §§ 502(b)(10) (which requires the knowing introduction of “computer instructions that are designed to . . . damage.”), 502(c)(4) (which requires a showing that a defendant knowingly accessed and without permission added, altered, damaged, deleted, or destroyed any data, computer software, or computer programs which resided or existed internally or externally to a computer, computer system, or computer network)).

⁶³*See Facebook, Inc. v. Power Ventures, Inc.*, 252 F. Supp. 3d 765 (N.D. Cal. 2017). The court entered the following permanent injunction under the CFAA:

1. Defendants, their agents, officers, contractors, directors, shareholders, employees, subsidiary companies or entities, affiliated or related companies and entities, assignees, and successors-in-interest, and those in active concert or participation with them, are permanently enjoined from:

A. Accessing or using, or directing, aiding, facilitating, causing, or conspiring with others to use or access the Facebook website or servers for any commercial purpose, without Facebook’s prior permission, including by way of example and not limitation for the purpose of sending or assisting others in sending, or procuring the sending, of unsolicited commercial electronic text messages via the Facebook website or service.

“[n]umerous courts have found that unauthorized access of computers and the acquisition of data in violation of the CFAA constitute irreparable harm.”⁶⁴ The court subsequently

B. Using any data, including without limitation Facebook-user data and data regarding Facebook’s website or computer networks, obtained as a result of the unlawful conduct for which Defendants’ have been found liable.

C. Developing, using, selling, offering for sale, or distributing, or directing, aiding, or conspiring with others to develop, sell, offer for sale, or distribute, any software that allows the user to engage in the conduct found to be unlawful.

2. Defendants, their agents, officers, contractors, directors, shareholders, employees, subsidiary companies or entities, affiliated or related companies and entities, assignees, and successors-in-interest, and those in active concert or participation with them shall destroy any software, script(s) or code designed to access or interact with the Facebook website, Facebook users, or the Facebook service. They shall also destroy Facebook data and/or information obtained from Facebook or Facebook’s users, or anything derived from such data and/or information.

3. Within three calendar days of entry of this permanent injunction and order, Defendants shall affirm that they already have notified, or shall notify, their current and former officers, agents, servants, employees, successors, and assigns, and any persons acting in concert or participation with them of this permanent injunction.

4. Within seven calendar days of entry of this injunction and order, Defendants shall certify in writing, under penalty of perjury, that they have complied with the provision of this order, and state how notification of this permanent injunction in accordance with paragraph 3 above was accomplished, including the identities of all email accounts (if any) used for notification purposes.

5. The Court shall continue to retain jurisdiction over the parties for the purpose of enforcing this injunction and order.

Id. at 785-86.

⁶⁴See *Facebook, Inc. v. Power Ventures, Inc.*, 252 F. Supp. 3d 765, 782-86 (N.D. Cal. 2017), citing *TracFone Wireless, Inc. v. Adams*, 98 F. Supp. 3d 1243, 1252-53, 1255-56 (S.D. Fla. 2015) (permanently enjoining the defendant pursuant to the CFAA and Lanham Act where, among other things, “TracFone would be irreparably harmed because Adams’ actions, if allowed to persist, will continue to cause TracFone to suffer harm by impairing the integrity of TracFone’s proprietary computer system and wireless telecommunications network.”); *Reliable Property Services, LLC v. Capital Growth Partners, LLC*, 1 F. Supp. 3d 961, 965 (D. Minn. 2014) (preliminarily enjoining a copyright owner that accessed the plaintiff—snow removal service’s computer system, disabled a program, and stole confidential customer information, finding a “substantial threat of irreparable harm” based on the public dissemination of information after the defendant “unlawfully took volumes of detailed data” in violation of the CFAA); *Enargy Power Co. v. Wang*, Civil Action No. 13-11348-DJC, 2013 WL 6234625, at *10 (D. Mass. Dec. 3, 2013) (preliminarily enjoining the defendant subject to a \$10,000 bond because, among other things, “prevent[ing] Enargy from enjoying the uninterrupted use of its property. . . constitutes irreparable harm. Furthermore, Plaintiffs’ inability to make use of the PH Project files has hampered Enargy. . .”) (internal citations omitted); see also *Tagged, Inc. v. Does 1 through 10*, No. C 09-01713 WHA,

awarded Facebook \$145,028.40 in attorneys' fees under section 502.⁶⁵

In *Ticketmaster LLC v. Prestige Entertainment West, Inc.*,⁶⁶ the court, following *Power Ventures*, denied defendants' motion to dismiss plaintiff's claims under the CFAA and California Computer Data Access and Fraud Act, where Ticketmaster had sent the defendant a letter demanding that it stop accessing the Ticketmaster site using bots to make automated purchases of thousands of tickets to the popular show *Hamilton*, but the defendant continued to do so anyway.

In *Teva Pharmaceuticals USA, Inc. v. Sandhu*,⁶⁷ the court held that plaintiff failed to state a claim against Barinder Sandhu, an employee who was permitted to access plaintiff's computer in the course of her employment to access information in its database, including the information she allegedly shared with the other defendants, but that Teva could state a CFAA claim against those other defendants, who, as outsiders, were "akin to hackers," and who plausibly could be held liable for acting in concert with Sandhu, for directing, encouraging, or inducing her to access the Teva computer system, which they were unauthorized to do.

In *Craigslist Inc. v. 3Taps Inc.*,⁶⁸ an earlier opinion, a district court in the Ninth Circuit also found revocation of consent under the CFAA based on a cease and desist letter. In that case, the court ruled that Craigslist had stated a claim for a CFAA violation against a company that scraped classified ads from its website. The court considered that by making classified advertisements publicly available on its website, Craigslist had authorized the world to access

2010 WL 370331, at *12 (N.D. Cal. Jan. 25, 2010) (permanently enjoining the defendant in part because of the likelihood that the defendant might violate the CFAA and Cal. Penal Code § 502 again in the future).

⁶⁵See *Facebook, Inc. v. Power Ventures, Inc.*, Case No. 08-CV-05780, 2017 WL 3394754, at *6-8 (N.D. Cal. Aug. 8, 2017). Judge Koh wrote in dicta that section 502 allows only prevailing plaintiffs to recover fees. See *id.* at *6. For purposes of section 502, a party is a "prevailing party" if they have a "net monetary recovery" within the meaning of Cal. Code Civ. Proc. § 1032(a)(4). 2017 WL 3394754, at *6.

⁶⁶*Ticketmaster LLC v. Prestige Entertainment West, Inc.*, 315 F. Supp. 3d 1147, 1167-76 (C.D. Cal. 2018).

⁶⁷*Teva Pharmaceuticals USA, Inc. v. Sandhu*, 291 F. Supp. 3d 659, 669-71 (E.D. Pa. 2018).

⁶⁸*Craigslist Inc. v. 3Taps Inc.*, 964 F. Supp. 2d 1178 (N.D. Cal. 2013).

Craigslist.org.⁶⁹ However, the court found that Craigslist revoked its authorization when it sent the defendant a cease and desist letter banning it from using the site. Judge Breyer ruled that a website owner has the power to revoke authorization to access its site and therefore the defendant acted without authorization when it continued to scrape data from Craigslist's website after the time it had received Craigslist's cease and desist letter.

In *hiQ Labs, Inc. v. LinkedIn Corp.*,⁷⁰ however, Judge Edward Chen of the same district held that hiQ raised serious questions about LinkedIn's entitlement to relief under the CFAA (and California Penal Code § 502), in granting a preliminary injunction prohibiting LinkedIn from blocking hiQ's access, copying or use of public profiles on LinkedIn's website (information which LinkedIn members had designated as public) or blocking or putting in place technical or legal mechanisms to block hiQ's access to these public profiles. In that case, LinkedIn had sent a cease and desist letter demanding that hiQ, a company that provided information to businesses about their workforces based on statistical analyses of public profiles on LinkedIn, stop using bots to automatically scrape its site. hiQ instead filed suit for injunctive relief, claiming that its data analytics business was "wholly dependent" on LinkedIn's public data, which it alleged that it had been accessing for more than five years prior to receiving the cease and desist letter. In expressing serious questions about LinkedIn's entitlement to relief under the CFAA, the court held that the CFAA "was not intended to police traffic to publicly available websites on the Internet" ⁷¹

In so ruling, Judge Chen looked beyond the language of the statute and even its legislative history to focus on "the

⁶⁹*Craigslist Inc. v. 3Taps Inc.*, 964 F. Supp. 2d 1178, 1182 (N.D. Cal. 2013), citing *Pulte Homes, Inc. v. Laborer's Int'l Union of North America*, 648 F.3d 295, 304 (6th Cir. 2011) (stating that the public presumptively was authorized to access an "unprotected website"); see also *CollegeSource, Inc. v. AcademyOne, Inc.*, Civil Action No. 10-3542, 2012 WL 5269213, at *14 (E.D. Pa. Oct. 25, 2012) (holding that documents available to the general public on the plaintiff's website could not be accessed without authorization).

⁷⁰*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099 (N.D. Cal. 2017).

⁷¹*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1109 (N.D. Cal. 2017).

Act's theoretical underpinning . . . as a statute addressing the problem of computer 'trespass' . . .⁷² and its "historical context . . ."⁷³ Relying on a law review article by Professor Orin Kerr, Judge Chen explained that "because the Web is generally perceived as 'inherently open,' in that it 'allows anyone in the world to publish information that can be accessed by anyone else without requiring authentication,' . . . 'authorization,' in the context of the CFAA, should be tied to an authentication system, such as password protection"⁷⁴

Although hiQ was not required to use a password to access public information from LinkedIn's site, LinkedIn required use of CAPTCHA — a program designed to allow humans, but not bots, to access its site—and had challenged both hiQ's access *and* its automatic scraping of data. LinkedIn alleged that hiQ used bots to automatically access its site and circumvent CAPTCHA restrictions. Judge Chen, however, construed *authorization* to refer to "the *identity* of the person accessing the computer or website, not *how* access occurs."⁷⁵ Citing Professor Kerr, Judge Chen concluded that by circumventing CAPTCHA, hiQ did not access LinkedIn *without authorization* within the meaning of the CFAA because:

Unlike a password gate, a CAPTCHA does not limit access to certain individuals; it is instead intended "as a way to slow

⁷²*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1111 (N.D. Cal. 2017).

⁷³*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1109 (N.D. Cal. 2017).

⁷⁴*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1112 (N.D. Cal. 2017), *citing* Orin S. Kerr, *Norms of Computer Trespass*, 116 Colum. L. Rev. 1143, 1161-62 (2016). Judge Chen also noted that the U.S. Supreme Court, in *Packingham v. North Carolina*, 137 S. Ct. 1730 (2017), analogized the Internet in general, and social networking sites in particular, to the "modern public square" *Id.* at 1737. In that case, Judge Chen explained, the Court struck down a North Carolina law making it a felony for a registered sex offender to access social media websites like Facebook and Twitter, writing that "at present, social media sites are for many people 'the principal sources for knowing current events, checking ads for employment, speaking and listening in the modern public square, and otherwise exploring the vast realms of human thought and knowledge.'" 2017 WL 3473663, at *7, *quoting* *Packingham v. North Carolina*, 137 S. Ct. 1730, 1737 (2017). This First Amendment restriction on a state's ability to restrict access to websites, needless to say, would not transfer directly to private disputes where there is no government action.

⁷⁵*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1113 (N.D. Cal. 2017) (emphasis in original).

. . . a user's access rather than as a way to deny authorization to access." . . . Other measures taken by website owners to block or limit access to bots may be thought of in the same way. A user does not "access" a computer "without authorization" by using bots, even in the face of technical countermeasures, when the data it accesses is otherwise open to the public. Thus, under Professor Kerr's analysis, hiQ's circumvention of LinkedIn's measures to prevent use of bots and implementation of IP address blocks does not violate the CFAA because hiQ accessed only publicly viewable data not protected by an authentication gateway.⁷⁶

In holding that he had serious reservations about whether LinkedIn's revocation of permission to access public portions of its site rendered hiQ's access "without authorization," Judge Chen distinguished *Power Ventures* and *Nosal* as cases that did not involve public data.⁷⁷

For the same reasons the court had reservations about the viability of LinkedIn's CFAA claim, Judge Chen concluded that hiQ raised serious questions about whether LinkedIn had a claim under "the California analog to the CFAA," California Penal Code § 502, which prohibits *knowing access*.⁷⁸

It remains to be seen whether the Ninth Circuit or courts in other circuits will accept Judge Chen's narrow construction of what constitutes *without authorization* given that the distinctions he drew are not apparent from the face of the statute or even the CFAA's legislative history, that CAPTCHA could be viewed as a way of limiting authorized access to a website, and that a court could consider that the owner of a company's website should be allowed to condition access to and use of its site. Nevertheless, for the moment, *hiQ v.*

⁷⁶*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1113 (N.D. Cal. 2017) (footnote omitted), *citing* Orin S. Kerr, *Norms of Computer Trespass*, 116 Colum. L. Rev. 1143, 1170 (2016).

⁷⁷*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1113 (N.D. Cal. 2017).

⁷⁸*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1115 n.13 (N.D. Cal. 2017) (distinguishing *Chrisman v. City of Los Angeles*, 155 Cal. App. 4th 29, 34, 65 Cal. Rptr. 3d 701 (2007) (noting that "[s]ection 502 defines 'access' in terms redolent of 'hacking' or breaking into a computer")). Judge Chen wrote that "[t]hrough the statute also includes a provision that prohibits 'knowingly access[ing] and without permission tak[ing], copy[ing], or mak[ing] use of any data from a computer, computer system, or computer network,' Cal. Pen. Code § 502(c)(2), the Court similarly concludes there are serious questions about whether these provisions criminalize viewing public portions of a website." 273 F. Supp. 3d at 1115 n.13.

LinkedIn may be viewed as a potential impediment for database owners seeking to sue under the CFAA for unauthorized access when a third party accesses a site to scrape publicly available data.

hiQ v. LinkedIn would not limit a CFAA claim where material being scraped is protected by a password or other technical means (other than CAPTCHA or equivalent technologies that in Judge Chen's analysis arguably slow, rather than prevent, access to material).

Although Judge Chen enjoined LinkedIn from blocking access to hiQ, he wrote that websites like LinkedIn could employ "anti-bot measures to prevent . . . harmful intrusions or attacks on its server."⁷⁹ In view of the injunction he entered, those measures presumably would have had to prevent harmful intrusions, rather than merely deprive hiQ of access to the site.⁸⁰

Judge Chen also noted that, in addition to technological self-help, LinkedIn could pursue other legal remedies, if available.⁸¹ Those remedies presumably could include the other claims addressed in this chapter, including, if applicable, breach of contract and unfair competition. Circumventing CAPTCHA also could constitute a violation of the anti-circumvention provisions of the Digital Millennium Copyright Act,⁸² whether or not actionable under the CFAA.

Ultimately, these cases suggest that a database owner

⁷⁹*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1113 (N.D. Cal. 2017).

⁸⁰In applying *hiQ Labs, Inc. v. LinkedIn Corp.*, it is important to keep in mind the procedural posture of the case. First, suit was filed by hiQ, seeking to enjoin enforcement, rather than by LinkedIn, to obtain affirmative relief under the CFAA. Second, because hiQ sought a preliminary injunction, the court merely found substantial questions going to the applicability of the CFAA—not ultimate liability. Third, for injunctive relief to obtain, the court was required to find irreparable harm, which in *hiQ* arose in part because hiQ's business depended on access to public profiles on LinkedIn, which may not be true in every case. Fourth, in weighing the balance of hardships, which a court must do in evaluating requests for injunctive relief, the court considered that hiQ alleged that it had had access to public profiles on LinkedIn for five years, which was also a unique factor that might not be present in every case.

⁸¹*hiQ Labs, Inc. v. LinkedIn Corp.*, 273 F. Supp. 3d 1099, 1113 n.11 (N.D. Cal. 2017).

⁸²*See, e.g., Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1055–57 (N.D. Cal. 2010) (entering a default judgment for violations of sections 1201(a)(2) and 1201(b)(1) and awarding \$470,000 in statutory

may restrict access by contract, employment agreement, Terms of Use, cease and desist letter, or otherwise, but in at least the Second, Fourth and Ninth Circuits it may not maintain a claim for exceeding authorized access based on a use restriction. *Power Ventures* and *Craigslist* also underscore that a simple letter may be the most effective way to establish that access is not authorized (or, if it was authorized, that authorization has been revoked). A letter may not be effective in establishing that access is unauthorized, however, under *hiQ v. LinkedIn*, where the recipient of the letter is merely accessing publicly available information.

Use restrictions may also be difficult to enforce in certain jurisdictions or before judges who are hostile to Terms of Use or other unilateral contracts. A district court in an older criminal case, for example, ruled that unauthorized access defined by Terms of Service rendered the CFAA void for vagueness, at least in the case of a misdemeanor prosecution.⁸³ Some courts have also held that Terms of Use cannot define access to a site unless express assent is obtained or other measures have been taken to prevent access,⁸⁴ but this analysis is likely mistaken because a contract

damages under the DMCA where the defendant marketed products that circumvented plaintiff's CAPTCHA software and telephone verification security measures); *Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1111–12 (C.D. Cal. 2007) (holding that Ticketmaster was likely to prevail on its DMCA claims under section 1201(a)(2) (trafficking in a device that circumvents technological measures that control access to a protected work) and 1201(b)(1) (for trafficking in a device that circumvents technological measures that protect the rights of a copyright owner in a work) relating to circumvention of CAPTCHA); see generally *infra* § 5.07[1] (addressing the anti-circumvention provisions of the DMCA).

⁸³See *U.S. v. Drew*, 259 F.R.D. 449 (C.D. Cal. 2009) (granting Fed. R. Crim. Proc. 29(c) motion and holding that a misdemeanor conviction under sections 1030(a)(2)(C) and 1030(c)(2)(A) based on the defendant's exceeding the scope of authorized access as defined by MySpace's Terms of Service agreement void for vagueness, but suggesting that a different result might have obtained in the case of a felony conviction because of the scienter requirements imposed for felony prosecutions). *Drew* influenced Judge Kozinski's decision in the first *Nosal* case. See *U.S. v. Nosal*, 676 F.3d 854, 862 (9th Cir. 2012) (*en banc*); see generally *infra* § 44.08 (discussing *Drew* and its impact on *Nosal* in connection with a more thorough analysis of the CFAA).

⁸⁴See, e.g., *Cvent, Inc. v. Eventbrite, Inc.*, 739 F. Supp. 2d 927, 932–33 (E.D. Va. 2010) (dismissing plaintiff's CFAA claim based on the defendant-competitor's scraping the contents of its website in violation of a TOU pro-

may be former by implied, as well as express assent.⁸⁵ Needless to say, a CFAA claim may not be based on a TOU or EULA that does not expressly prohibit access.⁸⁶

Some courts have held that intentionally targeting email or phone calls to a company to prevent it from receiving calls, emails or voicemail messages may be actionable under the CFAA.⁸⁷

In addition to corporate exposure, personal liability also may be imposed under the CFAA. In *Facebook, Inc. v. Power*

vision that stated that “No competitors or future competitors are permitted access to our site or information, and any such access by third parties is unauthorized . . .” because the website took no affirmative steps to screen competitors from accessing its site and does not password protect its database or require express assent from users to its Terms of Use); *Koch Industries, Inc. v. Does 1–25*, No. 2:10CV1275DAK, 2011 WL 1775765, at *7-8 (D. Utah May 9, 2011) (following *Cvent* in dismissing a CFAA claim brought against political protestors who accessed and copied Koch Industries’ genuine website to set up a fake one at a different location to protest its political advocacy on global warming issues).

Cvent is a controversial case in which the court dismissed a complaint for breach of contract based on posted website Terms of Use, concluding that the plaintiff could not state a claim based on implied assent (notice and subsequent conduct). As analyzed in section 21.03, there is no legal basis to conclude that a contract may only be formed by express (click-through) assent. The *Cvent* court’s holding (and those of courts following *Cvent*) that posted TOU cannot define whether access to a site is authorized under the CFAA (or whether authorized access has been exceeded, in those circuits where a claim may be based on a use restriction) should be viewed in the context of the *Cvent* court’s flawed analysis of contract formation. See generally *infra* § 21.03 (analyzing both express and implied assent in the formation of Terms of Use and other unilateral contracts).

⁸⁵See *infra* § 21.03 (analyzing express and implied assent).

⁸⁶See *Earthcam, Inc. v. Oxblue Corp.*, 49 F. Supp. 3d 1210, 1231–32 (N.D. Ga. 2014) (granting summary judgment for the defendant on plaintiff’s CFAA claim premised on unauthorized access to login credentials to a website where the OxBlue Defendants had accessed the account webpage with the Earthcam customer’s authorization and nothing in plaintiff’s EULA applicable to the customer’s account prohibited the customer from sharing its login credentials and there was no evidence that the OxBlue Defendants were aware of any EULA restrictions), *aff’d*, 703 F. App’x 803, 806-10 (11th Cir. 2017).

⁸⁷See *Pulte Homes, Inc. v. Laborers’ Int’l Union of North America*, 648 F.3d 295 (6th Cir. 2011); see also *U.S. v. Carlson*, 209 F. App’x 181, 185 (3d Cir. 2006) (upholding a conviction where the defendant admitted sending thousands of email messages to plaintiff).

Ventures, Inc.,⁸⁸ the Ninth Circuit affirmed the entry of judgment for Facebook against Power Ventures and its CEO. The appellate court ruled that corporate officers or directors, in general, are personally liable for all torts which they authorize or direct or in which they participate, notwithstanding that they acted as agents for the corporation and not on their own behalf.⁸⁹ The appellate court noted, however, that personal liability typically is only imposed on corporate officers where the officer has been the *guiding spirit* behind the wrongful conduct.⁹⁰

A claim under the CFAA must be brought within two years of the date of the act complained of or the date of discovery of the damage.⁹¹

In limited circumstances a claim brought against a platform, intermediary or other interactive computer service for the misconduct of users could be preempted by the Communications Decency Act.⁹²

The CFAA is analyzed in greater detail (along with

⁸⁸*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1069-70 (9th Cir. 2016).

⁸⁹*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1069 (9th Cir. 2016), citing *Committee for Idaho's High Desert, Inc. v. Yost*, 92 F.3d 814, 823 (9th Cir. 1996).

⁹⁰*Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1069 (9th Cir. 2016), citing *Davis v. Metro Products, Inc.*, 885 F.2d 515, 523 n.10 (9th Cir. 1989).

⁹¹18 U.S.C.A. § 1030(g); *Sewell v. Bernardin*, 795 F.3d 337 (2d Cir. 2015) (holding claims by the plaintiff against her ex-boyfriend that were filed on January 2, 2014 time barred with respect to her AOL email account, where the plaintiff first found she could not log into her account on August 1, 2011, but not time barred with respect to access to her Facebook account, where she discovered that she could not log on, because her password had been altered, on February 24, 2012).

⁹²See 47 U.S.C.A. § 230(c); *Holomaxx Technologies v. Microsoft Corp.*, 783 F. Supp. 2d 1097 (N.D. Cal. 2011) (dismissing as preempted by section 230(c)(2) (with leave to amend) plaintiff's claim under the Computer Fraud and Abuse Act); *Holomaxx Technologies v. Yahoo!, Inc.*, No. CV-10-4926-JF, 2011 WL 865794 (N.D. Cal. Mar. 11, 2011) (ruling the same way in dismissing Holomaxx's virtually identical complaint against Yahoo!); *e360Insight, LLC v. Comcast Corp.*, 546 F. Supp. 2d 605 (N.D. Ill. 2008) (granting judgment on the pleadings in favor of Comcast under the section 230(c)(2) on plaintiff's claims for violations of the Computer Fraud and Abuse Act, 18 U.S.C.A. § 1030, and unfair practices barred by the Illinois Consumer Fraud Act, arising out of Comcast's blocking email from e360, a bulk emailer, to Comcast subscribers); see generally *infra* § 37.05 (analyzing the CDA and the scope of its preemption in greater detail).

conflicting lines of cases construing the statute) in section 44.08. CFAA claims in data privacy class action suits are analyzed in section 26.15.

In addition to CFAA claims, a number of states have enacted computer crime statutes that may provide additional remedies. Some of these statutes have been construed to be consistent with the CFAA,⁹³ while others are more akin to trespass statutes.⁹⁴ State statutes may afford relief not available under the CFAA, such as in a case where the \$5,000 damage threshold cannot be met.

5.07 DMCA and BOTS Act Claims

The Digital Millennium Copyright Act (DMCA) and Better Online Ticket Sales Act of 2016 (colloquially known as the BOTS Act) both proscribe circumvention of technological measures under different circumstances. The DMCA applies generally to circumvention of access and control mechanisms that protect copyrighted works.¹ The BOTS Act prohibits circumvention of a security measure, access control system,

⁹³See, e.g., N.J. Stat. Ann. § 2A:38A-3; *In re Nickelodeon Consumer Privacy Litig.*, 827 F.3d 262, 277-78 (3d Cir. 2016) (affirming the district court's dismissal with prejudice of plaintiffs' claims under the New Jersey Computer Related Offenses Act (CROA), N.J. Stat. Ann. § 2A:38A-3, an anti-hacking statute, because plaintiffs could not "allege that they had been 'damaged in business or property,' as the plain text of the New Jersey Act requires"), *cert. denied*, 137 S. Ct. 624 (2017). New Jersey courts, the panel noted, construe the statute as requiring the same type of evidence of damage as that required by the federal Computer Fraud and Abuse Act, 18 U.S.C.A. § 1030. *Id.* at 278.

⁹⁴See, e.g., Mo. Rev. Stat. §§ 537.525, 569.095 (the Missouri Computer Tampering Act, which prohibits, among other things, the knowing unauthorized receipt, use or disclosure of data and authorizes a civil action by the data owner); N.C. Gen. Stat. § 14-458 (making it illegal for any person "to use a computer or computer network without authority and with the intent" to undertake various alternative acts including removing, altering, erasing or disabling computer data, programs or software or making an unauthorized copy); *Spirax Sarco, Inc. v. SSI Engineering, Inc.*, 122 F. Supp. 3d 408, 417-18 (E.D.N.C. 2015) (holding that the plaintiff stated a claim under section 14-458(a) where it alleged that the defendant "intentionally used his Spirax-issued laptop to download vast quantities of computer files to his own media devices and Dropbox account, without authorization and in contravention of Spirax policies, and he also deleted vast quantities of computer files from the Spirax-issued laptop without authorization.").

[Section 5.07]

¹See 17 U.S.C.A. §§ 1201 *et seq.*; *infra* § 5.07[1].

or other technological control or measure that an Internet site or service uses to enforce posted event ticket purchasing limits or to maintain the integrity of posted online ticket purchasing order rules.² The BOTS Act applies only to online ticket sales, whereas the DMCA generally applies to content protection and access control measures used to prevent copying of copyrighted works. The DMCA also proscribes removing copyright management information (CMI) and using false CMI.³

5.07[1] DMCA Anti-Circumvention Provisions

The anti-circumvention provisions of the Digital Millennium Copyright Act (DMCA) may provide remedies to a database owner where a third party attempts to circumvent measures intended to protect a copyrighted database or access to the database.¹ These provisions are analyzed in greater detail in section 4.21.

The anti-circumvention provisions of the DMCA protect access control and copy protection mechanisms. Section 1201(a) protects against circumvention of access controls, while 1201(b) proscribes circumvention of technologies that protect against copying. Sections 1201(a)(2) and 1201(b)(1) collectively are referred to as anti-trafficking provisions. Section 1201(b)(1) addresses trafficking in technologies that circumvent technical measures that prevent copying. Section 1201(a)(2) prohibits trafficking in technologies that circumvent technological measures that effectively control access to protected works. Section 1201(a)(1), in turn, prohibits the actual circumvention of access controls. There is no corresponding prohibition on the actual circumvention of copy protection mechanisms, which may in some circumstances amount to a fair use under copyright law.

One court characterized the elements necessary to state a DMCA claim as: (1) ownership of a valid copyright; (2) circumvention of a technological measure designed to protect the copyrighted material; (3) unauthorized access by third parties; (4) infringement because of the circumvention; and (5) the circumvention was achieved through software that

²See 15 U.S.C.A. § 45c; *infra* § 5.07[3].

³See 17 U.S.C.A. § 1202; *infra* § 5.07[2].

[Section 5.07[1]]

¹17 U.S.C.A. §§ 1201 *et seq.*

the defendant either (i) designed or produced primarily for circumvention; (ii) made available despite only limited commercial significance other than circumvention; or (iii) marketed for use in circumvention of the controlling technological measure.²

In *Ticketmaster LLC v. RMG Technologies, Inc.*,³ the court ruled that Ticketmaster was likely to prevail on its claim under section 1201(a)(2) (trafficking in a device that circumvents technological measures that control access to a protected work) that the defendant violated the DMCA by offering a software tool that allowed its customers to circumvent technological measures, such as CAPTCHA,⁴ employed by Ticketmaster to block automated access to its site. The court also found that Ticketmaster was likely to prevail on its claim under section 1201(b)(1) (for trafficking in a device that circumvents technological measures that protect the rights of a copyright owner in a work). The court reasoned that CAPTCHA both controls access to a protected work because a user could not proceed to copyright protected webpages without solving CAPTCHA and protects rights of a copyright owner because, by preventing automated access to the ticket purchase webpage, CAPTCHA prevents users from copying those pages.⁵

Similarly, in *Facebook, Inc. v. Power Ventures, Inc.*,⁶ Judge Jeremy Fogel of the Northern District of California denied defendant's motion to dismiss plaintiff's DMCA claim, which was premised on the defendants allegedly circumventing technical measures intended to prevent them from accessing Facebook's website to copy user data in violation of the site's Terms of Use agreement. In so ruling, he rejected the defendants' argument that their access was not unautho-

²*Facebook, Inc. v. Power Ventures, Inc.*, 91 U.S.P.Q.2d 1430, 2009 WL 1299698 (N.D. Cal. May 11, 2009), citing *Chamberlain Group, Inc. v. Skylink Technologies, Inc.*, 381 F.3d 1178, 1203 (Fed. Cir. 2004).

³*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096 (C.D. Cal. 2007).

⁴CAPTCHA (an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart") is a computer security program that is designed to distinguish between human users and computer programs, and thereby prevent automated devices from accessing a site.

⁵*Ticketmaster LLC v. RMG Technologies, Inc.*, 507 F. Supp. 2d 1096, 1111–12 (C.D. Cal. 2007).

⁶*Facebook, Inc. v. Power Ventures, Inc.*, 91 U.S.P.Q.2d 1430, 2009 WL 1299698 (N.D. Cal. May 11, 2009).

rized because defendants merely provided a tool that their users (who were also Facebook users) used to access their own content. Judge Fogel concluded, however, that Facebook's Terms of Use agreement barred users from using automated programs to access the site.⁷

The court also rejected defendants' argument that there was no copyrighted work at issue because Facebook did not own a copyright to user content, which ultimately is the information that defendants' software sought to extract. Judge Fogel found that to access this data the defendants made a cached copy of the entire Facebook website and copied a user's entire Facebook profile page, simply to access the user's data. While Facebook did not own a copyright in individual user data, it likely did own one in the compilation of user data found on the Facebook site.

Although not discussed explicitly in the court's brief opinion, to effectively make such a claim, a site owner or service provider must establish that access was unauthorized (which in Facebook was based on exceeding the scope of permissible access, as set forth in the Terms of Use agreement) and that the defendant circumvented technical measures intended to block this access.

Other courts also have entered judgment for database and website owners based on the DMCA,⁸ although no claim may be made absent circumvention.⁹ Using software that ordinarily is only subject to an access control measure is not the

⁷See *supra* § 5.03[2] (discussing specific provisions of the Terms of Use agreement at issue in Facebook).

⁸See, e.g., *Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1055–57 (N.D. Cal. 2010) (entering a default judgment for violations of sections 1201(a)(2) and 1201(b)(1) and awarding \$470,000 in statutory damages under the DMCA where the defendant marketed products that circumvented plaintiff's CAPTCHA software and telephone verification security measures).

⁹See, e.g., *Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey*, 497 F. Supp. 2d 627, 642–46 (E.D. Pa. 2007) (granting summary judgment for the defendants in a case where plaintiffs sued the law firm that previously had represented an opposing party in a trademark infringement suit, alleging that defendants obtained archived copies of its website from the Wayback Machine, www.Archive.org, where the copies were only accessible because of a computer malfunction that caused the Archive.org site to ignore the Robots.txt files on plaintiff's site that would otherwise have resulted in the archived pages being made publicly inaccessible; the defendants did not circumvent the Robots.txt file and plaintiff's inference that defendants should have known they were “not al-

same thing as establishing that the defendant in fact engaged in circumvention, as opposed to merely gaining access to it after someone else did so. “Because § 1201(a)(1) is targeted at circumvention, it does not apply to the use of copyrighted works *after* the technological measure has been circumvented.”¹⁰

Where a company markets a product that allows users to employ bots to access a site (for example, to play an online game generating virtual goods or currency), liability could arise under the DMCA to the extent the program allows users to circumvent technological measures intended to defeat the use of bots or for breach of contract, tortious interference with contract or other grounds.¹¹

There presently are circuit splits on a number of issues, including whether section 1201(a) created a new statutory anti-circumvention right distinct from infringement or whether circumvention of an access control is only actionable if it facilitates infringement. Under either view, the work accessed must be a copyrighted work to state a claim under section 1201(a). The difference is that according to the Ninth Circuit, sections 1201(a)(1) and 1201(a)(2) created “a new form of protection, i.e., the right to prevent circumvention of access controls, broadly to. . . copyrighted works[.]”¹² whereas the Federal Circuit requires a further showing of a nexus between the circumvention and infringement, which places circumvention undertaken to enable fair use or other

lowed to view the archived images via the Wayback Machine was both unreasonable and irrelevant” to the DMCA claim).

¹⁰*MGE UPS Systems, Inc. v. GE Consumer and Indus., Inc.*, 622 F.3d 361, 366 (5th Cir. 2010) (*en banc*) (affirming dismissal of plaintiff’s claim; emphasis in original).

¹¹*See MDY Industries, LLC v. Blizzard Entertainment, Inc.*, 629 F.3d 928 (9th Cir. 2011) (affirming entry of judgment for Blizzard under section 1201(a)(2), but reversing judgment on its copyright infringement and section 1201(b) claims and reversing based on disputed facts the entry of summary judgment on plaintiff’s claim for tortious interference with the plaintiff’s Terms of Use); *see also MDY Industries, LLC v. Blizzard Entertainment, Inc.*, No. CV-06-2555-PHX-DGC, 2011 WL 2533450 (D. Ariz. June 27, 2011) (declining to vacate the permanent injunction pursuant to section 1201(a)(2) on remand); *see generally infra* § 51.02[3] (analyzing the case in greater detail in the context of virtual goods and currency).

¹²*MDY Industries, LLC v. Blizzard Entertainment, Inc.*, 629 F.3d 928, 945 (9th Cir. 2011).

noninfringing uses outside of section 1201(a)'s reach.¹³

Anti-circumvention case law is analyzed in substantially greater detail in section 4.21[2].

5.07[2] Removing, Altering or Falsifying Copyright Management Information

Removing, altering or falsifying copyright management information (CMI) is potentially actionable under section 1202 of the Digital Millennium Copyright Act.¹

Section 1202(a) prohibits any person from knowingly and with the intent to induce, enable, facilitate or conceal infringement,

- (1) providing CMI that is false or
- (2) distributing or importing for distribution CMI that is false.

Section 1202(b), in turn, prohibits anyone, without the authority of the copyright owner or the law, from

- intentionally removing or altering any CMI;
- distributing or importing for distribution CMI knowing that the CMI has been removed or altered without authority of the copyright owner or the law; or
- distributing, importing for distribution, or publicly performing works, copies of works, or phonorecords, knowing that CMI has been removed or altered without the authority of the copyright owner or the law,

knowing (or in the case of a civil suit pursuant to section 1203, "having reasonable ground to know") that it will induce, enable, facilitate or conceal copyright infringement.²

Absent knowledge, there can be no violation of section

¹³See *Storage Tech. Corp. v. Custom Hardware Engineering & Consulting, Inc.*, 421 F.3d 1307, 1318-19 (Fed. Cir. 2005) (applying First Circuit law); *Chamberlain Group, Inc. v. Skylink Technologies, Inc.*, 381 F.3d 1178, 1192-1203 (Fed. Cir. 2004) (applying Seventh Circuit law), *cert. denied*, 544 U.S. 923 (2005).

[Section 5.07[2]]

¹See 17 U.S.C.A. § 1202(c)(6); see generally *supra* § 4.21[3] (analyzing the statute in greater detail and discussing case law).

²17 U.S.C.A. § 1202(b). Certain limited exceptions involving broadcast stations and cable systems are set forth in section 1202(e).

1202.³ At least two district courts further have held that a claim under 1202(b) must be premised on removal of CMI from the “body” or “area around” a work to violate the DMCA⁴ and that a claim may not proceed if it is based merely on a general copyright notice appears on an entirely different webpage than the page where the work at issue appears.⁵ The rationale for this rule is to prevent “a ‘gotcha’ system where a picture or piece of text has no CMI near it but the plaintiff relies on a general copyright notice buried elsewhere on the website.”⁶ As a practical matter, however, the requirement that a plaintiff establish knowledge or intent eliminates the risk of “gotcha” liability being imposed.

While a 1202(a) prohibits false CMI, section 1202(b) prohibits removal. To state a claim under section 1202(b) a plaintiff therefore must allege *removal*. Merely copying information into a different form (such as taking notes of an oral lecture and incorporating them into a note package)⁷ does not amount to removal. Similarly, a claim for false CMI under section 1202(a) requires a showing that an alteration was made to an original work and may not be maintained where allegedly infringing material is merely incorporated into a new product that with information that identifies the new product.⁸ By contrast, where attribution information is replaced, a claim may be stated under section 1202(a) where

³See *Gordon v. Nextel Communications*, 345 F.3d 922, 926-27 (6th Cir. 2003) (affirming summary judgment for the defendant on plaintiff’s claim under 1202(b)(3) where there was no evidence of that defendants knew that CMI had been removed or altered without the authority of the copyright owner).

⁴*Schiffer Publishing, Ltd. v. Chronicle Books, LLC*, No. 03 C 4962, 2004 WL 2583817, at *4, 14 (E.D. Pa. Nov.12, 2004) (finding no DMCA violation where a book contained 118 copyrighted photos with no CMI near them and the defendant merely had a general copyright notice on the whole book).

⁵See *Personal Keepsakes, Inc. v. Personalizationmall.com, Inc.*, No. 11 C 5177, 2012 WL 414803, at *7 (N.D. Ill. Feb. 8, 2012).

⁶*Personal Keepsakes, Inc. v. Personalizationmall.com, Inc.*, No. 11 C 5177, 2012 WL 414803, at *7 (N.D. Ill. Feb. 8, 2012). Judge Virginia M. Kendall premised this ruling on the definition of CMI, which requires that CMI be conveyed with a copyrighted work. See *id.*

⁷See *Faulkner Press, LLC v. Class Notes, LLC*, 756 F. Supp. 2d 1352, 1359 (N.D. Fla. 2010).

⁸See *Faulkner Press, LLC v. Class Notes, LLC*, 756 F. Supp. 2d 1352, 1359-60 (N.D. Fla. 2010) (holding that the defendant did not add false CMI by printing “Einstein’s Notes ©” on its note packages where the note package was a different work from Class Notes even if, as the plaintiff al-

a plaintiff can allege (a) knowledge that the CMI information is false, and (2) intent to induce, enable, facilitate or conceal an infringement of any right under Title 17.⁹

Courts disagree about whether *Copyright Management Information* must involve technical measures of automated systems, which is suggested by the legislative history. As analyzed in section 4.21[3], the trend is to construe CMI broadly based on the plain terms of the statute, rather than more narrowly based on legislative history.

In *Associated Press v. All Headline News Corp.*,¹⁰ for example, a court in the Southern District of New York held that the plaintiff had stated a claim under 17 U.S.C.A. § 1202 where it alleged that defendants took Associated Press articles from the Internet and removed information that identified the AP as the owner and author, before reproducing them on their website. In so ruling, however, the court disagreed with courts in New Jersey and California that had held that to state a claim under section 1202 a plaintiff must alter, remove or falsify technological measures of automated systems, which the court in *All Headline News* criticized as being based on legislative history rather than the plain text of the statute.

Following *All Headline News*, the court in *Cable v. Agence France Press*¹¹ denied the defendant's motion to dismiss, holding that the plaintiff's name and a link ("Photos © 2009 wayne cable, selfmadephoto.com"), which allegedly had been removed from plaintiff's photos before they were included without authorization in the defendant's photo database, constituted *copyright management information* "in the absence of evidence to the contrary, which may be considered in the context of future dispositive motions"

leged, it included material from the plaintiff's work).

⁹See, e.g., *Agence France Presse v. Morel*, 769 F. Supp. 2d 295, 304-05 (S.D.N.Y. 2011) (holding that the plaintiff pled a claim for falsification under section 1202(a) by alleging that Agence France Presse labeled his photographs with the credit lines "AFP/Getty/Daniel Morel" and "AFP/Getty/Lisandro Suero" which the plaintiff alleged were false and intended to facilitate infringement); *Ward v. National Geographic Society*, 208 F. Supp. 2d 429, 449 (S.D.N.Y. 2002) (granting summary judgment for the defendant because a defendant's knowledge may not be imputed).

¹⁰*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454 (S.D.N.Y. 2009).

¹¹*Cable v. Agence France Presse*, 728 F. Supp. 2d 977, 981 (N.D. Ill. 2010).

A party's copying and reuse of content from a database with CMI removed potentially could violate section 1202. If CMI is not removed, however, it potentially could lead to exposure under the Lanham Act to the extent that its inclusion is likely to cause confusion or dilution.¹²

Case law on removal, alteration and falsification of Copyright Management Information is analyzed in greater detail in section 4.21[3].

5.07[3] BOTS Act Anticircumvention

The Better Online Ticket Sales Act of 2016 (colloquially known as the BOTS Act)¹ makes it unlawful to “circumvent a security measure, access control system, or other technological control or measure on an Internet website or online service that is used by the ticket issuer² to enforce posted event³ ticket⁴ purchasing limits or to maintain the integrity

¹²See *infra* § 5.08.

[Section 5.07[3]]

¹See 15 U.S.C.A. § 45c.

²While the definition of *ticket issuer* was not codified, a note to the statute includes the definition contained in the law enacted by Congress, which is what the FTC will follow. *Ticket issuer* means

any person who makes event tickets available, directly or indirectly, to the general public, and may include—

- (A) the operator of the venue;
- (B) the sponsor or promoter of an event;
- (C) a sports team participating in an event or a league whose teams are participating in an event;
- (D) a theater company, musical group, or similar participant in an event; and
- (E) an agent for any such person.”

15 U.S.C.A. § 45c note; Pub. L. 114–274 § 3, 130 Stat. 1403 (Dec. 14, 2016).

³While the definition of *event* was not codified, a note to the statute includes the definition contained in the law enacted by Congress, which is what the FTC will follow. *Event* means “any concert, theatrical performance, sporting event, show, or similarly scheduled activity, taking place in a venue with a seating or attendance capacity exceeding 200 persons that—(A) is open to the general public; and (B) is promoted, advertised, or marketed in interstate commerce or for which event tickets are generally sold or distributed in interstate commerce.” 15 U.S.C.A. § 45c note; Pub. L. 114–274 § 3, 130 Stat. 1403 (Dec. 14, 2016).

⁴While the definition of *event ticket* was not codified, a note to the statute includes the definition contained in the law enacted by Congress, which is what the FTC will follow. The term *event ticket* means “any physical, electronic, or other form of a certificate, document, voucher, token, or

of posted online ticket purchasing order rules”⁵ The Act also makes it unlawful “to sell or offer to sell any event ticket in interstate commerce” obtained in violation of this prohibition if the person selling or offering to sell the ticket “participated directly in or had the ability to control the conduct”⁶ or “knew or should have known that the event ticket was acquired in violation of” this prohibition.⁷

The Act creates exceptions allowing a person to create or use any computer software or system “to investigate, or further the enforcement or defense, of any alleged violation of this section or other statute or regulation”⁸ or “to engage in research necessary to identify and analyze flaws and vulnerabilities of measures, systems, or controls . . . if these research activities are conducted to advance the state of knowledge in the field of computer system security or to assist in the development of computer security product.”⁹

Violations of the BOTS Act may be enforced by the Federal Trade Commission as an unfair or deceptive practice,¹⁰ pursuant to the Federal Trade Commission Act,¹¹ or by State Attorneys General.¹²

The BOTS Act neither authorizes a private cause of action nor expressly prohibits one. The Act preempts certain

other evidence indicating that the bearer, possessor, or person entitled to possession through purchase or otherwise has— (A) a right, privilege, or license to enter an event venue or occupy a particular seat or area in an event venue with respect to one or more events; or (B) an entitlement to purchase such a right, privilege, or license with respect to one or more future events.” 15 U.S.C.A. § 45c note; Pub. L. 114–274 § 3, 130 Stat. 1403 (Dec. 14, 2016).

In the view of the FTC, an *event ticket* does not include online travel tickets for bus, train or airline travel. See FTC, *BOTS Act: That’s The Ticket!* (Apr. 7, 2017), <https://www.ftc.gov/news-events/blogs/business-blog/2017/04/bots-act-thats-ticket> (Apr. 13, 2017 5:12 PM Comment).

⁵15 U.S.C.A. § 45c(a)(1)(A).

⁶15 U.S.C.A. § 45c(a)(1)(B)(i).

⁷15 U.S.C.A. § 45c(a)(1)(B)(ii).

⁸15 U.S.C.A. § 45c(a)(2)(A).

⁹15 U.S.C.A. § 45c(a)(2)(B).

¹⁰15 U.S.C.A. § 45c(b)(1).

¹¹15 U.S.C.A. §§ 45, 46; see generally *infra* § 25.02 (analyzing FTC jurisdiction and regulation of Internet and mobile activities).

¹²15 U.S.C.A. § 45c(c).

enforcement by State Attorneys General¹³ but does not purport to create or preempt civil remedies. It is therefore possible that a violation of the BOTS Act, while not independently actionable, could form the basis of a state law unfair competition claim in those states, such as California, that allow claims to be brought for violations under statutes that do not afford a private cause of action¹⁴ (unless it were determined that Congress sought to occupy the field, resulting in “field preemption” of state law claims¹⁵).

5.08 Lanham Act Remedies

Where material copied from a database includes logos or other branding, a claim potentially may be asserted under the Lanham Act.¹ Often, however, a screen scraper is smart

¹³See 15 U.S.C.A. § 45c(c)(4).

¹⁴See, e.g., Cal. Bus. & Prof. §§ 17200 *et seq.* Section 17200 “borrows” violations from other laws by making them independently actionable as unfair competitive practices. *Korea Supply Co. v. Lockheed Martin Corp.*, 29 Cal. 4th 1134, 1143–45, 131 Cal. Rptr. 2d 29 (Cal. 2003). Under section 17200, “[u]nlawful acts are ‘anything that can properly be called a business practice and that at the same time is forbidden by law . . . be it civil, criminal, federal, state, or municipal, statutory, regulatory, or court-made,’ where court-made law is, ‘for example a violation of a prior court order.’” *Sybersound Records, Inc. v. UAV Corp.*, 517 F.3d 1137, 1151–52 (9th Cir. 2008) (citations omitted); see generally *infra* §§ 6.12[6] (analyzing unfair practices laws in greater detail), 25.04[3] (addressing unfair competition in connection with an overview of consumer protection laws).

¹⁵States “are precluded from regulating conduct in a field that Congress, acting within its proper authority, has determined must be regulated by its exclusive governance.” *Arizona v. United States*, 567 U.S. 387, 399 (2012). Preemption may be express, as it is in some statutes, or “[t]he intent to displace state law altogether can be inferred from a framework of regulation ‘so pervasive . . . that Congress left no room for the States to supplement it’ or where there is a ‘federal interest . . . so dominant that the federal system will be assumed to preclude enforcement of state laws on the same subject.’” *Id.*, quoting *Rice v. Santa Fe Elevator Corp.*, 331 U.S. 218, 230 (1947). Given that Congress expressly chose to include a preemption provision applicable to State Attorneys General actions but not to civil claims, an argument may be advanced that Congress did not intend to occupy the field (unless there is a contradictory intention expressed in the legislative history).

[Section 5.08]

¹See 15 U.S.C.A. § 1125(a); see generally *infra* § 6.12. For example, in *Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039, 1058–60 (N.D. Cal. 2010), the court entered a default judgment for trademark infringement under the Lanham Act and California law based on the defendants’ display of the Craigslist mark in the text and in the headings of sponsored

enough to only post unprotectable data, not branding that may have been included with the data. If scraped material is entitled to copyright protection and includes copyright management information (CMI), however, removing the information to avoid exposure under the Lanham Act potentially could result in liability for removing CMI under the Digital Millennium Copyright Act.²

Where a database owner's marks are displayed on a competitor's site, the database owner potentially may maintain a claim for trademark infringement, unfair competition, false designation of origin or passing off, provided likelihood of confusion may be shown.³ If the association of a database owner's marks with a given site tarnishes or blurs the mark, and if the mark may qualify as "famous," a claim also may be maintained for dilution.⁴ If the site merely identifies the database owner, the reference may be a nominative fair use.⁵ Where the fact of display is a function of fair use copying, the display also could be deemed a fair

links advertising products to automate the process of posting listings to Craigslist, in advertising their products and on their website.

In *Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1239–43 (D. Colo. 2009), the court denied the defendant's motion to dismiss plaintiff's trademark infringement claim where the issues of likelihood of confusion and nominative fair use (considered in the context of likelihood of confusion) presented factual issues that could not be resolved on a motion to dismiss. In that case, Health Grades sued a hospital alleging that it breached its click-through license agreement with the plaintiff by commercially reproducing, modifying and/or distributing its healthcare provider award and ranking information from plaintiff's website, including its trademarks, in press releases and other marketing materials.

²17 U.S.C.A. § 1202; *supra* § 5.07[2] (database-related DMCA claims); see generally *supra* § 4.21[3] (analyzing the statute in depth).

³See *infra* chapter 6.

⁴See *infra* § 6.11.

⁵The nominative fair use defense permits certain uses of a trademark to refer to the trademarked product. See *infra* § 6.14[3]. For example, in *Comparison Medical Analytics, Inc. v. Prime Healthcare Services, Inc.*, Case No. 2:14-CV-3448 SVW (MANx), 2015 WL 12746228, at *1-5 (C.D. Cal. Apr. 14, 2015), the court entered summary judgment for the defendant on claims for trademark infringement and unfair competition under the Lanham Act and common law unfair competition based on nominative fair use, in a case brought by a company that "grants to hospitals awards, and then sells them the right to publicize the awards . . .," where the plaintiff gave the defendant "numerous awards . . . [and] then sued Prime for posting news of the awards on its website . . ." without a license to do so.

use in limited circumstances.⁶

Where branding information or credits are edited out of material from a factual database or other compilation that is unprotectable, the database owner will not be permitted to maintain a Lanham Act claim for what amounts to a disguised claim for copyright (which, if actionable, must be brought under the Copyright Act, not the Lanham Act).⁷ Where the material is protectable, however, a claim for re-

⁶See, e.g., *Sega Enterprises Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992) (Sega marks displayed because they were embedded in code needed to be used to make Accolade's game compatible with Sony's player that used a proprietary format); *Health Grades, Inc. v. Robert Wood Johnson University Hosp., Inc.*, 634 F. Supp. 2d 1226, 1239–43 (D. Colo. 2009) (denying defendant's motion to dismiss where the plaintiff's marks were used in connection with a description of the defendant's ranking and the awards it received from the plaintiff, where the court held that nominative fair use was an element of likelihood of confusion, which the plaintiff was required to prove in order to prevail); see generally *infra* §§ 6.12[3][C], 6.14.

⁷See *Dastar Corp. v. Twentieth Century Fox Film Corp.*, 539 U.S. 23 (2003); see also, e.g., *General Universal Systems, Inc. v. Lee*, 379 F.3d 131, 148–49 (5th Cir. 2004) (holding a software developer/licensor's reverse palming off claim preempted where it rested on the defendant's copying ideas, concepts, structures and sequences embodied in his copyrighted work, rather than palming off tangible copies); *Phoenix Entertainment Partners, LLC v. Rumsey*, 829 F.3d 817, 827–31 (7th Cir. 2016) (holding, on different facts from *Dastar*, that plaintiff's passing off claim for copying digital karaoke music files was preempted where the claim was directed at the creative content contained in the file); *Slep-Tone Entertainment Corp. v. Wired for Sound Karaoke & DJ Services, LLC*, 845 F.3d 1246 (9th Cir. 2017) (affirming dismissal under *Dastar* of plaintiff's trademark and trade dress claims as disguised claims alleging copying; following the Seventh Circuit's decision in *Rumsey*); *infra* § 6.12[1] (analyzing *Dastar*). But see *Cvent, Inc. v. Eventbrite, Inc.*, 739 F. Supp. 2d 927, 935–36 (E.D. Va. 2010) (denying defendant's motion to dismiss a database owner's reverse passing off claim brought against a screen scraper to the extent that the plaintiff “does not assert that Eventbrite has passed off its ideas as its own, but rather than Eventbrite has re-branded and re-packaged its product (the CSN venue database) and sold it as its own.”); *Cable v. Agence France Presse*, 728 F. Supp. 2d 977, 981 (N.D. Ill. 2010) (denying defendant's motion to dismiss where the plaintiff alleged that Agence France Presse took plaintiff's photos and repackaged them as its own in its own database, without revision); *Experian Marketing Solutions, Inc. v. U.S. Data Corp.*, No. 8:09cv24, 2009 WL 2902957 (D. Neb. Sept. 9, 2009) (holding that plaintiff's claim, involving an unauthorized copy of consumer data files, was not preempted because the database constituted a tangible good); see generally *infra* § 6.12[1] (analyzing case law).

verse passing off potentially could be maintained.⁸

In *Register.com, Inc. v. Verio, Inc.*,⁹ the Second Circuit declined in part to rule as moot and reversed in part the lower court's holding that the plaintiff was likely to prevail on its Lanham Act claims. In that case, Verio had used bots to repeatedly copy from Register.com's website the WHOIS database (which lists the contact information for all domain name registrants in Top Level Domains for which Register.com acts as a registrar).¹⁰ Verio used this information to contact new registrants soliciting their interest in services that it offered in competition with Register.com and its co-brand and private label partners. The court ruled that the portion of the injunction barring Verio from using Register.com's marks was moot because Verio had already agreed not to use Register.com's marks any longer and Register.com had agreed to modify the preliminary injunction to delete this part of the order. The other aspect of the preliminary injunction barred the phrasing of solicitations that the district court found misleading but which did not include any reference to Register.com or its marks, and which the Second Circuit therefore found was neither false nor misleading.¹¹

In *Facebook, Inc. v. Power Ventures, Inc.*,¹² the court denied defendant's motion to dismiss plaintiff's Lanham Act claim where the plaintiff alleged that the defendants—operators of a website that used screen scraping tools to allow users to view email and social network accounts from a single location—sent Facebook users a screen shot advertising its service, which displayed Facebook's mark and appeared to have originated with or have been endorsed by Facebook.

In *Associated Press v. All Headline News Corp.*,¹³ the court granted the defendants' motion to dismiss plaintiff's Lanham

⁸See 15 U.S.C.A. § 1125(a); see generally *infra* § 6.12.

⁹*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393 (2d Cir. 2004).

¹⁰The WHOIS database and the roles of domain name registries and registrars are described in detail in chapter 7.

¹¹*Register.com, Inc. v. Verio, Inc.*, 356 F.3d 393, 440–43 (2d Cir. 2004). The telemarketing script referred to the person's recent registration of a domain name but did not name Register.com and included the truthful representation that the caller worked for Verio.

¹²*Facebook, Inc. v. Power Ventures, Inc.*, 91 U.S.P.Q.2d 1430, 2009 WL 1299698 (N.D. Cal. May 11, 2009).

¹³*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454

Act claims, in a case where the plaintiff alleged that defendants copied AP breaking news reports and reprinted its news stories on their All Headline News (AHN) website, either as AP reports or AHN content. In rejecting plaintiff's conclusory allegations, Judge Castel wrote that "trademark law 'generally does not prevent one who trades a branded product from accurately describing it by its brand name, so long as the trader does not create confusion by implying an affiliation with the owner of the product.'"¹⁴ Likewise, Judge Castel dismissed AP's false advertising claim, concluding that AHN's description of its site as a "news service" even though AHN did not do any original news reporting was not actionable because the term "news service" did not lend itself to absolute criteria and defendants' characterization of their service appeared to be permissible puffery.¹⁵ The court denied defendants' motion to dismiss AP's state law unfair competition claim based on passing off, however, finding that the plaintiff had stated a claim by alleging that AHN passed off AP content as its own.¹⁶

By contrast, in *Cable v. Agence France Press*,¹⁷ the court denied defendants' motion to dismiss plaintiff's reverse passing off claim (as well as its copyright management information claim under the DMCA)¹⁸ where the plaintiff alleged that Agence France Press removed his name and copyright notice and the link to his website ("Photos © 2009 wayne cable, selfmadephoto.com") and republished his photos in its online photo database, ImageForum.¹⁹

Where a party copies material without attribution but modifies it in some way, a reverse passing off claim may not be brought.²⁰ Where attribution along with copyright management information is removed, however, a database

(S.D.N.Y. 2009).

¹⁴*Associated Press v. All Headline News Corp.*, 608 F. Supp. 2d 454, 462 (S.D.N.Y. 2009), quoting *Dow Jones & Company, Inc. v. International Securities Exchange, Inc.*, 451 F.3d 295, 308 (2d Cir. 2006) (quotation marks omitted).

¹⁵608 F. Supp. 2d at 463; see generally *infra* § 6.12[5] (advertising, including puffery).

¹⁶608 F. Supp. 2d at 464.

¹⁷*Cable v. Agence France Presse*, 728 F. Supp. 2d 977 (N.D. Ill. 2010).

¹⁸17 U.S.C.A. § 1202; see *supra* §§ 5.07[2], 4.21.

¹⁹For further discussion of this aspect of the case, see *infra* § 6.12[2].

²⁰*Dastar Corp. v. Twentieth Century Fox Film Corp.*, 539 U.S. 23

owner may be able to maintain a claim under the DMCA.²¹

In addition to claims under the Lanham Act, database owners may be able to assert equivalent claims for trademark infringement, dilution, false designation of origin, false advertising, passing off or unfair competition under state law.²²

Except for dilution and certain false advertising claims, to prevail under the Lanham Act, a plaintiff must establish likelihood of confusion.²³ Mere copying is not actionable. Thus, while the Lanham Act may provide remedies in certain cases where material from a database is republished elsewhere, it does not proscribe access to data or copying.

5.09 Trade Secret Protection

Most commercially available databases typically are not treated as trade secrets. Where the contents of a database constitute trade secrets, however, a database owner potentially may seek injunctive relief and damages for trade secret misappropriation provided the defendant in fact misappropriated the data and did not merely obtain it lawfully from a third party (in some jurisdictions, without knowledge that the material is a trade secret). Assuming that the contents of the database qualify for trade secret protection and are adequately protected as such, a database owner potentially could state a claim.¹ Misappropriation of trade secret claims generally are not preempted so long as an extra element—such as misappropriation or breach of a duty or trust—is alleged.²

(2003); *see generally infra* § 6.12.

²¹17 U.S.C.A. § 1202; *see generally supra* § 5.07[2].

²²*See infra* §§ 6.04 (state trademarks), 6.11[7] (state law dilution), 6.12[6] (unfair competition); *supra* § 5.04 (misappropriation and unfair competition).

²³*See infra* §§ 6.08, 6.11, 6.12[5].

[Section 5.09]

¹*See infra* chapter 10.

²*See, e.g., Dun & Bradstreet Software Services, Inc. v. Grace Consulting, Inc.*, 307 F.3d 197, 218 (3d Cir. 2002) (holding that a misappropriation claim based on breach of a duty or trust would not be preempted, while one based solely on copying would be preempted), *cert. denied*, 538 U.S. 1032 (2003); *Huckshold v. HSSL, LLC*, 344 F. Supp. 2d 1203 (E.D. Mo. 2004) (holding trade secret misappropriation and breach of contract claims not preempted where the plaintiff alleged that the defendant owed a duty

If the content of a database is a trade secret, other state law claims potentially may be preempted in states that have enacted section 7 of the Uniform Trade Secrets Act.³ Section 7 preempts conflicting tort, restitutionary, and other state laws providing civil remedies for misappropriation of a trade secret but does not preempt contractual remedies, whether or not based on misappropriation, and other civil remedies that are not based on misappropriation of a trade secret.⁴ A copy of the UTSA is reprinted in the Appendix to chapter 10.

By contrast, a claim brought under the federal Defend Trade Secrets Act (DTSA)⁵ would not preempt other state claims that may be brought in addition to or instead of a claim under the DTSA.⁶ It likewise does not limit the avail-

to protect the confidentiality of plaintiffs' trade secrets and breached its contract by allowing a third party to copy the software in violation of their agreement (and was not merely a claim that the defendant itself copied the software), which thus involved an extra element, but finding plaintiff's tortious interference claim preempted where the only element needed to be shown to establish liability was copying); *see generally supra* § 4.18[1] (analyzing copyright preemption).

³As of October 2018, 48 states—every state except New York and North Carolina—as well as the District of Columbia, Puerto Rico and the U.S. Virgin Islands had adopted a version of the Uniform Trade Secrets Act (UTSA). *See infra* § 10.02.

⁴UTSA § 7; *infra* § 10.17 (addressing UTSA preemption). A copy of the UTSA is reprinted in the appendix to chapter 10. Section 7 preemption—in some of the jurisdictions where it has been enacted into state law—preempts claims regardless of whether the underlying information is a trade secret, and may have particular application to claims involving business information and data. *See, e.g., Heller v. Cepia, LLC*, No. C 11-01146 JSW, 2012 WL 13572, at *7 (N.D. Cal., Jan. 4, 2012) (dismissing claims for common law misappropriation, conversion, unjust enrichment, and trespass to chattels, because these claims, “premised on the wrongful taking and use of confidential business and proprietary information, regardless of whether such information constitutes trade secrets, are superseded by the CUTSA.”); *see generally infra* § 10.17 (discussing conflicting lines of cases on whether a claim is preempted even if based on information that may not be protectable as a trade secret).

⁵18 U.S.C.A. §§ 1830 to 1839; *see generally infra* § 10.12[2].

⁶18 U.S.C.A. § 1838. That section states that except as set forth in section 1833(b) (which provides immunity from liability for the confidential disclosure of a trade secret to the government or an attorney for the purpose of reporting a violation of law, in a court filing under seal or in connection with an anti-retaliation lawsuit, provided the material is filed under seal and is not disclosed except pursuant to court order), the DTSA “shall not be construed to preempt or displace any other remedies, whether civil or criminal, provided by United States Federal, State, commonwealth,

ability of remedies under other federal statutes,⁷ such as the Computer Fraud and Abuse Act.⁸

Claims asserted against an interactive computer service provider for merely hosting database content alleged to incorporate trade secrets (as opposed to direct conduct by the site or service itself) may be preempted by the Good Samaritan exemption of the Telecommunications Act of 1996, also known as the Communications Decency Act, at least in the Ninth Circuit⁹ for state law claims, and generally under the DTSA.

Claims for trade secret misappropriation are analyzed in chapter 10.

5.10 EU Database Directive

5.10[1] Overview

The European Parliament and the Council of the European Union enacted the EU Database Directive in 1996, which compels member states to afford fifteen-year *sui generis* protection for databases where there has been “qualitatively and/or quantitatively a substantial investment in either the obtaining, verification, or presentation of the contents.”¹ A database is defined as “a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other

possession, or territory law for the misappropriation of a trade secret, or to affect the otherwise lawful disclosure of information by any Government employee under . . .” the Freedom of Information Act, 5 U.S.C.A. § 552. *See* 18 U.S.C.A. § 1838; H.R. Rep. 114-529, 114th Cong. 2d Sess. 5 (2016) (stating that the DTSA does not preempt variations of the UTSA enacted in 48 states but “offers a complementary Federal remedy . . .”).

⁷See 18 U.S.C.A. § 1838; *see generally infra* § 10.12[2].

⁸18 U.S.C.A. § 1030; *see generally infra* § 44.08.

⁹47 U.S.C.A. § 230(c); *see generally infra* § 37.05[5][B].

[Section 5.10[1]]

¹Commission Directive 7934/95 of March 11, 1996, on the Legal Protection of Databases, 1996 O.J. (L077) 20, at chapter III (“Database Directive”), Art. 7(1). This provision grew out of a 1992 European Economic Community initiative intended as a rejection of the analysis employed in *Feist Publications, Inc. v. Rural Telephone Service Co.*, 499 U.S. 340 (1991). *See* Jack E. Brown, “Proposed International protection of Electronic Databases,” *The Computer Law*, Jan. 1997, at 17, 18. A copy of the Directive may be obtained at <http://guagua.echo.lu.legal/en/ipr/database>.

means.”²

Rights to the *sui generis* protection created by the Directive are granted based on the location of the owner or author of the work, not the location where it was created. As discussed in subsection 5.10[3] below, the Directive potentially allows database makers or rights holders to extend protection indefinitely.

Sui generis rights are granted in addition to copyright protection, which may be available to database authors within the European Union based on original selection or arrangement.³ Even under EU copyright law, however, case law has established that skill and labor (like the U.S. “sweat of the brow” doctrine) are insufficient to confer copyright protection on a database and it is the selection and arrangement of data in the database, not in the creation of the data, that determines whether copyright protection may be available for a database. The European Court of Justice has explained that the “criterion of originality is satisfied when, through the selection or arrangement of the data which it contains, its author expresses his creative ability in an original manner by making free and creative choices . . . and thus stamps his ‘personal touch.’” This criterion is “not satisfied when the setting up of the database is dictated by technical considerations, rules or constraints which leave no room for creative freedom.”⁴

The Database Directive also does not extend protection to computer programs used in the making or operation of a database, however, which is the subject of a separate EU directive.⁵

Despite its potentially broad coverage, court opinions to date have construed the scope of the Database Directive

²Database Directive Art. 1(2); Preamble ¶ 17. Elements of a database, however, need not be “physically stored in an organized manner.” Jack E. Brown, “Proposed International protection of Electronic Databases,” *The Computer Law*, Jan. 1997, at 21. Excluded from the definition of a database are recordings or audiovisual, cinematographic, literary or musical works. Jack E. Brown, “Proposed International protection of Electronic Databases,” *The Computer Law*, Jan. 1997, at 17.

³The EU Copyright Directive is separately addressed in chapter 4. See *supra* § 4.20.

⁴See *Football Dataco Ltd. v. Yahoo UK Ltd.*, Case C-604/10 (European Court of Justice 2012).

⁵See Database Directive Arts. 1, 2.

more narrowly.⁶

The legal protection conferred by the Database Directive is not applicable to a database which is neither eligible for copyright nor *sui generis* protection. Even where a database is entitled to *sui generis* protection, the Directive establishes mandatory rights for lawful users of a database (which are akin to fair use rights).⁷ These rights can be limited by contract, however, if permitted under the applicable national law.⁸

5.10[2] Copyright Protection for Databases

The Directive was intended in part to harmonize copyright protection for databases within the European Union. The EU Database Directive compels protection of databases under the copyright laws of member states which “by reason of the selection or arrangement of their contents, constitute the author’s own intellectual creation”¹ No other criteria—beyond an author’s selection or arrangement—may be applied by EU member countries in granting copyright protection to databases.² The protection afforded by the Directive, however, does not extend to the contents of protected databases and must be provided “without prejudice to any rights subsisting in those contents”³

The Directive recognizes an author’s exclusive rights to reproduction, translation, adaptation, arrangement and “any other alteration,” public distribution (subject to first sale within the community) and “communication, display or performance to the public,” as well as reproduction, distribution, communication, display or performance to the public as a result of translation, adaptation, arrangement or other

⁶See *British Horseracing Board Ltd. v. William Hill Organization Ltd.*, Case C-203/02 (European Court of Justice 2004); *Fixtures Mktg, Ltd. v. Oy Veikkaus Ab*, Case C-46/02 (European Court of Justice 2004).

⁷See Database Directive Arts. 6(1), 8, 15.

⁸See *Ryanair Ltd v. PR Aviation BV*, Case C-30/14 (European Court of Justice 2015).

[Section 5.10[2]]

¹Database Directive Art. 3(1).

²See Database Directive Art. 3(1); *Football Dataco Ltd. v. Yahoo UK Ltd.*, Case C-604/10 (European Court of Justice 2012).

³Database Directive Art. 3(2).

alteration.⁴ While there is no express fair use provision, the Directive allows member states, at their option, to allow for exceptions traditionally recognized under national law or use “for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.”⁵ Member states also may allow for exceptions “for the purposes of public security” or “an administrative or judicial procedure.”⁶ These exceptions are narrower than the U.S. fair use defense.⁷

5.10[3] *Sui Generis* Protection

5.10[3][A] In General

In addition to copyright protection, the EU Database Directive compels member states to provide *sui generis* protection for at least fifteen years from the date of completion of the database¹ to the “maker of a database”² who can show “a substantial investment in either the obtaining, verification or presentation of the contents” to “prevent extraction³ and/or re-utilization⁴ of the whole or of a substantial

⁴See Database Directive Art. 5.

⁵See Database Directive Art. 6.

⁶See Database Directive Art. 6(2)(c).

⁷See *supra* § 4.10.

[Section 5.10[3][A]]

¹See Database Directive Art. 10(1).

²The term “maker of a database” is not defined, but presumably is not necessarily the same entity as the author in whom copyright protection under the Directive vests. Although maker is not defined, the making of a database is said to require “the investment of considerable human, technical and financial resources . . .” See Database Directive Art. 10(1), Preamble ¶ 7. By extension, a maker would be a person or entity that invests such resources. According to one commentator, the term could be broad enough to include companies that provide, but do not themselves compile, data. See David Mirchin, “EU Database Directive Has Global Ramifications”, Nat. L.J., June 9, 1997.

³*Extraction* is defined as “the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form.” Database Directive Art. 7(2)(a).

⁴*Re-utilization* means “any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission.” Database Directive Art. 7(2)(b). First sale within the community, however, exhausts

part of the contents of that database.”⁵ Whether a substantial investment has been made, or a substantial part . . . extracted or re-utilized, may be evaluated qualitatively and/or quantitatively.⁶ The *sui generis* database rights created by the Directive may be transferred, assigned or granted by license.⁷ These rights exist independently of any copyright protection which may be available for the database or its contents.⁸

The Directive contains a limited right akin to fair use under U.S. law. Lawful users must be allowed to extract and/or re-utilize insubstantial parts of the contents of a database (judged qualitatively and/or quantitatively).⁹ This right is tempered, however, by the requirement that member states prohibit “[t]he repeated and systematic extraction and/or re-utilization of insubstantial parts of the contents of the database implying acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database.”¹⁰ The right to insubstantial extraction or re-utilization, while perhaps more limited, may be easier to apply than fair use under U.S. law, which is determined by a balancing test that focuses on other factors beyond merely the amount and nature of the portion copied.¹¹

Member states are further permitted to allow “extraction for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.”¹² Exceptions also are recognized for extraction or re-utilization “for the purposes of public security or an administrative or judicial procedure.”¹³

the right to control resale within the EU. Database Directive Art. 7(2)(a).

⁵See Database Directive Art. 7(1).

⁶See Database Directive Art. 7(2)(a). The term substantial, however, is not defined.

⁷See Database Directive Art. 7(3).

⁸See Database Directive Art. 7(4).

⁹Database Directive Art. 8.

¹⁰Database Directive Art. 7(5).

¹¹See *supra* § 4.10.

¹²Database Directive Art. 9(b).

¹³Database Directive Art. 9(c).

5.10[3][B] Territorial Scope of Protection

Protection is available under the EU Database Directive based on the residency of the database owner, not the country or territory where it is created. Rights under the Directive are granted to makers or rightholders who are nationals of an EU member state or who have their habitual residence within the European Union.¹ A company or firm formed in accordance with the laws of a member state may benefit from the Directive, but only if (1) its “central administration or principal place of business” is located within the EU or (2) its “registered office” is located within the EU and “its operations . . . [are] genuinely linked on an ongoing basis with the economy of a Member State.”² The Council also is authorized to extend protection to databases made in third countries by persons or entities not entitled to benefit from the Directive, where reciprocal protections are recognized.³

5.10[3][C] Term of Protection

The term of *sui generis* protection, while stated as fifteen years, actually is always longer than fifteen years, and potentially may be extended indefinitely.¹ While a database is protected as of the date of completion, the term of protection only expires fifteen years from the first day of January of the year following the date when the database was completed.²

The term of protection may be extended by making the database available to the public “in any manner” before the expiration of the initial term. In such case, protection shall be extended to fifteen years from the first day of January of the year following the date when the database was first made available to the public.³ A database made available to the public just prior to the expiration of its initial term therefore

[Section 5.10[3][B]]

¹Database Directive Art. 11(1).

²Database Directive Art. 11(2).

³Database Directive Art. 11(3), Preamble 56.

[Section 5.10[3][C]]

¹The term for copyright protection in a database, however, is not extended by the Directive. *See* Database Directive Art. 2(c).

²*See* Database Directive Art. 10(1).

³*See* Database Directive Art. 10(2).

could enjoy potentially up to almost thirty-one years of protection (depending on when during the year preceding the commencement of the first fifteen-year term the database was completed).

As a practical matter, protection for a database could be extended indefinitely. If a substantial change is made to the contents of a database, including changes that result “from the accumulation of successive additions, deletions or alterations, which would result in the database being considered to be a substantial new investment,” then the database would be entitled to a new term of protection.⁴ The substantiality of any changes or investments may be evaluated qualitatively or quantitatively.⁵ Given the extent of revisions required to keep most commercial databases current, it seems likely that owners could extend the term of protection indefinitely, and then still be entitled to over fifteen years of protection after the database is retired from use and no longer updated.

Limited retroactive protection also may exist for certain databases. Specifically, protection must be recognized for any database completed after December 31, 1982, that otherwise complied with the requirements for protection under the Directive on January 1, 1998.⁶ Retroactive protection, however, must be granted “without prejudice to any acts concluded and rights acquired” prior to January 1, 1998.⁷

5.11 Sample Injunction Order

5.11[1] Overview

The following is a form created from the actual order for preliminary and permanent injunctive relief entered by the court in *Craigslist, Inc. v. Naturemarket, Inc.*,¹ where the court granted a default judgment to Craigslist on claims for

⁴See Database Directive Art. 10(3).

⁵Database Directive Art. 10(3).

⁶See Database Directive Art. 14(3). The initial term of protection would expire fifteen years from the first of January in the year following the date on which the database was completed. See Database Directive Art. 14(5).

⁷See Database Directive Art. 14(4).

[Section 5.11[1]]

¹*Craigslist, Inc. v. Naturemarket, Inc.*, 694 F. Supp. 2d 1039 (N.D. Cal. 2010).

copyright infringement (based on exceeding the scope of the license for Craigslist's website, set forth in its TOU), DMCA violations (for circumvention of CAPTCHA and other security measures), the Computer Fraud and Abuse Act (for exceeding authorized access as defined by Craigslist's TOU for purposes of employing, implementing and updating software to allow for automated postings on Craigslist), California Penal Code § 502 (for knowingly accessing a computer without permission and causing damage), trademark infringement (in connection with its use of the Craigslist mark in sponsored links),² common law trademark infringement (based on use of the mark in advertising defendants' services and auto posting software), breach of contract (the TOU agreement), inducing breach of contract and intentional interference with contractual relations, and fraud. In that case, the defendants sold software products that allowed users to automatically post material to Craigslist's site, in violation of its TOU, and harvest email addresses from the site. Because of the number of claims on which judgment was entered, the order may provide a useful form. Like any form, it must be tailored to the specific facts and claims at issue in a given case.

5.11[2] FORM

Upon consideration of plaintiff's motion for injunctive relief and defendants' opposition, the court hereby orders that defendants —, their employees, representatives, agents and all persons or entities acting in concert with them, are preliminarily and permanently enjoined from:

(a) manufacturing, developing, creating, adapting, modifying, exchanging, offering, distributing, selling, providing, importing, trafficking in, or using any automated device or computer program (including but not limited to, any technology, product, service, device, component, or part thereof) that enables postings on — (the "Website") without each posting being entered manually;

(b) manufacturing, developing, creating, adapting, modifying, exchanging, offering, distributing, selling, providing, importing, making available, trafficking in, or using content that uses automated means (including, but not limited to, spiders, robots, crawlers, data mining tools, and data scrap-

²See generally *infra* § 9.11 (analyzing the law of sponsored links).

ing tools) to download or otherwise obtain data from the Website;

(c) copying, distributing, displaying, creating derivative works or otherwise using protected elements of plaintiff's copyrighted website (located at www._), including but not limited to, the website's post to classifieds, account registration and account log in expressions and compilations, and from inducing, encouraging, causing or materially contributing to any other person or entity doing the same;

(d) circumventing technological measures that control access to plaintiff's copyrighted website and/or portions thereof (including, but not limited to, CAPTCHAs and RECAPTCHAs), and from inducing, encouraging, causing or materially contributing to any other person or entity doing the same;

(e) manufacturing, developing, creating, adapting, modifying, exchanging, offering, selling, distributing, providing, importing, trafficking in, or using technology, products, services, devices, components, or parts thereof, that are primarily designed or produced for the purpose of circumventing technological measures and/or protection afforded by technological measures that control access to plaintiff's copyrighted website and/or portions thereof, and from inducing, encouraging, causing or materially contributing to any other person or entity doing the same;

(f) accessing or attempting to access plaintiff's computers, computer systems, computer network, computer programs, and data, without authorization or in excess of authorized access, including, but not limited to, creating accounts or posting content on the Website, and from inducing, encouraging, causing, materially contributing to, aiding or abetting any other person or entity to do the same;

(g) manufacturing developing, creating, adapting, modifying, exchanging, offering, selling, distributing, providing, importing, trafficking in, purchasing, acquiring, transferring, marketing or using any program, device, or service designed to provide an automated means of accessing the Website, automated means of creating accounts on the Website or with plaintiff, or automated means of posting ads or other content on the Website, including, but not limited to, any program, device, or service that is, in whole or in part, designed to circumvent security measures on the Website;

(h) repeatedly posting the same or similar content on the

Website, posting the same item or service in more than one category on the Website, posting the same item or service in more than one geographic area on the Website, and from inducing, encouraging, causing, assisting, aiding, abetting or contributing to any other person or entity doing the same;

(i) posting ads on behalf of others, causing ads to be posted on behalf of others, and accessing the Website to facilitate posting ads on behalf of others;

(j) using, offering, selling or otherwise providing a third-party agent, service, or intermediary to post content to the Website;

(k) misusing or abusing plaintiff, the Website and plaintiff's services in any way, including, but not limited to, violating plaintiff's TOU;

(l) accessing or using the Website for any commercial purpose whatsoever, and;

(m) using the — mark and any confusingly similar designation in Internet advertisements and otherwise in commerce in any manner likely to confuse consumers as to their association, affiliation, endorsement or sponsorship with or by the plaintiff.

IT IS SO ORDERED

JUDGE

5.12 Anti-Scraping Measures Pursuant to the Cybersecurity Information Sharing Act

The Cybersecurity Information Sharing Act (CISA)¹ permits companies, including database owners, to take certain measures to protect the security of their systems. It also limits a database owner's ability to treat as a cybersecurity threat a Terms of Use violation.

The Act permits companies to take defensive measures to protect their information systems.² However, the Act narrowly circumscribes what constitutes a *defensive measure*. A *defensive measure* is defined narrowly as one that addresses

[Section 5.12]

¹6 U.S.C.A. §§ 1501 to 1510.

²6 U.S.C.A. § 1503(b)(1); *infra* § 27.04[1.5].

the purpose of the statute.³ The statute also expressly excludes the possibility of relying solely on Terms of Use or other consumer license agreements as a basis for taking a defensive measure against a cybersecurity threat.⁴ Thus, while CISA empowers database owners to take added measures to protect the security of their information systems, it also limits their ability to treat contractual violations as cybersecurity threats under the Act.

CISA is analyzed in section 27.04[1.5] of chapter 27.

5.13 Checklist of Potential Ways to Protect Database Content

In General

- Database owners should restrict access and use by contract
- Database owners should employ technological means (including security, access controls and copy protection mechanisms) to block access to material
- Database owners should organize their databases and materials to maximize potential protection under copyright, trademark, trade secret and patent laws
- Database owners should set no scraping tags pursuant to the Robots Exclusion Standard and consider taking other technical measures to restrict scraping

Copyright

- Is the database, as a compilation, entitled to protection based on the selection, arrangement or organization of the compilation?
- Are the individual components of a database independently protectable (such as photos, articles, music and videos) or merely unprotectable data (or material otherwise in the public domain such as court opinions)?
 - If the “contributions” to the collective work are separately protectable, who owns the copyrights to these works?
 - Does the database owner have rights to the underlying components of the database (ei-

³See 6 U.S.C.A. § 1503(b)(2); *infra* § 27.04[1.5].

⁴6 U.S.C.A. § 1501(5)(B); *infra* § 27.04[1.5].

ther through ownership or a license) or is there a potential *Tasini* problem?¹

- Is an implied license defense available?
- Does the extent of copying rise to the level of substantial similarity or virtual identity?
- Does the copying qualify as fair use intermediate copying?

Contract

- Is there privity of contract between the database owner and the party against whom the agreement is sought to be enforced?
 - If not, can a claim be asserted for tortious interference with contract or interference with prospective economic advantage, based on the third party providing the means for its users to breach their contracts (or could the database owner allege that it is an intended third party beneficiary of the contract)?²
- Are the terms presented in a manner in which they are likely to be deemed enforceable?³
 - Agreements presented as click-through contracts are more likely to be enforced than terms that are merely posted on a site
 - Is the agreement susceptible to being challenged as unconscionable?⁴
- May the agreement be characterized as a license or mere contract?
 - Is the compilation protectable under copyright law?
 - Are licensees given access to software or other protectable material in addition to factual data or other material in the public domain?

[Section 5.13]

¹See *supra* § 5.01.

²See *supra* § 5.03[5].

³See *infra* § 21.03; see generally *infra* chapters 21, 22 (analyzing Terms of Use and enforceable unilateral contracts). Some database companies obtain signed, written contracts, or negotiate the terms of an agreement, which eliminates the formation issues addressed here.

⁴See *infra* §§ 21.05 (unconscionability and checklist), 22.05[2][M] (arbitration and class action provisions and unconscionability), 22.05[4] (draftsmanship).

- Is the agreement susceptible to being challenged under the copyright misuse doctrine?⁵
- Do the terms of the contract adequately protect the database owner? Common terms include prohibitions on:
 - Commercial use of the database or website, including duplication and downloading of content;
 - The use of bots, scripts, executable code, intelligent agent software, spiders, crawlers or other automated means of accessing the site or extracting data;
 - Accessing the site more than a set number of times in a given time period;
 - Taking any action that imposes an unreasonable burden or disproportionately uses system resources;⁶
 - Using the site in any manner not expressly licensed; and
 - Use after termination of the agreement
- Does the agreement include carve-outs that could allow certain uses, either because what is being licensed is narrowly defined or particular uses or data (such as material in the public domain) are excluded?⁷

Common Law Misappropriation

- Does the claim allege an extra element beyond mere copying?
- Was copying undertaken to provide information sooner than when users might otherwise receive it (i.e., “hot news”)?⁸

Trespass

- Was access unauthorized based on contractual terms or notice?
- May a claim for trespass to chattels be brought for trespass to an intangible under applicable statute law?
- If so, what level of damage can be shown to server capacity or system resources?

⁵See *supra* § 5.03[1].

⁶See *supra* § 5.03[2].

⁷See *supra* § 5.03[3].

⁸See *supra* § 5.04.

- Injury to a business is not recoverable; the damage must be to the chattel⁹

Conversion

- Was data destroyed or deleted, or were intangible assets taken, or was data merely copied?

Computer Fraud and Abuse Act

- Was access unauthorized (or was authorized access exceeded) based on contractual terms or notice?
 - If authorization was given, has it been revoked?
- Can \$5,000 in damages be shown?¹⁰

DMCA

- Did access involve circumvention of access controls?
- Was copyright management information (CMI) deleted or false CMI added?¹¹

BOTS Act

- Did someone circumvent a security measure, access control system, or other technological control or measure that is used by a ticket issuer to enforce posted event ticket purchasing limits or to maintain the integrity of posted online ticket purchasing order rules?

Lanham Act

- Does the use of information from the database include use of the database owner's marks?¹²
 - Could the use be deemed a nominative fair use (use not in a trademark sense)?
 - Does the use involve intermediate copying?
 - Does the use involve passing off goods or services as belonging to the database owner or falsely suggesting sponsorship, affiliation or endorsement by the owner?
 - Is the use likely to cause confusion or dilution?

Trade Secret

- Was confidential information entitled to protection as a

⁹See *supra* § 5.05.

¹⁰See *supra* § 5.06.

¹¹See *supra* § 5.07.

¹²See *supra* § 5.08.

trade secret misappropriated by someone under a duty to keep it confidential?¹³

Cybersecurity Information Sharing Act

- Is the database owner protecting its information systems pursuant to the Act?¹⁴

Preemption Considerations

- Are potential claims against third parties preempted by the Good Samaritan Exemption to the Communications Decency Act, 47 U.S.C.A. § 230(c)?¹⁵
- Are potential claims preempted by the Copyright Act,¹⁶ the Lanham Act,¹⁷ section 7 of the Uniform Trade Secrets Act (where enacted as state law)¹⁸ or the Patent Act?¹⁹

5.14 Checklist for Ethical Scraping Practices

The following is a checklist of measures to undertake to mitigate the risk of liability under U.S. law for scraping third party content without permission:¹

- Don't access a site whose TOU or EULA prohibit scraping or non-commercial use of the site;²
- Don't scrape any site whose owner or operator has notified you that you are not permitted to access the site or to scrape content or that your access rights have been limited or revoked;

¹³See *supra* § 5.09.

¹⁴See *supra* § 5.12.

¹⁵See *infra* § 37.05.

¹⁶See *supra* § 5.04.

¹⁷See *supra* § 5.08.

¹⁸See *supra* § 5.09; *infra* § 10.17.

¹⁹See *supra* § 5.04[3].

[Section 5.14]

¹This list reflects general “best practices” but is neither a comprehensive list for avoiding exposure nor a statement of legal principles that if violated necessarily would result in liability. As underscored throughout this chapter, what is lawful in the area of database protection and screen scraping depends on the nature of the database, the type of information copied or used by a third party, how it was accessed, and what is being done with it. Businesses engaged in screen scraping may need to make many nuanced decisions to structure their affairs to avoid liability.

²See *supra* § 5.03.

- Scraped content, where possible, should be used for internal analysis only;
- Original and creative images or text (as opposed to facts), which may be entitled to copyright protection, should not be used commercially or copied to websites or other publicly available documents or locations without permission (unless a fair use);³
- Scraping should be undertaken during off hours and in a manner to ensure that there is no substantial impairment to the target site's server capacity;⁴
- Avoid using scraped data to "scoop" competitors with time sensitive information;⁵
- Do not circumvent technical measures or access controls to scrape content⁶ and do not scrape material from private locations (or non-public locations where access is restricted) or seek to gain access to a proprietary location to scrape content under false pretenses;
- Don't circumvent a security measure, access control system, or other technological control or measure that is used by a ticket issuer to enforce posted event ticket purchasing limits or to maintain the integrity of a posted online ticket purchasing order rule;⁷ and
- Honor the Robot Exclusion Standard and similar protocols or tags that alert third parties not to access a site using bots or intelligent agents.⁸

5.15 Managing the IP Risks of Artificial Intelligence By Contract

Artificial intelligence raises a host of policy questions. From an IP perspective, however, when AI is deployed, liability likely would fall on the person or entity that empowered the agent, much as liability falls on a principal for the conduct of an agent. If software agents using AI scrape particular data, it would still likely be the responsibility of the entity that programmed the agent, much in the same way

³See *supra* § 5.02. If copyrighted material is published, copyright management information (CMI) and logos or other trademark-protected material generally should not be removed. See *supra* §§ 5.07[2], 5.08.

⁴See *supra* § 5.05[1].

⁵See *supra* § 5.04[1].

⁶See *supra* § 5.07[1].

⁷See *supra* § 5.07[3].

⁸See *supra* § 5.05[1].

that a business is responsible for any misconduct by an employee acting within the scope of employment.

Most IP-related AI issues today can be addressed by contract. For example, if one party provides tools to deploy an intelligent avatar using the attributes of a third party, permission would be required to use those attributes.¹ Likewise, it would be prudent for the parties to determine in advance what their respective rights will be with respect to the creative output of the intelligent agent. The same is true for developers of intelligent software. Publishers can elect to provide a tool that creates intellectual property owned by a customer or the publisher, or both, with exclusive or non-exclusive license rights carved out.

There are open questions about the extent to which AI can generate intellectual property. Copyright law, for example, only extends protection to original and creative expression.² Whether original expression created through artificial intelligence is the work product of the agent or the human who created the agent remains to be fully fleshed out. At a minimum, however, only humans have standing to sue for copyright infringement.³

Liability for agents empowered by AI may be varied by contract or subject to indemnification obligations or insurance, but the party that deployed an agent in most cases will be liable for the actions it directed, preprogrammed, or enabled through AI. With respect to screen scraping by bots, whether intelligent agents or agents using artificial intelligence, the liability regime established by contract, common law rules such as trespass and statutes such as the Computer Fraud and Abuse Act will largely be the same.

[Section 5.15]

¹Rights of publicity are analyzed in chapter 12.

²*See supra* § 5.02.

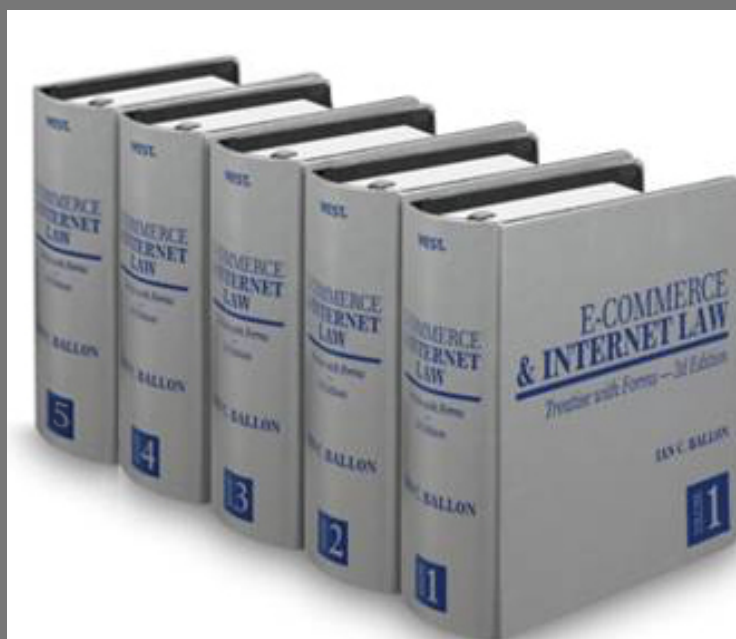
³*See, e.g., Naruto v. Slater*, 888 F.3d 418, 426 (9th Cir. 2018) (holding, in the Monkey Selfie case, that “animals other than humans . . . lack statutory standing to sue under the Copyright Act.”).

E-COMMERCE & INTERNET LAW: TREATISE WITH FORMS 2D 2019

Ian C. Ballon

**NEW AND
IMPORTANT
FEATURES
FOR 2019
NOT FOUND
ELSEWHERE**

**THE PREEMINENT
INTERNET AND
MOBILE LAW
TREATISE FROM A
LEADING INTERNET
LITIGATOR — NOW A
5 VOLUME SET!**



To order call **1-888-728-7677**
or visit **legalsolutions.thomsonreuters.com**

Key Features of E-Commerce & Internet Law

- ◆ The California Consumer Privacy Act, GDPR, California IoT security statute, Vermont data broker registration law, Ohio safe harbor statute and other important privacy and cybersecurity laws
- ◆ Understanding conflicting law on mobile contract formation, unconscionability and enforcement of arbitration and class action waiver clauses
- ◆ The most comprehensive analysis of the TCPA's application to text messaging and its impact on litigation found anywhere
- ◆ Complete analysis of the Cybersecurity Information Sharing Act (CISA), state security breach statutes and regulations, and Defend Trade Secrets Act (DTSA) and their impact on screen scraping and database protection, cybersecurity information sharing and trade secret protection, privacy obligations and the impact that Terms of Use and other internet and mobile contracts may have in limiting the broad exemption from liability otherwise available under CISA
- ◆ Comprehensive and comparative analysis of the platform liability of Internet, mobile and cloud site owners, and service providers, for user content and misconduct under state and federal law
- ◆ Understanding the laws governing SEO and SEM and their impact on e-commerce vendors, including major developments involving internet advertising and embedded and sponsored links
- ◆ AI, screen scraping and database protection
- ◆ Strategies for defending cybersecurity breach and data privacy class action suits
- ◆ Copyright and Lanham Act fair use, patentable subject matter, combating genericide, right of publicity laws governing the use of a person's images and attributes, initial interest confusion, software copyrightability, damages in internet and mobile cases, the use of icons in mobile marketing, new rules governing fee awards, and the applicability and scope of federal and state safe harbors and exemptions
- ◆ How to enforce judgments against foreign domain name registrants
- ◆ Valuing domain name registrations from sales data
- ◆ Compelling the disclosure of the identity of anonymous and pseudonymous tortfeasors and infringers
- ◆ Exhaustive statutory and case law analysis of the Digital Millennium Copyright Act, the Communications Decency Act (including exclusions created by FOSTA-SESTA), the Video Privacy Protection Act, and Illinois Biometric Privacy Act
- ◆ Analysis of the CLOUD Act, BOTS Act, SPEECH Act, Consumer Review Fairness Act, N.J. Truth-in-Consumer Contract, Warranty and Notice Act, Family Movie Act and more
- ◆ Practical tips, checklists and forms that go beyond the typical legal treatise
- ◆ Clear, concise, and practical analysis

AN ESSENTIAL RESOURCE FOR ANY INTERNET AND MOBILE, INTELLECTUAL PROPERTY OR DATA PRIVACY/ CYBERSECURITY PRACTICE

E-Commerce & Internet Law is a comprehensive, authoritative work covering law, legal analysis, regulatory issues, emerging trends, and practical strategies. It includes practice tips and forms, nearly 10,000 detailed footnotes, and references to hundreds of unpublished court decisions, many of which are not available elsewhere. Its unique organization facilitates finding quick answers to your questions.

The updated new edition offers an unparalleled reference and practical resource. Organized into five sectioned volumes, the 59 chapters cover:

- Sources of Internet Law and Practice
- Intellectual Property
- Licenses and Contracts
- Data Privacy, Cybersecurity and Advertising
- The Conduct and Regulation of E-Commerce
- Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption
- Obscenity, Pornography, Adult Entertainment and the Protection of Children
- Theft of Digital Information and Related Internet Crimes
- Platform liability for Internet Sites and Services (Including Social Networks, Blogs and Cloud services)
- Civil Jurisdiction and Litigation

Distinguishing Features

- ◆ Clear, well written and with a practical perspective based on how issues actually play out in court (not available anywhere else)
- ◆ Exhaustive analysis of circuit splits and changes in the law combined with a common sense, practical approach for resolving legal issues, doing deals, documenting transactions and litigating and winning disputes
- ◆ Covers laws specific to the Internet and explains how the laws of the physical world apply to internet and mobile transactions and liability risks
- ◆ Addresses both law and best practices
- ◆ Comprehensive treatment of intellectual property, data privacy and mobile and Internet security breach law

Volume 1

Part I. Sources of Internet Law and Practice: A Framework for Developing New Law

- Chapter* 1. Context for Developing the Law of the Internet
 2. A Framework for Developing New Law
 3. [Reserved]

Part II. Intellectual Property

4. Copyright Protection in Cyberspace
 5. Database Protection, Screen Scraping and the Use of Bots and Artificial Intelligence to Gather Content and Information
 6. Trademark, Service Mark, Trade Name and Trade Dress Protection in Cyberspace
 7. Rights in Internet Domain Names

Volume 2

- Chapter* 8. Internet Patents
 9. Unique Intellectual Property Issues in Search Engine Marketing, Optimization and Related Indexing, Information Location Tools and Internet and Social Media Advertising Practices
 10. Misappropriation of Trade Secrets in Cyberspace
 11. Employer Rights in the Creation and Protection of Internet-Related Intellectual Property
 12. Privacy and Publicity Rights of Celebrities and Others in Cyberspace
 13. Idea Protection and Misappropriation

Part III. Licenses and Contracts

14. Documenting Internet Transactions: Introduction to Drafting License Agreements and Contracts
 15. Drafting Agreements in Light of Model and Uniform Contract Laws: UCITA, the UETA, Federal Legislation and the EU Distance Sales Directive
 16. Internet Licenses: Rights Subject to License and Limitations Imposed on Content, Access and Development
 17. Licensing Pre-Existing Content for Use Online: Music, Literary Works, Video, Software and User Generated Content Licensing Pre-Existing Content
 18. Drafting Internet Content and Development Licenses
 19. Website Development and Hosting Agreements
 20. Website Cross-Promotion and Cooperation: Co-Branding, Widget and Linking Agreements
 21. Obtaining Assent in Cyberspace: Contract Formation for Click-Through and Other Unilateral Contracts
 22. Structuring and Drafting Website Terms and Conditions
 23. ISP Service Agreements

Volume 3

- Chapter* 24. Software as a Service: On-Demand, Rental and Application Service Provider Agreements

Part IV. Privacy, Security and Internet Advertising

25. Introduction to Consumer Protection in Cyberspace
 26. Data Privacy
 27. Cybersecurity: Information, Network and Data Security
 28. Advertising in Cyberspace

Volume 4

- Chapter* 29. Email and Text Marketing, Spam and the Law of Unsolicited Commercial Email and Text Messaging

30. Online Gambling

Part V. The Conduct and Regulation of Internet Commerce

31. Online Financial Transactions and Payment Mechanisms
 32. Online Securities Law
 33. Taxation of Electronic Commerce
 34. Antitrust Restrictions on Technology Companies and Electronic Commerce
 35. State and Local Regulation of the Internet
 36. Best Practices for U.S. Companies in Evaluating Global E-Commerce Regulations and Operating Internationally

Part VI. Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption

37. Defamation, Torts and the Good Samaritan Exemption (47 U.S.C.A. § 230)
 38. Tort and Related Liability for Hacking, Cracking, Computer Viruses, Disabling Devices and Other Network Disruptions
 39. E-Commerce and the Rights of Free Speech, Press and Expression In Cyberspace

Part VII. Obscenity, Pornography, Adult Entertainment and the Protection of Children

40. Child Pornography and Obscenity
 41. Laws Regulating Non-Obscene Adult Content Directed at Children
 42. U.S. Jurisdiction, Venue and Procedure in Obscenity and Other Internet Crime Cases

Part VIII. Theft of Digital Information and Related Internet Crimes

43. Detecting and Retrieving Stolen Corporate Data
 44. Criminal and Related Civil Remedies for Software and Digital Information Theft
 45. Crimes Directed at Computer Networks and Users: Viruses and Malicious Code, Service Disabling Attacks and Threats Transmitted by Email

Volume 5

- Chapter* 46. Identity Theft

47. Civil Remedies for Unlawful Seizures

Part IX. Liability of Internet Sites and Service (Including Social Networks and Blogs)

48. Assessing and Limiting Liability Through Policies, Procedures and Website Audits
 49. The Liability of Platforms (including Website Owners, App Providers, eCommerce Vendors, Cloud Storage and Other Internet and Mobile Service Providers) for User Generated Content and Misconduct
 50. Cloud, Mobile and Internet Service Provider Liability and Compliance with Subpoenas and Court Orders
 51. Web 2.0 Applications: Social Networks, Blogs, Wiki and UGC Sites

Part X. Civil Jurisdiction and Litigation

52. General Overview of Cyberspace Jurisdiction
 53. Personal Jurisdiction in Cyberspace
 54. Venue and the Doctrine of Forum Non Conveniens
 55. Choice of Law in Cyberspace
 56. Internet ADR
 57. Internet Litigation Strategy and Practice
 58. Electronic Business and Social Network Communications in the Workplace, in Litigation and in Corporate and Employer Policies
 59. Use of Email in Attorney-Client Communications

“Should be on the desk of every lawyer who deals with cutting edge legal issues involving computers or the Internet.”

Jay Monahan

General Counsel, ResearchGate

ABOUT THE AUTHOR

IAN C. BALLON

Ian Ballon is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property and Technology Practice Group and is a litigator based in the firm's Silicon Valley and Los Angeles offices. He defends data privacy, cybersecurity breach, TCPA, and other Internet and mobile class action suits and litigates copyright, trademark, patent, trade secret, right of publicity, database and other intellectual property matters, including disputes involving Internet-related safe harbors and exemptions and platform liability.



Mr. Ballon was the recipient of the 2010 Vanguard Award from the State Bar of California's Intellectual Property Law Section. He also has been recognized by *The Los Angeles and San Francisco Daily Journal* as one of the Top 75 Intellectual Property litigators, Top Cybersecurity and Artificial Intelligence (AI) lawyers, and Top 100 lawyers in California.

In 2017 Mr. Ballon was named a "Groundbreaker" by *The Recorder* at its 2017 Bay Area Litigation Departments of the Year awards ceremony and was selected as an "Intellectual Property Trailblazer" by the *National Law Journal*.

Mr. Ballon was named as the Lawyer of the Year for information technology law in the 2019, 2018, 2016 and 2013 editions of *The Best Lawyers in America* and is listed in Legal 500 U.S., *The Best Lawyers in America* (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also serves as Executive Director of Stanford University Law School's Center for E-Commerce in Palo Alto.

Mr. Ballon received his B.A. *magna cum laude* from Tufts University, his J.D. *with honors* from George Washington University Law School and an LLM in international and comparative law from Georgetown University Law Center. He also holds the C.I.P.P./U.S. certification from the International Association of Privacy Professionals (IAPP).

In addition to *E-Commerce and Internet Law: Treatise with Forms 2d edition*, Mr. Ballon is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009), published by Thomson West (www.IanBallon.net).

He may be contacted at BALLON@GTLAW.COM and followed on Twitter and LinkedIn (@IanBallon).

Contributing authors: Parry Aftab, Ed Chansky, Francoise Gilbert, Tucker McCrady, Josh Raskin, Tom Smedinghoff and Emilio Varanini.

NEW AND IMPORTANT FEATURES FOR 2019

- > A comprehensive analysis of the **California Consumer Information Privacy Act**, **California's Internet of Things (IoT) security statute**, **Vermont's data broker registration law**, **Ohio's safe harbor** for companies with written information security programs, and other new state laws governing cybersecurity (chapter 27) and data privacy (chapter 26)
- > An exhaustive analysis of **FOSTA-SESTA** and what companies should do to maximize CDA protection in light of these new laws (chapter 37)
- > The **CLOUD Act** (chapter 50)
- > Understanding **the TCPA after ACA Int'l** and significant new cases & circuit splits (chapter 29)
- > Fully updated **50-state compendium** of security breach notification laws, with a **strategic approach** to handling notice to consumers and state agencies (chapter 27)
- > **Platform liability and statutory exemptions and immunities** (including a comparison of "but for" liability under the CDA and DMCA, and the latest law on secondary trademark and patent liability) (chapter 49)
- > Applying **the single publication rule** to websites, links and uses on social media (chapter 37)
- > The complex array of potential liability risks from, and remedies for, **screen scraping, database protection and use of AI to gather data and information online** (chapter 5)
- > State online dating and revenge porn laws (chapter 51)
- > **Circuit splits on Article III standing in cybersecurity litigation** (chapter 27)
- > Revisiting **sponsored link, SEO and SEM practices and liability** (chapter 9)
- > **Website and mobile accessibility** (chapter 48)
- > **The Music Modernization Act's Impact on copyright preemption and DMCA protection for pre-1972 musical works** (chapter 4)
- > **Compelling the disclosure of passwords and biometric information to unlock a mobile phone, tablet or storage device** (chapter 50)
- > Cutting through the jargon to make sense of **clickwrap, browsewrap, scrollwrap and sign-in wrap agreements (and what many courts and lawyers get wrong about online contract formation)** (chapter 21)
- > The latest case law, trends and strategy for **defending cybersecurity and data privacy class action suits** (chapters 25, 26, 27)
- > **Click fraud** (chapter 28)
- > Updated **Defend Trade Secrets Act** and **UTSA** case law (chapter 10)
- > **Drafting enforceable arbitration clauses and class action waivers** (with new sample provisions) (chapter 22)
- > **Applying the First Sale Doctrine to the sale of digital goods and information** (chapter 16)
- > **The GDPR, ePrivacy Directive and transferring data from the EU/EEA** (by Francoise Gilbert) (chapter 26)
- > **Patent law** (updated by Josh Raskin) (chapter 8)
- > **Music licensing** (updated by Tucker McCrady) (chapter 17)
- > **Mobile, Internet and Social Media contests & promotions** (updated by Ed Chansky) (chapter 28)
- > **Conducting a risk assessment and creating a Written Information Security Assessment Plan (WISP)** (by Thomas J. Smedinghoff) (chapter 27)

SAVE 20% NOW!!

To order call **1-888-728-7677**
or visit legalsolutions.thomsonreuters.com,
enter promo code **WPD20** at checkout

List Price: \$2,567.50
Discounted Price: \$2,054

LITIGATION RISKS AND COMPLIANCE OBLIGATIONS UNDER THE CALIFORNIA CONSUMER PRIVACY ACT

Excerpted from Chapter 26 (Data Privacy) of
E-Commerce and Internet Law: Legal Treatise with Forms 2d Edition
A 5-volume legal treatise by Ian C. Ballon (Thomson/West Publishing, www.IanBallon.net)

ARTIFICIAL INTELLIGENCE AND ROBOTICS
NATIONAL INSTITUTE
AMERICAN BAR ASSOCIATION
SANTA CLARA, CA
JANUARY 9-10, 2020

Ian C. Ballon
Greenberg Traurig, LLP

Silicon Valley: 1900 University Avenue, 5th Fl. East Palo Alto, CA 914303 Direct Dial: (650) 289-7881 Direct Fax: (650) 462-7881	Los Angeles: 1840 Century Park East, Ste. 1900 Los Angeles, CA 90067 Direct Dial: (310) 586-6575 Direct Fax: (310) 586-0575
---	--

Ballon@gtlaw.com
<www.ianballon.net>
LinkedIn, Twitter, Facebook: IanBallon



Ian C. Ballon

Shareholder

Internet, Intellectual Property & Technology Litigation

Admitted: California, District of Columbia and Maryland
Second, Third, Fourth, Fifth, Seventh, Ninth, Eleventh and Federal
Circuits

U.S. Supreme Court

JD, LLM, CIPP/US

Ballon@gtlaw.com

LinkedIn, Twitter, Facebook: IanBallon

Silicon Valley

1900 University Avenue

5th Floor

East Palo Alto, CA 94303

T 650.289.7881

F 650.462.7881

Los Angeles

1840 Century Park East

Los Angeles, CA 90067

T 310.586.6575

F 310.586.0575

Ian Ballon is a litigator who is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property & Technology Practice and represents Internet, technology, mobile and other companies in intellectual property and internet- and mobile-related litigation, including the defense of data privacy, security breach, and TCPA class action suits. He is also the author of the leading treatise on Internet law, *E-Commerce and Internet Law: Treatise with Forms 2d edition*, the 5-volume set published by West (www.IanBallon.net). In addition, he is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009). He also serves as Executive Director of Stanford University Law School's Center for the Digital Economy, which hosts the annual Best Practices Conference where lawyers, scholars and judges are regularly featured and interact. A list of recent cases may be found at <http://www.gtlaw.com/Ian-C-Ballon-experience>.

Mr. Ballon was named the Lawyer of the Year for Information Technology Law in the 2019, 2018, 2016 and 2013 editions of Best Lawyers in America. In both 2018 and 2019 he was recognized as one of the Top 1,000 trademark attorneys in the world for his litigation practice by *World Trademark Review*. In addition, in 2019 he was named one of the top 20 Cybersecurity lawyers in California and in 2018 one of the Top Cybersecurity/Artificial Intelligence lawyers in California by the *Los Angeles and San Francisco Daily Journal*. He received the "Trailblazer" Award, Intellectual Property, 2017 from *The National Law Journal* and he has been recognized as a "Groundbreaker" in *The Recorder's* 2017 Litigation Departments of the Year Awards. In addition, he was the 2010 recipient of the State Bar of California IP Section's Vanguard Award for significant contributions to the development of intellectual property law (<http://ipsection.calbar.ca.gov/IntellectualPropertyLaw/IPVanguardAwards.aspx>). He is listed in Legal 500 U.S., The Best Lawyers in America (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also has been recognized by *The Daily Journal* as one of the Top 75 IP litigators in California in every year that the list has been published, from 2009 through 2019, and has been listed as a Northern California Super Lawyer every year from 2004 through 2018 and as one of the Top 100 lawyers in California. Mr. Ballon also holds the CIPP/US certification from the International Association of Privacy Professionals (IAPP).

- (8) “Security purpose” means the purpose of preventing shoplifting, fraud, or any other misappropriation or theft of a thing of value, including tangible and intangible goods, services, and other purposes in furtherance of protecting the security or integrity of software, accounts, applications, online services, or any person.

Wash. Rev. Code Ann. § 19.001.003**Legislative findings**

- (1) The legislature finds that the practices covered by this chapter are matters vitally affecting the public interest for the purpose of applying the consumer protection act, chapter 19.86 RCW. A violation of this chapter is not reasonable in relation to the development and preservation of business and is an unfair or deceptive act in trade or commerce and an unfair method of competition for the purpose of applying the consumer protection act, chapter 19.86 RCW.
- (2) This chapter may be enforced solely by the attorney general under the consumer protection act, chapter 19.86 RCW.

Wash. Rev. Code Ann. § 19.001.004**Application of chapter**

- (1) Nothing in this act applies in any manner to a financial institution or an affiliate of a financial institution that is subject to Title V of the federal Gramm-Leach-Bliley act of 1999 and the rules promulgated thereunder.
- (2) Nothing in this act applies to activities subject to Title V of the federal health insurance privacy and portability act of 1996 and the rules promulgated thereunder.
- (3) Nothing in this act expands or limits the authority of a law enforcement officer acting within the scope of his or her authority including, but not limited to, the authority of a state law enforcement officer in executing lawful searches and seizures.

26.13A California Consumer Privacy Act (CCPA)¹**In General**

[Section 26.13A]

¹This section was co-authored with Greenberg Traurig attorney

The California Consumer Privacy Act (CCPA)² was hastily enacted in 2018 to avoid a more inflexible ballot initiative that would have been next to impossible to amend.³ The CCPA was influenced by the GDPR,⁴ which took effect in the European Union and European Economic Area in May 2018, as well as prior California data privacy and consumer laws. The statute was amended in September 2018 and it is expected that it will be amended again at least one more time before it takes effect on January 1, 2020.⁵ It is also possible that the enactment of CCPA could prompt Congress to adopt a federal consumer privacy law to preempt state laws so that there is a uniform national standard, as has occurred in the past with other laws such as the CAN-SPAM Act,⁶ which was enacted after California enacted a very strict email marketing law. Absent federal preemption, other states may enact similar regulatory schemes—potentially with variations that could make it more complex for companies to comply. A copy of the CCPA as amended in September 2018 is reprinted at the end of this chapter at Appendix 8.

Rebekah Guyon.

²Cal. Civ. Code §§ 1798.100 to 1798.196.

³Real estate millionaire Alastair Mactaggart had spent \$2 million to obtain enough signatures for a ballot initiative that would have created a comprehensive consumer privacy law, enforced through litigation. Because laws enacted through ballot initiatives in California require a supermajority to amend—and therefore are effectively almost impossible to revise—legislative and business leaders worked together to enact a somewhat better version of the law by the deadline set by Mactaggart—5 P.M. on June 28, 2018—which was the last date by which the initiative could be withdrawn from the 2018 California ballot. *See, e.g.,* Nicholas Confessore, *The Unlikely Activist Who Took On Silicon Valley—and Won*, N.Y. Times, Aug. 14, 2018. Mactaggart had an incentive to cut a deal because advertising for ballot initiatives is very costly and, even when enacted, many initiatives are subject to legal challenge. The rush to cut a deal with the millionaire backer of the consumer privacy initiative, however, resulted in a statute that was more than 10,000 words long, complex, and contained numerous errors and ambiguities. *See, e.g.,* Eric Goldman, *A First (But Very Incomplete) Crack at Inventorying the California Consumer Privacy Act's Problems*, Technology & Marketing Law Blog, July 24, 2018, available at <https://blog.ericgoldman.org/archives/2018/07/a-first-but-very-incomplete-crack-at-inventorying-the-california-consumer-privacy-acts-problems.htm>.

⁴*See supra* § 26.04.

⁵*See* Cal. Civ. Code § 1798.198(a) (setting the operative date of the statute as January 1, 2020, subject to the withdrawal of a ballot initiative that in fact was withdrawn).

⁶15 U.S.C.A. §§ 7701 to 7713; *see infra* § 29.03.

The CCPA imposes certain statutory obligations, which will be supplemented by regulations that the California Attorney General will issue in 2019. Subject to enumerated exclusions discussed later in this section (including businesses subject to federal financial services and health care privacy regulations), it broadly addresses the use of personal information about California residents—not merely consumers.⁷ Rather than regulating the use, collection and dissemination of information obtained *by companies from consumers*, as past consumer laws did, the CCPA focuses on information *about* state residents, and therefore regulates privacy more broadly than—and addresses perceived loopholes that existed in—prior consumer privacy laws. The statute requires not simply that businesses amend their privacy policies to account for the law, but that specific notices be placed on a business’s website, written contracts be entered into with service providers, and ultimately that internal practices and procedures be adjusted to ensure compliance with the statute, for those businesses that are subject to it.

The CCPA is intended to impose compliance obligations on larger business entities and those involved in selling customer information. It applies to a business “that collects⁸ consumers’ personal information, or on the behalf of which such information is collected and that alone, or jointly with others, determines the purposes and means of the processing of consumers’ personal information, that does business in the State of California.”⁹ A business is subject to the CCPA only if it:

- (1) has “annual gross revenues in excess of twenty-five million dollars”
- (2) buys, receives for commercial purposes, or sells the personal information of 50,000 or more consumers, households, or devices or

⁷Cal. Civ. Code § 1798.140(g) (“‘Consumer’ means a natural person who is a California resident, as defined in Section 17014 of Title 18 of the California Code of Regulations, as that section read on September 1, 2017, however identified, including by any unique identifier.”).

⁸*Collects, collected, or collection* means “buying, renting, gathering, obtaining, receiving, or accessing any personal information pertaining to a consumer by any means. This includes receiving information from the consumer, either actively or passively, or by observing the consumer’s behavior.” Cal. Civ. Code § 1798.140(e).

⁹Cal. Civ. Code § 1798.140(c)(1).

- (3) “[d]erives 50 percent or more of its annual revenues from selling consumers’ personal information.”¹⁰

The collection or sale of personal information that takes place “wholly outside of California” is not subject to the CCPA.¹¹

By contract, businesses subject to the CCPA must impose use and deletion obligations with respect to personal information on *service providers*. A service provider is “a sole proprietorship, partnership, limited liability company, corporation, association, or other legal entity that is organized or operated for the profit or financial benefit of its shareholders or other owners, that processes information on behalf of a business and to which the business discloses a consumer’s personal information for a business purpose”¹²

¹⁰Cal. Civ. Code § 1798.140(c) (defining a *business*). The law also applies to an entity “that controls or is controlled by a business . . . and that shares common branding with the business.” *Id.* § 1798.140(c)(2). *Control* or *controlled* means “ownership of, or the power to vote, more than 50 percent of the outstanding shares of any class of voting security of a business; control in any manner over the election of a majority of the directors, or of individuals exercising similar functions; or the power to exercise a controlling influence over the management of a company. *Id.* *Common branding* means “a shared name, servicemark, or trademark.” *Id.*

¹¹Cal. Civ. Code § 1798.145(a)(6).

¹²*Business purpose* means “the use of personal information for the business’s or a service provider’s operational purposes, or other notified purposes, provided that the use of personal information shall be reasonably necessary and proportionate to achieve the operational purpose for which the personal information was collected or processed or for another operational purpose that is compatible with the context in which the personal information was collected.” Cal. Civ. Code § 1798.140(d). The statute provides seven examples of *business purposes*, which presumably is a non-exclusive list of examples. Those examples are:

- (1) Auditing related to a current interaction with the consumer and concurrent transactions, including, but not limited to, counting ad impressions to unique visitors, verifying positioning and quality of ad impressions, and auditing compliance with this specification and other standards.
- (2) Detecting security incidents, protecting against malicious, deceptive, fraudulent, or illegal activity, and prosecuting those responsible for that activity.
- (3) Debugging to identify and repair errors that impair existing intended functionality.
- (4) Short-term, transient use, provided the personal information that [sic] is not disclosed to another third party and is not used to build a profile about a consumer or otherwise alter an individ-

pursuant to a written contract, provided that the contract prohibits the entity receiving the information from retaining, using, or disclosing the personal information for any purpose other than for the specific purpose of performing the services specified in the contract for the business, or as otherwise permitted by [the CCPA], including retaining, using, or disclosing the personal information for a commercial purpose other than providing the services specified in the contract with the business.”¹³ Thus, a *service provider* under the CCPA is broadly defined as an entity or person that processes information for a business, but only includes persons or entities operating for profit (or financial benefit), and requires that a written contract be in place restricting the service provider’s ability to retain, use or disclose personal information except as permitted by the contract or the CCPA. A service provider also must certify in its written contract with a business its compliance with the CCPA.¹⁴ A business that discloses personal information to a service provider will not be liable under the CCPA if the service provider uses the personal information in violation of the

ual consumer’s experience outside the current interaction, including, but not limited to, the contextual customization of ads shown as part of the same interaction.

- (5) Performing services on behalf of the business or service provider, including maintaining or servicing accounts, providing customer service, processing or fulfilling orders and transactions, verifying customer information, processing payments, providing financing, providing advertising or marketing services, providing analytic services, or providing similar services on behalf of the business or service provider.
- (6) Undertaking internal research for technological development and demonstration.
- (7) Undertaking activities to verify or maintain the quality or safety of a service or device that is owned, manufactured, manufactured for, or controlled by the business, and to improve, upgrade, or enhance the service or device that is owned, manufactured, manufactured for, or controlled by the business.

Id. Subsection 4 contains obvious typographical errors and likely was intended to refer to short-term, transient use, provided that personal information is not disclosed to a third party.

¹³Cal. Civ. Code § 1798.140(v).

¹⁴Cal. Civ. Code § 1798.140(w)(2)(a)(ii). The requirement that a service provider certify its compliance with the CCPA is not included in the statute’s definition for *service provider*, but is separately set forth as a requirement to avoid being classified as a “third party,” which would subject the business to potential liability under the CCPA. *Compare* Cal. Civ. Code § 1798.140(v) *with* Cal. Civ. Code § 1798.140(w).

CCPA, “provided that, at the time of disclosing the personal information, the business does not have actual knowledge, or reason to believe, that the service provider intends to commit such a violation.”¹⁵ A service provider will likewise not be liable under the CCPA for the obligations of a business for which it provides services.¹⁶ Service providers are subject to enforcement actions brought by the California Attorney General¹⁷ and presumably breach of contract actions brought by a contracting business.

Unlike a *third party*,¹⁸ a business is not required to dis-

¹⁵Cal. Civ. Code § 1798.145(h).

¹⁶Cal. Civ. Code § 1798.145(h).

¹⁷Cal. Civ. Code § 1798.155(b).

¹⁸A *third party* means a person who is not any of the following:

- (1) The business that collects personal information from consumers under this title.
- (2)
 - (A) A person to whom the business discloses a consumer’s personal information for a business purpose pursuant to a written contract, provided that the contract:
 - (i) Prohibits the person receiving the personal information from:
 - (I) Selling the personal information.
 - (II) Retaining, using, or disclosing the personal information for any purpose other than for the specific purpose of performing the services specified in the contract, including retaining, using, or disclosing the personal information for a commercial purpose other than providing the services specified in the contract.
 - (III) Retaining, using, or disclosing the information outside of the direct business relationship between the person and the business.
 - (ii) Includes a certification made by the person receiving the personal information that the person understands the restrictions in subparagraph (A) and will comply with them.
 - (B) A person covered by this paragraph that violates any of the restrictions set forth in this title shall be liable for the violations. A business that discloses personal information to a person covered by this paragraph in compliance with this paragraph shall not be liable under this title if the person receiving the personal information uses it in violation of the restrictions set forth in this title, provided that, at the time of disclosing the personal information, the business does not have actual knowledge, or reason to believe, that the person intends to commit such a violation.

Cal. Civ. Code § 1798.140(w).

close to consumers the categories of service providers to which it provides access to personal information.¹⁹ A third party is restricted from selling personal information about a consumer sold to it by a business “unless the consumer has received explicit notice and is provided an opportunity to exercise the right to opt-out pursuant to Section 1798.120.”²⁰

As amended in September 2018, the Act affords California residents the rights to:

- Notice of the personal information collected and the purpose for collecting each category of information, at or before the point at which the information is collected;
- Request that a business that collects a consumer’s personal information disclose the categories of personal information collected about a consumer and provide copies of the specific personal information collected;
- Request that a business that sells or discloses a consumer’s personal information disclose the categories of personal information sold or disclosed about a consumer;
- Opt-out of the collection of personal information (and, for minors not otherwise subject to the Child Online Privacy Protection Act (COPPA),²¹ affirmatively requires opt-in consent²²);
- Request that a business that collects a consumer’s personal information delete any personal information about the consumer that the business has collected.

The CCPA also prohibits a business from selling personal information purchased from another business without explicitly notifying the consumers whose information would be sold and providing an opportunity to opt out.²³

Personal information includes, but is not limited to, a non-exclusive list of specific data elements,²⁴ “if it identifies, relates to, describes, is capable of being associated with, or could be reasonably linked, directly or indirectly, with a par-

¹⁹Compare Cal. Civ. Code § 1798.115(a)(2) with Cal. Civ. Code § 1798.140(t).

²⁰Cal. Civ. Code § 1798.115(d).

²¹15 U.S.C.A. §§ 6501 to 6506; 16 C.F.R. §§ 312.1 to 312.13; *supra* § 26.13[2].

²²Cal. Civ. Code § 1798.120(c).

²³Cal. Civ. Code § 1798.115(d).

²⁴Cal. Civ. Code § 1798.140(o)(1).

ticular consumer or household. . . .” The data elements identified in the statute, which may be supplemented by regulation,²⁵ are:

- (A) Identifiers such as a real name, alias, postal address, unique personal identifier, online identifier, Internet Protocol address, email address, account name, social security number, driver’s license number, passport number, or other similar identifiers.
- (B) Any categories of personal information described in subdivision (e) of Section 1798.80.²⁶
- (C) Characteristics of protected classifications under California or federal law.
- (D) Commercial information, including records of personal property, products or services purchased, obtained, or considered, or other purchasing or consuming histories or tendencies.
- (E) Biometric information.
- (F) Internet or other electronic network activity information, including, but not limited to, browsing history, search history, and information regarding a consumer’s interaction with an Internet website, application, or advertisement.
- (G) Geolocation data.
- (H) Audio, electronic, visual, thermal, olfactory, or similar information.
- (I) Professional or employment-related information.
- (J) Education information, defined as information that is not publicly available personally identifiable information as defined in the Family Educational Rights and Privacy Act (20 U.S.C.A. § 1232g, 34 C.F.R. Part 99).
- (K) Inferences drawn from any of the information identi-

²⁵Cal. Civ. Code § 1798.185(a)(1).

²⁶Cal. Civ. Code § 1798.80 defines *personal information* as any information that identifies, relates to, describes, or is capable of being associated with, a particular individual, including, but not limited to, his or her name, signature, social security number, physical characteristics or description, address, telephone number, passport number, driver’s license or state identification card number, insurance policy number, education, employment, employment history, bank account number, credit card number, debit card number, or any other financial information, medical information, or health insurance information. “Personal information” does not include publicly available information that is lawfully made available to the general public from federal, state, or local government records.

Id.

fied in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes.²⁷

Personal information, however, excludes *publicly available information*.²⁸ But this exclusion is less than what meets the

²⁷Cal. Civ. Code § 1798.140(o)(1).

²⁸*Publicly available* means “information that is lawfully made available from federal, state, or local government records, if any conditions associated with such information.” Cal. Civ. Code § 1798.140(o)(2). Words such as “are complied with” appear to have been omitted from the end of this sentence, which plainly appears to be a clause limiting the scope of what may constitute publicly available information. This is made clear two sentences later in the same definitional section, which provides that “information is not ‘publicly available’ if that data is used for a purpose that is not compatible with the purpose for which the data is maintained and made available in the government records or for which it is publicly maintained.” *Id.*

The definition of what constitutes material *publicly available* also excludes “consumer information that is deidentified or aggregate consumer information.” *Id.* This may seem at face value as a perplexing exclusion because government data frequently includes deidentified or aggregate consumer information. One might assume that this exclusion was not intended to apply to data released by a government agency that has been deidentified, as opposed to information deidentified by a business. However, given the language of the statute and the obligations imposed on a business that use deidentified or aggregate consumer information, it appears that the definition broadly encompasses even deidentified or aggregate consumer data provided by a government agency.

The CCPA treats consumer information that has been deidentified or presented in aggregate form as a separate category of information and imposes special obligations on businesses that use it. A business that uses deidentified or aggregate consumer information must have:

- (1) implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain.
- (2) implemented business processes that specifically prohibit reidentification of the information.
- (3) implemented business processes to prevent inadvertent release of deidentified information.
- (4) made no attempt to reidentify the information.

Cal. Civ. Code § 1798.140(h). Otherwise, the information will be treated as *personal information* (because it will not qualify as *deidentified* or *aggregate consumer data* under the statute, and therefore will not be excluded from the definition of *personal information* as information that is *publicly available*). Hence, it appears that the legislature intended to exclude *deidentified* and *aggregate consumer data* from the definition of material that is *publicly available* to ensure that it was handled with the same

eye. Information is “not ‘publicly available’ if that data is used for a purpose that is not compatible with the purpose for which the data is maintained and made available in the government records or for which it is publicly maintained.”²⁹ Accordingly, unless a business intends to use public information for the same purpose as the government entity that maintains it, public information collected, sold, or disclosed may be subject to the CCPA’s disclosure and deletion requirements.

Publicly available also does not mean biometric information collected by a business about a consumer without the consumer’s knowledge³⁰ (and thus constitutes *personal information*).

The CCPA does not restrict a business’s collection, use, retention, sale, or disclosure of “deidentified” or “aggregate consumer information.”³¹ However, as noted earlier in connection with the definition of what constitutes information made *publicly available*, deidentified and aggregate consumer information could become *personal information* if a business fails to undertake the four protective measures included in section 1798.140(h).

Conversely, the CCPA generally does not require re-identification or de-anonymization of deidentified or aggregate consumer data so that the information would be subject to the requirements imposed on *personal information* under the law. The CCPA may not be construed to require “a business to reidentify or otherwise link information that is not maintained in a manner that would be considered personal

level of care by a business as information that has been de-anonymized by a private business.

²⁹Cal. Civ. Code § 1798.140(o)(2).

³⁰Cal. Civ. Code § 1798.140(o)(2).

³¹Cal. Civ. Code § 1798.145(a)(5). *Deidentified* is defined as “information that cannot reasonably identify, relate to, describe, or be capable of being associated with, or be linked, directly or indirectly, to a particular consumer,” provided that a business has implemented the four technical safeguards and business processes discussed earlier, to prevent reidentification of the information. *Id.* § 1798.140(h). *Aggregate consumer information* is defined as information that “relates to a group or category of consumers, from which individual consumer identities have been removed” and which is “not linked or reasonably linkable to any consumer or household, including via a device.” *Id.* § 1798.140(a). A collection of individual consumer records that have been deidentified, however, is not “[a]ggregate consumer information” under the CCPA. *Id.* § 1798.140(a).

information.”³²

Overall, the definition of *personal information* is quite broad. For example, the inclusion of “[i]nferences drawn from the information identified in this subdivision to create a profile about a consumer” means any time a company draws an inference about a user, the inferences themselves become personal information, subject to the statute. Likewise, the fact that public information can become *personal information* if used for a different purpose or if a business fails to treat deidentified or aggregate consumer data from a government agency as required by section 1798.140(h), reflects an expansive notion of what constitutes PII compared to other U.S. state and federal laws.

The CCPA directs the California Attorney General to issue regulations to provide greater clarity on a number of aspects of the law on or before January 1, 2020, and empowers the AG to enforce it six months after the publication of final regulations or by July 1, 2020, whichever is sooner.³³

The statute also creates a private right of action and provides for statutory damages for a security breach involving personal information that results from a business’s failure to implement and maintain reasonable security procedures, subject to a 30 day right to cure.³⁴

Existing California privacy laws in effect prior to the time the CCPA takes effect are analyzed in section 26.13[6].

Notice to consumers of the personal information collected and the purpose for its collection, at or before the point at which the information is collected

The CCPA requires that a business that collects personal information from consumers notify consumers, at or before the point at which information will be collected, what categories of personal information will be collected and the purposes for which each category of personal information will be used.³⁵ As a corollary to this rule, the CCPA provides that a business may not “collect additional categories of personal information or use personal information collected for additional purposes” without providing this notice to a

³²Cal. Civ. Code § 1798.145(i).

³³See Cal. Civ. Code § 1798.185.

³⁴Cal. Civ. Code § 1798.150(a).

³⁵Cal. Civ. Code § 1798.100(b).

consumer.³⁶

Disclosure requirements pursuant to consumer requests

The CCPA provides California residents with a right to request disclosures of the “categories” of their personal information that a business has collected, sold, and used.³⁷ The “categories” referred to in these disclosure requirements “follow the definition of personal information” in the statute, which are the same categories (A) through (K) noted earlier, and may be supplemented by the California Attorney General.³⁸

A business may only provide the required disclosures “upon receipt of a verifiable consumer request.”³⁹ It likewise is not required to provide personal information requested to a consumer more than twice in a 12-month period.⁴⁰

A business is required to disclose the information requested within “45 days of receiving a verifiable consumer request from the consumer.”⁴¹ The 45 day time period may be extended once by an additional 45 days.⁴² Additionally, a business may take up to “90 additional days where necessary, taking into account the complexity and number of the requests” to respond.⁴³ A business is required to notify the consumer of the extension within 45 days of receiving the request and, for extensions beyond the additional 45 days,

³⁶Cal. Civ. Code § 1798.100(b).

³⁷Cal. Civ. Code §§ 1798.100, 1798.110, 1798.115.

³⁸Cal. Civ. Code §§ 1798.130(c), 1798.140(o), 1798.185(a)(2).

³⁹Cal. Civ. Code § 1798.100(c), Cal. Civ. Code § 1798.130(a)(2). A *verifiable consumer request* is a request “by a consumer,” on his or her own behalf or on behalf of a minor child or other person authorized to act on the consumer’s behalf, “that the business can reasonably verify” pursuant to regulations that the Attorney General is required to implement no later than July 1, 2020. Cal. Civ. Code § 1798.140(y). A business is not required to produce personal information if it cannot verify the identity of the requesting party. *Id.* § 1798.140(y). However, it is unclear what steps a business will be required to take to verify a consumer request after the CCPA’s effective date if the Attorney General has not implemented the regulations provided in this section by then.

⁴⁰Cal. Civ. Code § 1798.100(d).

⁴¹Cal. Civ. Code § 1798.130(a)(2).

⁴²Cal. Civ. Code § 1798.130(a)(2).

⁴³Cal. Civ. Code § 1798.145(g)(1).

the business must provide the reason for the delay.⁴⁴

A business must deliver the information “free of charge to the consumer,” unless the requests are “manifestly unfounded or excessive . . . because of their repetitive character,” in which case a business may charge a “reasonable fee” for the disclosure.⁴⁵ The disclosure “shall cover the 12-month period preceding the business’s receipt of the verifiable consumer request”.⁴⁶ The information should be sent via a consumer’s “account with the business,” if one exists, and if not, it may be delivered by mail or electronically, at the consumer’s option.⁴⁷ The information must be “in a portable and, to the extent technically feasible, in a readily useable format that allows the consumer to transmit this information to another entity without hindrance.”⁴⁸

If a business does not “take action on the request of the consumer, the business shall inform the consumer, without delay and at the latest within the time period permitted” for its response “of the reasons for not taking action and any rights the consumer may have to appeal the decision to the business.”⁴⁹

A business must provide consumers “two or more designated methods for submitting” disclosure requests, which must include, “at a minimum, a toll-free telephone number, and if the business maintains an Internet Web site, a Web site address.”⁵⁰ A business must also ensure that its customer service representatives are “informed” of the CCPA’s requirements regarding disclosure of personal information collected, sold, and disclosed, and financial incentives offered for personal information, and how to “direct consumers to exercise” their disclosure rights under the CCPA.⁵¹

Right to the disclosure of the categories and specific pieces of personal information collected

The CCPA provides that a “consumer shall have the right

⁴⁴Cal. Civ. Code §§ 1798.130(a)(2), 1798.145(g)(1).

⁴⁵Cal. Civ. Code § 1798.100(d); Cal. Civ. Code § 1798.145(g)(3).

⁴⁶Cal. Civ. Code § 1798.130(a)(2).

⁴⁷Cal. Civ. Code §§ 1798.100(d), 1798.130(a)(2).

⁴⁸Cal. Civ. Code §§ 1798.100(d), 1798.130(a)(2).

⁴⁹Cal. Civ. Code § 1798.145(g)(2).

⁵⁰Cal. Civ. Code § 1798.130(a)(1).

⁵¹Cal. Civ. Code § 1798.130(a)(6).

to request that a business that collects a consumer's personal information disclose to that consumer the categories and specific pieces of personal information the business has collected."⁵² Pursuant to section 1798.110, a consumer has the right to request that the business disclose:

- (1) "the categories of personal information it has collected about that consumer"
- (2) "the categories of sources from which the personal information is collected"
- (3) "the business or commercial purpose for collecting or selling personal information"
- (4) "the categories of third parties with whom the business shares personal information"; and
- (5) "the specific pieces of personal information it has collected about that consumer."⁵³

As a limiting factor, however, a business is not required to "[r]eidentify or otherwise link any data that, in the ordinary course of business, is not maintained in a manner that would be considered personal information" to comply with these disclosure requirements.⁵⁴ Thus, the fact that information may be de-anonymized or re-personalized does not mean that it is in fact subject to the statute's disclosure requirements.

Likewise, section 1798.100 does not require a business "to retain any personal information collected for a single, one-time transaction, . . ." if the information "is not sold or retained by the business or [used] to reidentify or otherwise link information that is not maintained in a manner that would be considered personal information."⁵⁵ Although drafted inartfully, this section appears intended to obviate the need for a business to retain (and hence potentially produce) personal information collected for a single, one-time transaction, provided the information is not (1) sold to third parties, (2) retained by the business, or (3) used to reidentify (or repersonalize) aggregate data or otherwise link information that would not be considered *personal information*.⁵⁶

⁵²Cal. Civ. Code § 1798.100(a).

⁵³Cal. Civ. Code §§ 1798.110(a)(1)—(5).

⁵⁴Cal. Civ. Code § 1798.110(d)(1).

⁵⁵Cal. Civ. Code § 1798.100(e).

⁵⁶Similarly, section 1798.110, which further specifies a business's

Right to the disclosure of the categories of personal information sold or disclosed

The CCPA provides that a consumer “shall have the right to request that a business that sells the consumer’s personal information, or that discloses it for a business purpose” make certain disclosures to the consumer.⁵⁷ *Sell* is not limited in the statute to the exchange of personal information for money, but covers any transfer “by the business to another business or a third party for monetary or other valuable consideration.”⁵⁸ The CCPA further provides that courts

duty to disclose personal information collected, more broadly states that a business is not required to retain personal information “collected for a single one-time transaction if, in the ordinary course of business, that information about the consumer is not retained.” Cal. Civ. Code § 1798.110(d)(1).

⁵⁷Cal. Civ. Code § 1798.115(a). What constitutes a *business purpose* is discussed earlier in this section and defined in Cal. Civ. Code § 1798.140(d).

⁵⁸Cal. Civ. Code § 1798.140(t). The statute provides that a business does *not* sell personal information when:

- (A) A consumer uses or directs the business to intentionally disclose personal information or uses the business to intentionally interact with a third party, provided the third party does not also sell the personal information, unless that disclosure would be consistent with the provisions of this title. An intentional interaction occurs when the consumer intends to interact with the third party, via one or more deliberate interactions. Hovering over, muting, pausing, or closing a given piece of content does not constitute a consumer’s intent to interact with a third party.
- (B) The business uses or shares an identifier for a consumer who has opted out of the sale of the consumer’s personal information for the purposes of alerting third parties that the consumer has opted out of the sale of the consumer’s personal information.
- (C) The business uses or shares with a service provider personal information of a consumer that is necessary to perform a business purpose if both of the following conditions are met:
 - (i) The business has provided notice that information being used or shared in its terms and conditions consistent with Section 1798.135.
 - (ii) The service provider does not further collect, sell, or use the personal information of the consumer except as necessary to perform the business purpose.
- (D) The business transfers to a third party the personal information of a consumer as an asset that is part of a merger, acquisition, bankruptcy, or other transaction in which the third party assumes control of all or part of the business, provided that information is used or shared consistently with Sections 1798.110 and 1798.115. If a third party materially alters how it uses or shares the personal information of a consumer in a manner that

examining compliance with its provisions should take a liberal approach to determining whether a transaction is a sale subject to its regulation. The CCPA mandates that, where a series of “steps or transactions” are taken “with the intention of avoiding the reach of this title, including the disclosure of information by a business to a third party in order to avoid the definition of sell, a court shall disregard the intermediate steps or transactions for purposes of effectuating the purposes of this title.”⁵⁹

A consumer has the right to request disclosure of:

- (1) the “categories of personal information that the business collected about the consumer”;
- (2) the “categories of personal information that the business sold about the consumer and the categories of third parties to whom the personal information was sold,” broken down by “category or categories of personal information for each third party to whom the personal information was sold”; and
- (3) the “categories of personal information that the business disclosed about the consumer for a business purpose.”⁶⁰

A business that both sells and discloses personal information is required to separately list the categories of personal information sold and disclosed in response to a consumer request.⁶¹

Right to the deletion of personal information

The CCPA provides that a “consumer shall have the right to request that a business delete any personal information

is materially inconsistent with the promises made at the time of collection, it shall provide prior notice of the new or changed practice to the consumer. The notice shall be sufficiently prominent and robust to ensure that existing consumers can easily exercise their choices consistently with Section 1798.120. This subparagraph does not authorize a business to make material, retroactive privacy policy changes or make other changes in their privacy policy in a manner that would violate the Unfair and Deceptive Practices Act (Chapter 5 (commencing with Section 17200) of Part 2 of Division 7 of the Business and Professions Code).

Id. § 1798.140(t)(2).

⁵⁹Cal. Civ. Code § 1798.190.

⁶⁰Cal. Civ. Code §§ 1798.115(a)(1)—(3).

⁶¹Cal. Civ. Code § 1798.130(a)(4).

about the consumer which the business has collected from the consumer.”⁶² When a business receives a “verifiable consumer request from a consumer to delete the consumer’s personal information” the business must “delete the consumer’s personal information” not only from its own records, but the business must also direct any “service providers to delete the consumer’s personal information from their records” as well.⁶³

The CCPA carves out specific exceptions to the deletion requirement. Although not expansive, as written the exceptions allow a business to retain personal information when it is necessary for an ongoing business relationship with the consumer because the information is necessary to complete a transaction or provide a good or service that the consumer requested.⁶⁴ Additionally, a business may retain the information for internal use, as long as the use is “reasonably aligned with the expectations of the consumer based on the consumer’s relationship with the business” or “compatible with the context in which the consumer provided the information.”⁶⁵ A business may also retain consumer information for the purpose of detecting “security incidents,” protecting against or prosecuting malicious and fraudulent activity,⁶⁶ debugging,⁶⁷ and to comply with the California Electronic Communications Privacy Act, Cal. Penal Code § 1546 or another “legal obligation.”⁶⁸ Other statutory exclusions are less clear; a business may retain and use consumers’ personal information after a deletion request to “[e]xercise free speech,” or

⁶²Cal. Civ. Code § 1798.105(a).

⁶³Cal. Civ. Code § 1798.105(c). A *service provider* is a for-profit entity that “process information on behalf of a business and to which the business discloses a consumer’s personal information for a business purpose pursuant to a written contract.” *Id.* § 1798.140(v). The definition of “service provider” additionally requires a business subject to CCPA to specify in a written contract that the provider is prohibited from using the personal information for any purpose other than that outlined in the contract. *Id.* § 1798.140(v). Businesses thus must put in place written contracts with service providers. A service provider is also required to certify to its compliance with the CCPA in its written contract with a business. *Id.* § 1798.140(w)(2)(A)(ii).

⁶⁴Cal. Civ. Code § 1798.105(d)(1).

⁶⁵Cal. Civ. Code §§ 1798.105(d)(7), (9).

⁶⁶Cal. Civ. Code § 1798.105(d)(2).

⁶⁷Cal. Civ. Code § 1798.105(d)(3).

⁶⁸Cal. Civ. Code §§ 1798.105(d)(5), (8).

ensure another's right to exercise his or her free speech, or for the purpose of engaging in "public or peer-reviewed scientific, historical, or statistical research in the public interest."⁶⁹ Presumably, this is intended to allow an interactive computer service provider discretion to decline takedown requests directed at consumer review sites or other online discussion fora, and to protect free speech and the integrity of academic research. The exact contours of this exception, including the undefined term "public interest," have yet to be fleshed out.

**Right to opt-out of the sale of personal information/
minors' right to opt-in**

The CCPA gives California residents a right to opt-out of having their information sold, and requires affirmative opt-in consent from minors.

The statute provides that a "consumer shall have the right, at any time, to direct a business that sells personal information about the consumer to third parties not to sell the consumer's personal information," referred to as the "right to opt-out."⁷⁰

A business that sells consumers' personal information is required to notify California residents of their right to opt out. This notification must be provided through "a clear and conspicuous link on the business's Internet home page, titled 'Do Not Sell My Personal Information,' " which must link to an "Internet Web page that enables a consumer" "to opt out of the sale of the consumer's personal information."⁷¹ A business can maintain a separate homepage for California consumers with the required link if the business "takes reasonable steps to ensure that California consumers are directed" to that homepage and "not the homepage made available to the public generally."⁷² A business cannot require a consumer to create an account in order to opt-out.⁷³ A business that sells consumers' personal information must ensure

⁶⁹Cal. Civ. Code §§ 1798.105(d)(4), (6). *Research* is narrowly limited to studies "[c]ompatible with the business purpose for which the personal information was collected," and that are "[n]ot for any commercial purpose," among other limitations. Cal. Civ. Code § 1798.140(s).

⁷⁰Cal. Civ. Code § 1798.120(a).

⁷¹Cal. Civ. Code § 1798.135(a)(1).

⁷²Cal. Civ. Code § 1798.135(b).

⁷³Cal. Civ. Code § 1798.135(a)(1).

that its customer service representatives are aware of consumers' right to opt-out and how to exercise that right.⁷⁴ After a consumer has opted out, a business is prohibited from requesting that the consumer reauthorize the sale of his or her data for "at least 12 months."⁷⁵

With respect to minors, the CCPA prohibits businesses from selling personal information from consumers "if the business has actual knowledge that the consumer is less than 16 years of age," unless, for "consumers between 13 and 16 years of age" the consumer affirmatively authorizes the sale, or the parent or guardian of a consumer under 13 years of age affirmatively authorizes the sale.⁷⁶ The CCPA provides that a "business that willfully disregards the consumer's age shall be deemed to have actual knowledge of the consumer's age."⁷⁷ The CCPA refers to the prohibition on the sale of minors' personal information without consent as the "right to opt-in."⁷⁸

The requirement for parental consent for children under age 13 is consistent with the federal Child Online Privacy Protection Act (COPPA).⁷⁹ Federal law does not generally regulate child privacy for those aged 13 and older, although the FTC has identified minors in this age group as deserving of closer attention.⁸⁰ The CCPA provides special protection for this class of people, although inartful draftsmanship makes it unclear whether the law covers those who are 16 and on its face appears to exclude minors who are age 13, which presumably was not the drafters' intent. Presumably, the CCPA's opt-in right should apply to teenagers aged 13, 14 and 15, but the language of the statute is not entirely clear.⁸¹

⁷⁴Cal. Civ. Code § 1798.135(a)(3).

⁷⁵Cal. Civ. Code § 1798.135(a)(5).

⁷⁶Cal. Civ. Code § 1798.120(c).

⁷⁷Cal. Civ. Code § 1798.120(c).

⁷⁸Cal. Civ. Code § 1798.120(c).

⁷⁹15 U.S.C.A. §§ 6501 to 6506; 16 C.F.R. §§ 312.1 to 312.13; *see generally supra* § 26.13[2].

⁸⁰*See supra* § 26.13[2][H].

⁸¹*See* Cal. Civ. Code § 1798.120(c); Eric Goldman, *California Amends the Consumer Privacy Act (CCPA); Fixes About 0.01% of its Problems*, Technology & Marketing Law Blog (Oct. 4, 2018), available at <https://blog.ericgoldman.org/archives/2018/10/california-amends-the-consumer->

Nondiscrimination and Financial incentives

The CCPA generally prohibits businesses from discriminating against consumers based on their exercise of any rights provided in the statute.⁸² Discrimination includes denying a consumer goods or services, charging different prices or rates, providing a different level or quality of goods or services, and/or suggesting that a consumer will receive a different price, rate, or level or quality of goods or services.⁸³ However, the CCPA also provides that businesses are not prohibited from “charging a consumer a different price or rate, or from providing a different level or quality of goods or services to the consumer, if that difference is reasonably related to the value provided to the consumer by the consumer’s data.”⁸⁴

The CCPA also allows a business to offer “financial incentives” for the collection, sale, or deletion of personal information. These incentives may only be provided on an opt-in basis, and include “payments to consumers as compensation,” or a “different price, rate, level or quality of goods or services to the consumer if that price is directly related to the value provided to the consumer by the consumer’s data.”⁸⁵

In other words, a business may not discriminate against a consumer who declines to provide consent or requests deletion of personal information, but it may provide financial incentives for a consumer not to do so. Financial incentives must be correlated to the value of a consumer’s information.

privacy-act-ccpa-fixes-about-0-01-of-its-problems.htm (“The language is inconsistent about 16 year olds (or, if you read the restriction as applying only to 14 and 15 year olds, then it’s inconsistent about 13 year olds): ‘a business shall not sell the personal information of consumers if the business has actual knowledge that the consumer is **less than 16 years of age**, unless the consumer, in the case of consumers **between 13 and 16 years of age**, or the consumer’s parent or guardian, in the case of consumers who are **less than 13 years of age**, has affirmatively authorized the sale of the consumer’s personal information.’”).

⁸²Cal. Civ. Code § 1798.125(a).

⁸³Cal. Civ. Code § 1798.125(a)(1).

⁸⁴Cal. Civ. Code § 1798.125(a)(2). This sentence is inartfully worded but presumably speaks to any difference between the value provided, or price charged, to consumers, and the value of a consumer’s personal information. This meaning of this provision is likely to be fleshed out by the California Attorney General.

⁸⁵Cal. Civ. Code §§ 1798.125(b)(1), (3).

De minimis payments for information of great value thus are unlikely to pass muster. What constitutes fair value presumably will be clarified in regulations to be promulgated by the California Attorney General or through enforcement actions by the Attorney General.

Required privacy policy disclosures

The CCPA requires that businesses that collect, sell, or disclose California residents' personal information publicly inform consumers of their rights under the CCPA. These disclosures must be made in a business's "online privacy policy," "in any California-specific description of consumers' privacy rights," or, if the business does not maintain those policies, "on its Internet Web site."⁸⁶ A business must update these disclosures "at least once every 12 months."⁸⁷ The disclosure must include "one or more designated methods for submitting" disclosure requests under the statute.⁸⁸

Additionally, a business must disclose the categories of personal information that it has collected, sold, or disclosed in the previous 12 months.⁸⁹ A business that collects consumers' personal information is required to disclose:

- (1) the "categories of personal information it has collected about" consumers;
- (2) the "categories of sources from which the personal information is collected";
- (3) the "business or commercial purpose for collecting or selling personal information";
- (4) the "categories of third parties with whom the business shares personal information"; and
- (5) the "specific pieces of personal information the business has collected about that consumer."⁹⁰

A business that sells or discloses consumers' personal in-

⁸⁶Cal. Civ. Code § 1798.130(a)(5)

⁸⁷Cal. Civ. Code § 1798.130(a)(5).

⁸⁸Cal. Civ. Code § 1798.130(a)(5)(A).

⁸⁹The "categories of personal information" referred to in the privacy policy disclosure requirements "follow the definition of personal information in Section 1798.140." Cal. Civ. Code § 1798.130(c).

⁹⁰Cal. Civ. Code §§ 1798.110(c)(1)—(5). Although subsection (5) is technically included in the list of public disclosures that a business is required to make pursuant to section 1798.130, its inclusion is likely a mistake. The California legislature presumably did not intend for a business to publicly disclose "specific pieces of personal information" collected about an individual consumer. More likely, it intended to require disclosure

formation is required to disclose separately the categories of personal information that it has sold and disclosed within the last 12 months. Alternatively, if the business has not sold or disclosed consumer personal information in the preceding 12 months, it must “disclose that fact.”⁹¹

A business that sells consumers’ personal information must additionally include in its privacy policy, or in a California-specific description of privacy rights, a description of a consumer’s rights under the CCPA to opt-out and include the link titled “Do Not Sell My Personal Information” in the document.⁹²

A business that offers financial incentives for the collection, sale, or deletion of personal information must notify consumers of the incentives in its privacy policy or other public disclosure document.⁹³

Scope and exclusions

The California legislature mandated that the CCPA “be liberally construed to effectuate its purposes.”⁹⁴ It expressly preempts all rules, regulations, codes, ordinances, and other laws adopted by a city, county, city and county, municipality, or local agency regarding the collection and sale of consumers’ personal information by a business.⁹⁵ The CCPA is intended to supplement federal and state law, if permissible, but is not intended to apply if it would be preempted by, or in conflict with, federal law or the U.S. or California Constitution.⁹⁶

The CCPA provides that compliance with its obligations “shall not restrict a business’s ability” to comply with other applicable laws or a civil or criminal investigation, cooperate with law enforcement agencies, or exercise or defend legal claims.⁹⁷ The CCPA does not “apply where compliance by the business with the title would violate an evidentiary privilege

of the type of personal information it collects generally from consumers.

⁹¹Cal. Civ. Code §§ 1798.130(a)(5)(C)(i)—(ii).

⁹²Cal. Civ. Code § 1798.135(a)(2).

⁹³Cal. Civ. Code § 1798.125(b)(2); Cal. Civ. Code § 1798.130(a)(5)(A).

⁹⁴Cal. Civ. Code § 1798.194.

⁹⁵Cal. Civ. Code § 1798.180. Unlike the rest of the CCPA, which is set to take effect on January 1, 2020, this preemption provision became immediately effective upon enactment in 2018. *See id.* § 1798.199.

⁹⁶Cal. Civ. Code § 1798.196.

⁹⁷Cal. Civ. Code §§ 1798.145(a)(1)–(4).

under California law,” such as the attorney-client privilege, and the statute further does not apply to medical information or health information that is regulated by federal law, or information collected as part of a clinical trial subject to federal law.⁹⁸ The CCPA also does not apply to the sale of personal information “to or from a consumer reporting agency if that information is to be reported in, or used to generate, a consumer report” subject to the Fair Credit Reporting Act, 15 U.S.C. §§ 1681, *et seq.*, or to personal information collected, processed, sold or disclosed pursuant to the Gramm-Leahy-Bliley Act (Public Law 106-102), the California Financial Information Privacy Act, Cal. Fin. Code §§ 4050—4060, or the Driver’s Privacy Protection Act, 18 U.S.C. §§ 2721 *et seq.*⁹⁹

The CCPA also may not be applied to infringe upon the noncommercial free speech rights protected by the California Constitution.¹⁰⁰

Attorney General enforcement

⁹⁸Cal. Civ. Code § 1798.145(c)(1).

⁹⁹Cal. Civ. Code §§ 1798.145(d) - (f).

¹⁰⁰Cal. Civ. Code § 1798.145(k) (“The rights afforded to consumers and the obligations imposed on any business under this title shall not apply to the extent that they infringe on the noncommercial activities of a person or entity described in subdivision (b) of Section 2 of Article I of the California Constitution.”). Article I section 2(b) of the California Constitution provides that:

A publisher, editor, reporter, or other person connected with or employed upon a newspaper, magazine, or other periodical publication, or by a press association or wire service, or any person who has been so connected or employed, shall not be adjudged in contempt by a judicial, legislative, or administrative body, or any other body having the power to issue subpoenas, for refusing to disclose the source of any information procured while so connected or employed for publication in a newspaper, magazine or other periodical publication, or for refusing to disclose any unpublished information obtained or prepared in gathering, receiving or processing of information for communication to the public.

Nor shall a radio or television news reporter or other person connected with or employed by a radio or television station, or any person who has been so connected or employed, be so adjudged in contempt for refusing to disclose the source of any information procured while so connected or employed for news or news commentary purposes on radio or television, or for refusing to disclose any unpublished information obtained or prepared in gathering, receiving or processing of information for communication to the public.

As used in this subdivision, “unpublished information” includes information not disseminated to the public by the person from whom disclosure is sought, whether or not related information has been disseminated and includes, but is not limited to, all notes, outtakes, photographs, tapes or other data of whatever sort not itself disseminated to the public through a medium of communication,

The law delegates to the California Attorney General responsibilities analogous to those given the Federal Trade Commission by Congress under the Children's Online Privacy Protection Act (COPPA),¹⁰¹ Health Insurance Portability and Accountability Act (HIPAA)¹⁰² and Gramm-Leach-Bliley (GLB).¹⁰³ The Attorney General is delegated authority to adopt regulations,¹⁰⁴ provide opinions, and file suit to enforce the law (subject to affording businesses notices and an opportunity to cure within 30 days).¹⁰⁵ Given the number of ambiguities and drafting errors in the statute, and the limited nature of the private right of action (which only relates to security breaches), the Attorney General will have primary responsibility for interpreting and shaping enforcement priorities under the CCPA.

The statute contemplates that any business or third party may seek the opinion of the Attorney General for guidance on how to comply with the CCPA.¹⁰⁶

The law also authorizes the Attorney General to bring a civil action against businesses, service providers, or any other person that violates the CCPA.¹⁰⁷ A business "shall be in violation" if it "fails to cure any alleged violation within 30 days after being notified of noncompliance."¹⁰⁸ The Attorney General may seek injunctive relief and a civil penalty of not more than two thousand five hundred dollars (\$2,500) for each violation or seven thousand five hundred dollars (\$7,500) for each intentional violation.¹⁰⁹ While the penalties *per violation* are small, it remains to be seen how the Attorney General construes the term *violation*. Whether a violation is defined in terms of an incident or a single act or omission, for example, or the number of people impacted,

whether or not published information based upon or related to such material has been disseminated.

Cal. Const. Art. I § 2(b).

¹⁰¹See *supra* § 26.13[2][F].

¹⁰²See *supra* § 26.11.

¹⁰³See *supra* § 26.12[2]; see generally *supra* § 26.13[5] (analyzing FTC enforcement actions).

¹⁰⁴See Cal. Civ. Code § 1798.185.

¹⁰⁵See Cal. Civ. Code § 1798.155.

¹⁰⁶Cal. Civ. Code § 1798.155(a).

¹⁰⁷Cal. Civ. Code § 1798.155(b).

¹⁰⁸Cal. Civ. Code § 1798.155(a).

¹⁰⁹Cal. Civ. Code § 1798.155(b).

will be significant.

Revenue from litigation will be allocated to a Consumer Privacy Fund, which may be used exclusively to offset costs incurred by state courts and the California Attorney General in connection with the CCPA.¹¹⁰ This creates a potential conflict of interest, in that unless the legislature allocates funds expressly for all the new work to be done under the statute, there will be added pressure on the Attorney General's Office to pursue litigation—and to recover penalties in litigation.

Private right of action for data breaches

The CCPA creates a private right of action, with the possibility of recovering statutory damages, for consumers “whose nonencrypted or nonredacted personal information . . . is subject to an unauthorized access and exfiltration, theft, or disclosure as a result of the business's violation of the duty to implement and maintain reasonable security procedures and practices”¹¹¹ The private right of action created by the CCPA may be brought only for data

¹¹⁰Cal. Civ. Code § 1798.160.

¹¹¹Cal. Civ. Code § 1798.150(a)(1). *Personal information* in this section is defined by reference section 1798.81.5, which is narrower in scope than the CCPA's definition in section 1798.140(o). *Personal information* under section 1798.81.5 means either of the following:

- (A) An individual's first name or first initial and his or her last name in combination with any one or more of the following data elements, when either the name or the data elements are not encrypted or redacted:
 - (i) Social security number.
 - (ii) Driver's license number or California identification card number.
 - (iii) Account number, credit or debit card number, in combination with any required security code, access code, or password that would permit access to an individual's financial account.
 - (iv) Medical information.
 - (v) Health insurance information.
- (B) A username or email address in combination with a password or security question and answer that would permit access to an online account.

Cal. Bus. & Prof. Code § 1798.81.5(d)(1). *Personal information* does not include “publicly available information that is lawfully made available to the general public from federal, state, or local government records.” *Id.* § 1798.81.5(d)(4).

Medical information means any individually identifiable information, in electronic or physical form, regarding the individual's medical his-

breaches arising from a business's failure to maintain reasonable security measures, and not any other failures to comply with the CCPA.¹¹² What constitutes a *reasonable* security measure is not defined in the statute. Hence, unless the term is narrowed by regulations to be promulgated by the Attorney General, any time a California business suffers a security breach it will likely be sued in a lawsuit where plaintiffs will challenge both the security measures adopted and a business's adherence to those measures. In such cases, where the issue is contested, causation may raise factual questions that could make a case difficult to resolve on motion practice.

A person harmed by the data breach may bring an action to recover statutory damages in the range of \$100 - \$750 "per consumer per incident or actual damages, whichever is greater, injunctive or declaratory relief, and any other relief that a court deems proper."¹¹³ In assessing the amount of statutory damages, the court shall consider "any one or more of the relevant circumstances presented by any of the parties to the case, including, but not limited to, the nature and seriousness of the misconduct, the number of violations, the persistence of the misconduct, the length of time over which the misconduct occurred, the willfulness of the defendant's misconduct, and the defendant's assets, liabilities, and net worth."¹¹⁴ Nevertheless, a data breach impacting 100,000 consumers could invite putative class action suits seeking up to \$7,500,000, which seems disproportionate. And a breach impacting 1,000,000 state residents could result in a putative class action suit seeking \$750,000,000, where the plaintiffs, if successful, would be entitled to at least \$100,000,000. Given the wide range of exposure, the private cause of action created by the CCPA is likely to generate substantial litigation.

To bring a claim for statutory damages, either individually

tory or medical treatment or diagnosis by a health care professional. *Id.* § 1798.81.5(d)(2).

Health insurance information means an individual's insurance policy number or subscriber identification number, any unique identifier used by a health insurer to identify the individual, or any information in an individual's application and claims history, including any appeals records. *Id.* § 1798.81.5(d)(3).

¹¹²Cal. Civ. Code § 1798.150(c).

¹¹³Cal. Civ. Code § 1798.150(a)(1).

¹¹⁴Cal. Civ. Code § 1798.150(a)(2).

or as a putative class action suit, a consumer must provide a business “30 days’ written notice identifying the specific provisions of this title the consumer alleges have been or are being violated,” and allow the business 30 days to cure the violations. If within the 30 days the business actually cures the noticed violation (assuming a cure is possible) and provides the consumer an express written statement that the violations have been cured and that no further violations shall occur, then no action for individual statutory damages or class-wide statutory damages may be initiated against the business.¹¹⁵

This provision tracks the 30 day notice and cure period in the California Consumer Legal Remedies Act,¹¹⁶ a statute popular with class action counsel. Under that statute, some class action lawyers have become adept at framing claims for which a “cure” is impossible. It is unclear how, if at all, a breach which has occurred could be cured. Indeed, the statute acknowledges that possibility in framing requirements “[i]n the event a cure is possible”¹¹⁷ It remains to be seen whether the Attorney General will promulgate regulations to elaborate on the type of “cure” that would meet this requirement of the statute (such as measures to mitigate the consequences of a breach and minimize the risk of similar future breaches) or whether the issue will be fleshed out in litigation. Given the size of potential exposure and the ambiguity surrounding what constitutes *reasonable security*, a merely symbolic right to cure would be concerning.

If a business is able to cure and provides an express written statement to a consumer, but operates in breach of the express written statement, the consumer may initiate an action against the business to enforce the written statement and may pursue statutory damages for each breach of the express written statement, as well as any other violation of the title that postdates the written statement.¹¹⁸

No notice, however, is required for an individual consumer to initiate an action solely for actual pecuniary damages suf-

¹¹⁵Cal. Civ. Code § 1798.150(b).

¹¹⁶Cal. Civ. Code § 1782; *Laster v. T-Mobile USA, Inc.*, 407 F. Supp. 2d 1181, 1196 (S.D. Cal. 2005) (dismissing plaintiff’s claim with prejudice because of plaintiff’s failure to provide notice to defendants pursuant to section 1782(a)); *see generally supra* § 25.04[3].

¹¹⁷Cal. Civ. Code § 1798.150(b).

¹¹⁸Cal. Civ. Code § 1798.150(b).

ferred as a result of an alleged violation.¹¹⁹

Significantly, the cause of action established by section 1798.150 applies “only to violations as defined in subdivision (a) and shall not be based on violations of any other section of this title. Nothing in this title shall be interpreted to serve as the basis for a private right of action under any other law.”¹²⁰ What this means is that a violation of the statute could *not* form the basis for a claim under California’s notorious section 17200, which typically affords a cause of action for violation of other statutes, laws or regulations.¹²¹ The private enforcement right created by the CCPA thus is actually quite narrow. Nevertheless, the potential availability of statutory damages means that it will be heavily litigated by class action counsel seeking a generous settlement or award on behalf of a putative class of those whose information was exposed in a security breach. Further, the ambiguous nature of the standard of care—to “implement and maintain reasonable security procedures and practices”—means that regardless of culpability, any time a business experiences a security breach that exposes the information of California residents, class action counsel will have an incentive to file suit.

California law currently provides that any customer injured by a violation of its security breach notification statute may institute a civil action to recover damages¹²² or injunctive relief,¹²³ in addition to any other remedies that may be available.¹²⁴ Among other things, the breach of the notification statute itself could be actionable as an unfair trade practice under California law if damages can be

¹¹⁹Cal. Civ. Code § 1798.150(b).

¹²⁰Cal. Civ. Code § 1798.150(c).

¹²¹Cal. Bus. & Prof. §§ 17200 *et seq.* Section 17200 “borrows” violations from other laws by making them independently actionable as unfair competitive claims. *Korea Supply Co. v. Lockheed Martin Corp.*, 29 Cal. 4th 1134, 1143–45, 131 Cal. Rptr. 2d 29 (Cal. 2003). Under section 17200, “[u]nlawful acts are ‘anything that can properly be called a business practice and that at the same time is forbidden by law . . . be it civil, criminal, federal, state, or municipal, statutory, regulatory, or court-made,’ where court-made law is, ‘for example a violation of a prior court order.’” *Sybersound Records, Inc. v. UAV Corp.*, 517 F.3d 1137, 1151–52 (9th Cir. 2008) (citations omitted); *see generally supra* § 25.04[3].

¹²²Cal. Civil Code § 1798.84(b).

¹²³Cal. Civil Code § 1798.84(e).

¹²⁴Cal. Civil Code § 1798.84(g).

shown.¹²⁵ Absent any injury traceable to a company's failure to reasonably notify customers of a data breach, however, a plaintiff may not have standing to bring suit for a defendant's alleged failure to maintain reasonable security measures, at least in federal court.¹²⁶ The new cause of action created by the CCPA, by providing a remedy of statutory damages, will likely dramatically increase the number of California putative class action suits brought following a security breach. Given the liberal standing requirements for security breach cases in the Ninth Circuit,¹²⁷ some of these claims will be brought in federal court, although suits by California residents against California companies likely would need to be brought in state court, because of the lack of diversity jurisdiction, unless plaintiffs are able to also sue for violations of federal statutes.

The CCPA's requirement for contractual undertakings and obligations by service providers and third parties means it is also likely that the CCPA, if it takes effect, will result in litigation between or among *businesses*, *service providers* and *third parties*, as those terms are defined under the statute.

Data privacy class action litigation is analyzed in section

¹²⁵See Cal. Bus. & Prof. Code §§ 17200 *et seq.*; see generally *supra* §§ 27.01, 27.04[6] (discussing how the breach of an unrelated statute may be actionable under § 17200).

¹²⁶See, e.g., *Cahen v. Toyota Motor Corp.*, 717 F. App'x 720 (9th Cir. 2017) (affirming the lower court's ruling finding no standing to assert claims that car manufacturers equipped their vehicles with software that was susceptible to being hacked by third parties); *Antman v. Uber Technologies, Inc.*, Case No. 3:15-cv-01175-LB, 2018 WL 2151231 (N.D. Cal. May 10, 2018) (dismissing, with prejudice, plaintiff's claims, arising out of a security breach, for allegedly (1) failing to implement and maintain reasonable security procedures to protect Uber drivers' personal information and promptly notify affected drivers, in violation of Cal. Civ. Code §§ 1798.81, 1798.81.5, and 1798.82; (2) unfair, fraudulent, and unlawful business practices, in violation of California's Unfair Competition Law, Cal. Bus. & Prof. Code § 17200; (3) negligence; and (4) breach of implied contract, for lack of Article III standing, where plaintiff could not allege injury sufficient to establish Article III standing); see generally *infra* § 27.07 (analyzing claims raised in security breach litigation).

¹²⁷See, e.g., *In re Zappos.com, Inc.*, 888 F.3d 1020, 1023-30 (9th Cir. 2018) (holding that plaintiffs, whose information had been stolen by a hacker but who had not been victims of identity theft or financial fraud, nevertheless had Article III standing to maintain suit in federal court); see generally *infra* § 27.07 (comparing the relatively liberal standing requirements for security breach cases in the Ninth Circuit to case law from other circuits).

26.15. Security breach class action suits are analyzed in section 27.07.

26.14 Website Privacy Policies

26.14[1] In General¹

Privacy policies, statements or notices² are required for websites that collect personally identifying information³ in particular industries, when collected from children or pursuant to state law when collected from residents of states such as Texas and California. Sites that do not collect personal information (such as static websites that merely advertise products but have no interactive components) need not post policies. Likewise, many B2B sites will not need privacy statements because, even though they may collect confidential business information (and even trade secrets), they do not collect personally identifying information from consumers. By contrast, most websites targeted to consumers that operate on a national (or international) basis will need to have a privacy policy.

Privacy policies are required in the financial services⁴ and

[Section 26.14[1]]

¹This section addresses laws in effect in 2019. The California Consumer Privacy Act (CCPA), Cal. Civ. Code §§ 1798.100 to 1798.199, which is set to take effect on January 1, 2020 if not preempted by federal legislation, is separately analyzed in section 26.13A.

²Some privacy professionals prefer to refer to website privacy disclosures as privacy statements or notices, rather than policies, because a policy may dictate what a company's practices are or should be whereas a statement or notice merely sets forth those practices.

³There is no single definition of personally identifying or personally identifiable information (PII) that applies under all statutes and regulations. While companies may want (or need) to take more aggressive positions in litigation, for purposes of drafting a privacy policy it is generally a good idea to use the most expansive definition possible, which would cover any information that does or could identify a person or information about them. The FTC, in the context of behavioral advertising, suggested that the relevant criteria should be whether information reasonably could be associated with a particular consumer or device, not whether it is PII. See *supra* § 26.01. When in doubt greater transparency and more complete disclosures generally are advisable (subject to the caveat that a disclosure should not be so long and complicated that it is difficult for consumers to understand).

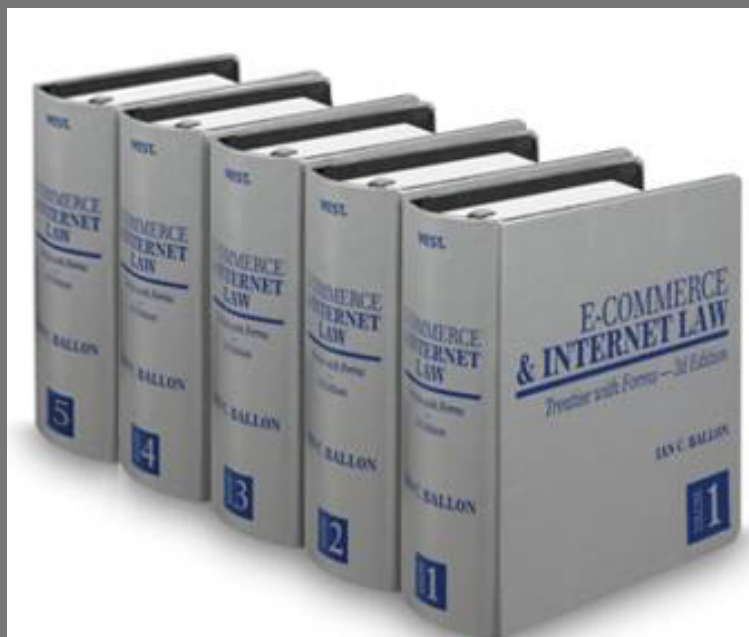
⁴See *supra* § 26.12[2].

E-COMMERCE & INTERNET LAW: TREATISE WITH FORMS 2D 2019

Ian C. Ballon

**NEW AND
IMPORTANT
FEATURES
FOR 2019
NOT FOUND
ELSEWHERE**

**THE PREEMINENT
INTERNET AND
MOBILE LAW
TREATISE FROM A
LEADING INTERNET
LITIGATOR — NOW A
5 VOLUME SET!**



To order call **1-888-728-7677**
or visit **legalsolutions.thomsonreuters.com**

Key Features of E-Commerce & Internet Law

- ◆ The California Consumer Privacy Act, GDPR, California IoT security statute, Vermont data broker registration law, Ohio safe harbor statute and other important privacy and cybersecurity laws
- ◆ Understanding conflicting law on mobile contract formation, unconscionability and enforcement of arbitration and class action waiver clauses
- ◆ The most comprehensive analysis of the TCPA's application to text messaging and its impact on litigation found anywhere
- ◆ Complete analysis of the Cybersecurity Information Sharing Act (CISA), state security breach statutes and regulations, and Defend Trade Secrets Act (DTSA) and their impact on screen scraping and database protection, cybersecurity information sharing and trade secret protection, privacy obligations and the impact that Terms of Use and other internet and mobile contracts may have in limiting the broad exemption from liability otherwise available under CISA
- ◆ Comprehensive and comparative analysis of the platform liability of Internet, mobile and cloud site owners, and service providers, for user content and misconduct under state and federal law
- ◆ Understanding the laws governing SEO and SEM and their impact on e-commerce vendors, including major developments involving internet advertising and embedded and sponsored links
- ◆ AI, screen scraping and database protection
- ◆ Strategies for defending cybersecurity breach and data privacy class action suits
- ◆ Copyright and Lanham Act fair use, patentable subject matter, combating genericide, right of publicity laws governing the use of a person's images and attributes, initial interest confusion, software copyrightability, damages in internet and mobile cases, the use of icons in mobile marketing, new rules governing fee awards, and the applicability and scope of federal and state safe harbors and exemptions
- ◆ How to enforce judgments against foreign domain name registrants
- ◆ Valuing domain name registrations from sales data
- ◆ Compelling the disclosure of the identity of anonymous and pseudonymous tortfeasors and infringers
- ◆ Exhaustive statutory and case law analysis of the Digital Millennium Copyright Act, the Communications Decency Act (including exclusions created by FOSTA-SESTA), the Video Privacy Protection Act, and Illinois Biometric Privacy Act
- ◆ Analysis of the CLOUD Act, BOTS Act, SPEECH Act, Consumer Review Fairness Act, N.J. Truth-in-Consumer Contract, Warranty and Notice Act, Family Movie Act and more
- ◆ Practical tips, checklists and forms that go beyond the typical legal treatise
- ◆ Clear, concise, and practical analysis

AN ESSENTIAL RESOURCE FOR ANY INTERNET AND MOBILE, INTELLECTUAL PROPERTY OR DATA PRIVACY/ CYBERSECURITY PRACTICE

E-Commerce & Internet Law is a comprehensive, authoritative work covering law, legal analysis, regulatory issues, emerging trends, and practical strategies. It includes practice tips and forms, nearly 10,000 detailed footnotes, and references to hundreds of unpublished court decisions, many of which are not available elsewhere. Its unique organization facilitates finding quick answers to your questions.

The updated new edition offers an unparalleled reference and practical resource. Organized into five sectioned volumes, the 59 chapters cover:

- Sources of Internet Law and Practice
- Intellectual Property
- Licenses and Contracts
- Data Privacy, Cybersecurity and Advertising
- The Conduct and Regulation of E-Commerce
- Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption
- Obscenity, Pornography, Adult Entertainment and the Protection of Children
- Theft of Digital Information and Related Internet Crimes
- Platform liability for Internet Sites and Services (Including Social Networks, Blogs and Cloud services)
- Civil Jurisdiction and Litigation

Distinguishing Features

- ◆ Clear, well written and with a practical perspective based on how issues actually play out in court (not available anywhere else)
- ◆ Exhaustive analysis of circuit splits and changes in the law combined with a common sense, practical approach for resolving legal issues, doing deals, documenting transactions and litigating and winning disputes
- ◆ Covers laws specific to the Internet and explains how the laws of the physical world apply to internet and mobile transactions and liability risks
- ◆ Addresses both law and best practices
- ◆ Comprehensive treatment of intellectual property, data privacy and mobile and Internet security breach law

Volume 1

Part I. Sources of Internet Law and Practice: A Framework for Developing New Law

- Chapter* 1. Context for Developing the Law of the Internet
 2. A Framework for Developing New Law
 3. [Reserved]

Part II. Intellectual Property

4. Copyright Protection in Cyberspace
 5. Database Protection, Screen Scraping and the Use of Bots and Artificial Intelligence to Gather Content and Information
 6. Trademark, Service Mark, Trade Name and Trade Dress Protection in Cyberspace
 7. Rights in Internet Domain Names

Volume 2

- Chapter* 8. Internet Patents
 9. Unique Intellectual Property Issues in Search Engine Marketing, Optimization and Related Indexing, Information Location Tools and Internet and Social Media Advertising Practices
 10. Misappropriation of Trade Secrets in Cyberspace
 11. Employer Rights in the Creation and Protection of Internet-Related Intellectual Property
 12. Privacy and Publicity Rights of Celebrities and Others in Cyberspace
 13. Idea Protection and Misappropriation

Part III. Licenses and Contracts

14. Documenting Internet Transactions: Introduction to Drafting License Agreements and Contracts
 15. Drafting Agreements in Light of Model and Uniform Contract Laws: UCITA, the UETA, Federal Legislation and the EU Distance Sales Directive
 16. Internet Licenses: Rights Subject to License and Limitations Imposed on Content, Access and Development
 17. Licensing Pre-Existing Content for Use Online: Music, Literary Works, Video, Software and User Generated Content Licensing Pre-Existing Content
 18. Drafting Internet Content and Development Licenses
 19. Website Development and Hosting Agreements
 20. Website Cross-Promotion and Cooperation: Co-Branding, Widget and Linking Agreements
 21. Obtaining Assent in Cyberspace: Contract Formation for Click-Through and Other Unilateral Contracts
 22. Structuring and Drafting Website Terms and Conditions
 23. ISP Service Agreements

Volume 3

- Chapter* 24. Software as a Service: On-Demand, Rental and Application Service Provider Agreements

Part IV. Privacy, Security and Internet Advertising

25. Introduction to Consumer Protection in Cyberspace
 26. Data Privacy
 27. Cybersecurity: Information, Network and Data Security
 28. Advertising in Cyberspace

Volume 4

- Chapter* 29. Email and Text Marketing, Spam and the Law of Unsolicited Commercial Email and Text Messaging

30. Online Gambling

Part V. The Conduct and Regulation of Internet Commerce

31. Online Financial Transactions and Payment Mechanisms
 32. Online Securities Law
 33. Taxation of Electronic Commerce
 34. Antitrust Restrictions on Technology Companies and Electronic Commerce
 35. State and Local Regulation of the Internet
 36. Best Practices for U.S. Companies in Evaluating Global E-Commerce Regulations and Operating Internationally

Part VI. Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption

37. Defamation, Torts and the Good Samaritan Exemption (47 U.S.C.A. § 230)
 38. Tort and Related Liability for Hacking, Cracking, Computer Viruses, Disabling Devices and Other Network Disruptions
 39. E-Commerce and the Rights of Free Speech, Press and Expression In Cyberspace

Part VII. Obscenity, Pornography, Adult Entertainment and the Protection of Children

40. Child Pornography and Obscenity
 41. Laws Regulating Non-Obscene Adult Content Directed at Children
 42. U.S. Jurisdiction, Venue and Procedure in Obscenity and Other Internet Crime Cases

Part VIII. Theft of Digital Information and Related Internet Crimes

43. Detecting and Retrieving Stolen Corporate Data
 44. Criminal and Related Civil Remedies for Software and Digital Information Theft
 45. Crimes Directed at Computer Networks and Users: Viruses and Malicious Code, Service Disabling Attacks and Threats Transmitted by Email

Volume 5

- Chapter* 46. Identity Theft

47. Civil Remedies for Unlawful Seizures

Part IX. Liability of Internet Sites and Service (Including Social Networks and Blogs)

48. Assessing and Limiting Liability Through Policies, Procedures and Website Audits
 49. The Liability of Platforms (including Website Owners, App Providers, eCommerce Vendors, Cloud Storage and Other Internet and Mobile Service Providers) for User Generated Content and Misconduct
 50. Cloud, Mobile and Internet Service Provider Liability and Compliance with Subpoenas and Court Orders
 51. Web 2.0 Applications: Social Networks, Blogs, Wiki and UGC Sites

Part X. Civil Jurisdiction and Litigation

52. General Overview of Cyberspace Jurisdiction
 53. Personal Jurisdiction in Cyberspace
 54. Venue and the Doctrine of Forum Non Conveniens
 55. Choice of Law in Cyberspace
 56. Internet ADR
 57. Internet Litigation Strategy and Practice
 58. Electronic Business and Social Network Communications in the Workplace, in Litigation and in Corporate and Employer Policies
 59. Use of Email in Attorney-Client Communications

“Should be on the desk of every lawyer who deals with cutting edge legal issues involving computers or the Internet.”

Jay Monahan

General Counsel, ResearchGate

ABOUT THE AUTHOR

IAN C. BALLON

Ian Ballon is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property and Technology Practice Group and is a litigator based in the firm's Silicon Valley and Los Angeles offices. He defends data privacy, cybersecurity breach, TCPA, and other Internet and mobile class action suits and litigates copyright, trademark, patent, trade secret, right of publicity, database and other intellectual property matters, including disputes involving Internet-related safe harbors and exemptions and platform liability.



Mr. Ballon was the recipient of the 2010 Vanguard Award from the State Bar of California's Intellectual Property Law Section. He also has been recognized by *The Los Angeles and San Francisco Daily Journal* as one of the Top 75 Intellectual Property litigators, Top Cybersecurity and Artificial Intelligence (AI) lawyers, and Top 100 lawyers in California.

In 2017 Mr. Ballon was named a "Groundbreaker" by *The Recorder* at its 2017 Bay Area Litigation Departments of the Year awards ceremony and was selected as an "Intellectual Property Trailblazer" by the *National Law Journal*.

Mr. Ballon was named as the Lawyer of the Year for information technology law in the 2019, 2018, 2016 and 2013 editions of *The Best Lawyers in America* and is listed in Legal 500 U.S., *The Best Lawyers in America* (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also serves as Executive Director of Stanford University Law School's Center for E-Commerce in Palo Alto.

Mr. Ballon received his B.A. *magna cum laude* from Tufts University, his J.D. *with honors* from George Washington University Law School and an LLM in international and comparative law from Georgetown University Law Center. He also holds the C.I.P.P./U.S. certification from the International Association of Privacy Professionals (IAPP).

In addition to *E-Commerce and Internet Law: Treatise with Forms 2d edition*, Mr. Ballon is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009), published by Thomson West (www.IanBallon.net).

He may be contacted at BALLON@GTLAW.COM and followed on Twitter and LinkedIn (@IanBallon).

Contributing authors: Parry Aftab, Ed Chansky, Francoise Gilbert, Tucker McCrady, Josh Raskin, Tom Smedinghoff and Emilio Varanini.

NEW AND IMPORTANT FEATURES FOR 2019

- > A comprehensive analysis of the **California Consumer Information Privacy Act**, **California's Internet of Things (IoT) security statute**, **Vermont's data broker registration law**, **Ohio's safe harbor** for companies with written information security programs, and other new state laws governing cybersecurity (chapter 27) and data privacy (chapter 26)
- > An exhaustive analysis of **FOSTA-SESTA** and what companies should do to maximize CDA protection in light of these new laws (chapter 37)
- > The **CLOUD Act** (chapter 50)
- > Understanding **the TCPA after ACA Int'l** and significant new cases & circuit splits (chapter 29)
- > Fully updated **50-state compendium** of security breach notification laws, with a **strategic approach** to handling notice to consumers and state agencies (chapter 27)
- > **Platform liability and statutory exemptions and immunities** (including a comparison of "but for" liability under the CDA and DMCA, and the latest law on secondary trademark and patent liability) (chapter 49)
- > Applying **the single publication rule** to websites, links and uses on social media (chapter 37)
- > The complex array of potential liability risks from, and remedies for, **screen scraping, database protection and use of AI to gather data and information online** (chapter 5)
- > State online dating and revenge porn laws (chapter 51)
- > **Circuit splits on Article III standing in cybersecurity litigation** (chapter 27)
- > Revisiting **sponsored link, SEO and SEM practices and liability** (chapter 9)
- > **Website and mobile accessibility** (chapter 48)
- > **The Music Modernization Act's Impact on copyright preemption and DMCA protection for pre-1972 musical works** (chapter 4)
- > **Compelling the disclosure of passwords and biometric information to unlock a mobile phone, tablet or storage device** (chapter 50)
- > Cutting through the jargon to make sense of **clickwrap, browsewrap, scrollwrap and sign-in wrap agreements (and what many courts and lawyers get wrong about online contract formation)** (chapter 21)
- > The latest case law, trends and strategy for **defending cybersecurity and data privacy class action suits** (chapters 25, 26, 27)
- > **Click fraud** (chapter 28)
- > Updated **Defend Trade Secrets Act** and **UTSA** case law (chapter 10)
- > **Drafting enforceable arbitration clauses and class action waivers** (with new sample provisions) (chapter 22)
- > **Applying the First Sale Doctrine to the sale of digital goods and information** (chapter 16)
- > **The GDPR, ePrivacy Directive and transferring data from the EU/EEA** (by Francoise Gilbert) (chapter 26)
- > **Patent law** (updated by Josh Raskin) (chapter 8)
- > **Music licensing** (updated by Tucker McCrady) (chapter 17)
- > **Mobile, Internet and Social Media contests & promotions** (updated by Ed Chansky) (chapter 28)
- > **Conducting a risk assessment and creating a Written Information Security Assessment Plan (WISP)** (by Thomas J. Smedinghoff) (chapter 27)

SAVE 20% NOW!!

To order call **1-888-728-7677**
or visit legalsolutions.thomsonreuters.com,
enter promo code **WPD20** at checkout

List Price: \$2,567.50
Discounted Price: \$2,054

CALIFORNIA'S IOT LAW ON THE SECURITY OF CONNECTED DEVICES

Excerpted from Chapter 27 (Cybersecurity: Information, Network and Data Security) of
E-Commerce and Internet Law: Legal Treatise with Forms 2d Edition
A 5-volume legal treatise by Ian C. Ballon (Thomson/West Publishing, www.IanBallon.net)

ARTIFICIAL INTELLIGENCE AND ROBOTICS
NATIONAL INSTITUTE
AMERICAN BAR ASSOCIATION
SANTA CLARA, CA
JANUARY 9-10, 2020

Ian C. Ballon
Greenberg Traurig, LLP

Silicon Valley: 1900 University Avenue, 5th Fl. East Palo Alto, CA 914303 Direct Dial: (650) 289-7881 Direct Fax: (650) 462-7881	Los Angeles: 1840 Century Park East, Ste. 1900 Los Angeles, CA 90067 Direct Dial: (310) 586-6575 Direct Fax: (310) 586-0575
---	--

Ballon@gtlaw.com
<www.ianballon.net>
LinkedIn, Twitter, Facebook: IanBallon



Ian C. Ballon

Shareholder

Internet, Intellectual Property & Technology Litigation

Admitted: California, District of Columbia and Maryland
Second, Third, Fourth, Fifth, Seventh, Ninth, Eleventh and Federal
Circuits

U.S. Supreme Court

JD, LL.M., CIPP/US

Ballon@gtlaw.com

LinkedIn, Twitter, Facebook: IanBallon

Silicon Valley

1900 University Avenue

5th Floor

East Palo Alto, CA 94303

T 650.289.7881

F 650.462.7881

Los Angeles

1840 Century Park East

Los Angeles, CA 90067

T 310.586.6575

F 310.586.0575

Ian Ballon is a litigator who is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property & Technology Practice and represents Internet, technology, mobile and other companies in intellectual property and internet- and mobile-related litigation, including the defense of data privacy, security breach, and TCPA class action suits. He is also the author of the leading treatise on Internet law, *E-Commerce and Internet Law: Treatise with Forms 2d edition*, the 5-volume set published by West (www.IanBallon.net). In addition, he is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009). He also serves as Executive Director of Stanford University Law School's Center for the Digital Economy, which hosts the annual Best Practices Conference where lawyers, scholars and judges are regularly featured and interact. A list of recent cases may be found at <http://www.gtlaw.com/Ian-C-Ballon-experience>.

Mr. Ballon was named the Lawyer of the Year for Information Technology Law in the 2019, 2018, 2016 and 2013 editions of Best Lawyers in America. In both 2018 and 2019 he was recognized as one of the Top 1,000 trademark attorneys in the world for his litigation practice by *World Trademark Review*. In addition, in 2019 he was named one of the top 20 Cybersecurity lawyers in California and in 2018 one of the Top Cybersecurity/Artificial Intelligence lawyers in California by the *Los Angeles and San Francisco Daily Journal*. He received the "Trailblazer" Award, Intellectual Property, 2017 from *The National Law Journal* and he has been recognized as a "Groundbreaker" in *The Recorder's* 2017 Litigation Departments of the Year Awards. In addition, he was the 2010 recipient of the State Bar of California IP Section's Vanguard Award for significant contributions to the development of intellectual property law (<http://ipsection.calbar.ca.gov/IntellectualPropertyLaw/IPVanguardAwards.aspx>). He is listed in Legal 500 U.S., The Best Lawyers in America (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also has been recognized by *The Daily Journal* as one of the Top 75 IP litigators in California in every year that the list has been published, from 2009 through 2019, and has been listed as a Northern California Super Lawyer every year from 2004 through 2018 and as one of the Top 100 lawyers in California. Mr. Ballon also holds the CIPP/US certification from the International Association of Privacy Professionals (IAPP).

Covered Entity's cybersecurity program" must be made available to the superintendent upon request.⁴³

A full set of the regulations is set forth in section 27.09[35]. Guidance on how to conduct a risk assessment and draft a written information security program is set forth in section 27.13.

27.04[6][L] California's IoT Law on the Security of Connected Devices

California's IoT data security law,¹ Cal. Civil Code §§ 1798.91.04 to 1798.91.06, which takes effect on January 1, 2020, will require a manufacturer of a connected device to equip the device with a reasonable security feature or features that are appropriate to the nature and function of the device, appropriate to the information it may collect, contain, or transmit, and designed to protect the device, and any information it contains, from unauthorized access, destruction, use, modification, or disclosure. Specifically, the law requires that as of January 1, 2020, a manufacturer² of

applicable and based on its Risk Assessment:

- (1) are designed to reconstruct material financial transactions sufficient to support normal operations and obligations of the Covered Entity; and
- (2) include audit trails designed to detect and respond to Cybersecurity Events that have a reasonable likelihood of materially harming any material part of the normal operations of the Covered Entity.

Id. The regulation also requires retention of records for three or five years, depending on the record. *See id.* § 500.06(b).

⁴³N.Y. Comp. Codes R. & Regs. tit. 23, § 500.02(d).

[Section 27.04[6][L]]

¹The Internet of Things (IoT) is a broad term used to refer to connected devices—such as smart refrigerators, smart televisions, wearable exercise monitors, self-driving cars, home security systems, and home or office climate control systems, among other things—that collect, store, or transfer information to other devices and networked computers, including personal data. *See generally supra* § 27.03B (explaining IoT).

²*Manufacturer* means “the person who manufactures, or contracts with another person to manufacture on the person's behalf, connected devices that are sold or offered for sale in California. For the purposes of this subdivision, a contract with another person to manufacture on the person's behalf does not include a contract only to purchase a connected device, or only to purchase and brand a connected device.” Cal. Civil Code § 1798.91.05(c).

a connected device³ equip the device with a reasonable security feature⁴ or features that are all of the following:

- (1) Appropriate to the nature and function of the device.
- (2) Appropriate to the information it may collect, contain, or transmit.
- (3) Designed to protect the device and any information contained therein from unauthorized access, destruction, use, modification, or disclosure.⁵

Subject to these requirements, if a connected device is equipped with a means for authentication⁶ outside a local area network, it will be deemed a *reasonable security feature* under the statute if either:

- (1) The preprogrammed password is unique to each device manufactured; or
- (2) The device contains a security feature that requires a user to generate a new means of authentication before access is granted to the device for the first time.⁷

The statute also includes four express exclusions. It may not be construed “to impose any duty upon the manufacturer of a connected device related to unaffiliated third-party software or applications that a user chooses to add to a connected device.”⁸

It may not be construed “to impose any duty upon a provider of an electronic store, gateway, marketplace, or other means of purchasing or downloading software or applications, to review or enforce compliance . . .” with the

³*Connected device* means “any device, or other physical object that is capable of connecting to the Internet, directly or indirectly, and that is assigned an Internet Protocol address or Bluetooth address.” Cal. Civil Code § 1798.91.05(b).

⁴A *security feature* is “a feature of a device designed to provide security for that device.” Cal. Civil Code § 1798.91.05(d).

⁵Cal. Civil Code § 1798.91.04(a). *Unauthorized access, destruction, use, modification, or disclosure* means “access, destruction, use, modification, or disclosure that is not authorized by the consumer.” *Id.* § 1798.91.05(e).

⁶*Authentication* means “a method of verifying the authority of a user, process, or device to access resources in an information system.” Cal. Civil Code § 1798.91.05(a).

⁷Cal. Civil Code § 1798.91.04(b).

⁸Cal. Civil Code § 1798.91.06(a).

statute.⁹

It may not be construed “to impose any duty upon the manufacturer of a connected device to prevent a user from having full control over a connected device, including the ability to modify the software or firmware running on the device at the user’s discretion.”¹⁰

And it may not be applied “to any connected device the functionality of which is subject to security requirements under federal law, regulations, or guidance promulgated by a federal agency pursuant to its regulatory enforcement authority.”¹¹

California’s IoT security law, which was the first U.S. statute to specifically address the security of information shared by connected devices, has been either applauded by security experts for taking a step in the right direction or criticized for focusing on adding “good” features instead of removing bad ones that subject devices to attacks.¹² It seems likely that other states or the federal government will seek to enact IoT regulations in the coming years.

27.05 The Payment Card Industry (PCI) Security Standard and Related State Laws

The Payment Card Industry Security Standards Council (PCI SSC) adopted the PCI Data Security Standard (PCI DDS) as a uniform set of security guidelines for major credit card companies, including American Express, Discover Financial Services, JCB International, MasterCard Worldwide and Visa Inc. International.¹ Any business processing, storing or transmitting payment card data must be PCI DSS compliant or risk being audited and fined and even losing the right to process credit card payments. Indeed, in 2006, Visa alone levied \$4.6 million in fines against businesses

⁹Cal. Civil Code § 1798.91.06(b).

¹⁰Cal. Civil Code § 1798.91.06(c).

¹¹Cal. Civil Code § 1798.91.06(d).

¹²See Adi Robertson, *California just became the first state with an Internet of Things cybersecurity law*, The Verge, Sept. 28, 2018 (quoting Robert Graham); Edward Kovacs, *California IoT Cybersecurity Bill Signed into Law*, SecurityWeek, Oct. 1, 2018 (quoting Graham as stating that the law “will do little [to] improve security, while doing a lot to impose costs and harm innovation.”).

[Section 27.05]

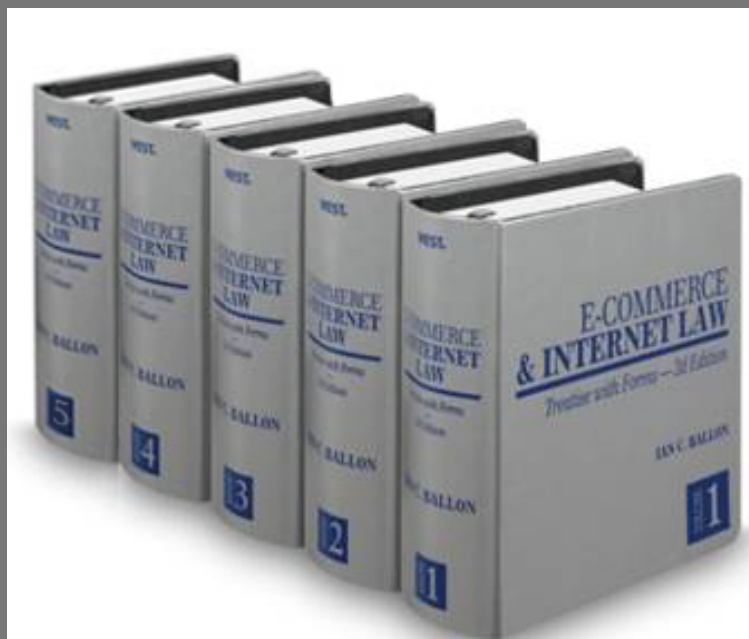
¹<https://www.pcisecuritystandards.org/>.

E-COMMERCE & INTERNET LAW: TREATISE WITH FORMS 2D 2019

Ian C. Ballon

**NEW AND
IMPORTANT
FEATURES
FOR 2019
NOT FOUND
ELSEWHERE**

**THE PREEMINENT
INTERNET AND
MOBILE LAW
TREATISE FROM A
LEADING INTERNET
LITIGATOR — NOW A
5 VOLUME SET!**



To order call **1-888-728-7677**
or visit **legalsolutions.thomsonreuters.com**

Key Features of E-Commerce & Internet Law

- ◆ The California Consumer Privacy Act, GDPR, California IoT security statute, Vermont data broker registration law, Ohio safe harbor statute and other important privacy and cybersecurity laws
- ◆ Understanding conflicting law on mobile contract formation, unconscionability and enforcement of arbitration and class action waiver clauses
- ◆ The most comprehensive analysis of the TCPA's application to text messaging and its impact on litigation found anywhere
- ◆ Complete analysis of the Cybersecurity Information Sharing Act (CISA), state security breach statutes and regulations, and Defend Trade Secrets Act (DTSA) and their impact on screen scraping and database protection, cybersecurity information sharing and trade secret protection, privacy obligations and the impact that Terms of Use and other internet and mobile contracts may have in limiting the broad exemption from liability otherwise available under CISA
- ◆ Comprehensive and comparative analysis of the platform liability of Internet, mobile and cloud site owners, and service providers, for user content and misconduct under state and federal law
- ◆ Understanding the laws governing SEO and SEM and their impact on e-commerce vendors, including major developments involving internet advertising and embedded and sponsored links
- ◆ AI, screen scraping and database protection
- ◆ Strategies for defending cybersecurity breach and data privacy class action suits
- ◆ Copyright and Lanham Act fair use, patentable subject matter, combating genericide, right of publicity laws governing the use of a person's images and attributes, initial interest confusion, software copyrightability, damages in internet and mobile cases, the use of icons in mobile marketing, new rules governing fee awards, and the applicability and scope of federal and state safe harbors and exemptions
- ◆ How to enforce judgments against foreign domain name registrants
- ◆ Valuing domain name registrations from sales data
- ◆ Compelling the disclosure of the identity of anonymous and pseudonymous tortfeasors and infringers
- ◆ Exhaustive statutory and case law analysis of the Digital Millennium Copyright Act, the Communications Decency Act (including exclusions created by FOSTA-SESTA), the Video Privacy Protection Act, and Illinois Biometric Privacy Act
- ◆ Analysis of the CLOUD Act, BOTS Act, SPEECH Act, Consumer Review Fairness Act, N.J. Truth-in-Consumer Contract, Warranty and Notice Act, Family Movie Act and more
- ◆ Practical tips, checklists and forms that go beyond the typical legal treatise
- ◆ Clear, concise, and practical analysis

AN ESSENTIAL RESOURCE FOR ANY INTERNET AND MOBILE, INTELLECTUAL PROPERTY OR DATA PRIVACY/ CYBERSECURITY PRACTICE

E-Commerce & Internet Law is a comprehensive, authoritative work covering law, legal analysis, regulatory issues, emerging trends, and practical strategies. It includes practice tips and forms, nearly 10,000 detailed footnotes, and references to hundreds of unpublished court decisions, many of which are not available elsewhere. Its unique organization facilitates finding quick answers to your questions.

The updated new edition offers an unparalleled reference and practical resource. Organized into five sectioned volumes, the 59 chapters cover:

- Sources of Internet Law and Practice
- Intellectual Property
- Licenses and Contracts
- Data Privacy, Cybersecurity and Advertising
- The Conduct and Regulation of E-Commerce
- Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption
- Obscenity, Pornography, Adult Entertainment and the Protection of Children
- Theft of Digital Information and Related Internet Crimes
- Platform liability for Internet Sites and Services (Including Social Networks, Blogs and Cloud services)
- Civil Jurisdiction and Litigation

Distinguishing Features

- ◆ Clear, well written and with a practical perspective based on how issues actually play out in court (not available anywhere else)
- ◆ Exhaustive analysis of circuit splits and changes in the law combined with a common sense, practical approach for resolving legal issues, doing deals, documenting transactions and litigating and winning disputes
- ◆ Covers laws specific to the Internet and explains how the laws of the physical world apply to internet and mobile transactions and liability risks
- ◆ Addresses both law and best practices
- ◆ Comprehensive treatment of intellectual property, data privacy and mobile and Internet security breach law

Volume 1

Part I. Sources of Internet Law and Practice: A Framework for Developing New Law

- Chapter* 1. Context for Developing the Law of the Internet
 2. A Framework for Developing New Law
 3. [Reserved]

Part II. Intellectual Property

4. Copyright Protection in Cyberspace
 5. Database Protection, Screen Scraping and the Use of Bots and Artificial Intelligence to Gather Content and Information
 6. Trademark, Service Mark, Trade Name and Trade Dress Protection in Cyberspace
 7. Rights in Internet Domain Names

Volume 2

- Chapter* 8. Internet Patents
 9. Unique Intellectual Property Issues in Search Engine Marketing, Optimization and Related Indexing, Information Location Tools and Internet and Social Media Advertising Practices
 10. Misappropriation of Trade Secrets in Cyberspace
 11. Employer Rights in the Creation and Protection of Internet-Related Intellectual Property
 12. Privacy and Publicity Rights of Celebrities and Others in Cyberspace
 13. Idea Protection and Misappropriation

Part III. Licenses and Contracts

14. Documenting Internet Transactions: Introduction to Drafting License Agreements and Contracts
 15. Drafting Agreements in Light of Model and Uniform Contract Laws: UCITA, the UETA, Federal Legislation and the EU Distance Sales Directive
 16. Internet Licenses: Rights Subject to License and Limitations Imposed on Content, Access and Development
 17. Licensing Pre-Existing Content for Use Online: Music, Literary Works, Video, Software and User Generated Content Licensing Pre-Existing Content
 18. Drafting Internet Content and Development Licenses
 19. Website Development and Hosting Agreements
 20. Website Cross-Promotion and Cooperation: Co-Branding, Widget and Linking Agreements
 21. Obtaining Assent in Cyberspace: Contract Formation for Click-Through and Other Unilateral Contracts
 22. Structuring and Drafting Website Terms and Conditions
 23. ISP Service Agreements

Volume 3

- Chapter* 24. Software as a Service: On-Demand, Rental and Application Service Provider Agreements

Part IV. Privacy, Security and Internet Advertising

25. Introduction to Consumer Protection in Cyberspace
 26. Data Privacy
 27. Cybersecurity: Information, Network and Data Security
 28. Advertising in Cyberspace

Volume 4

- Chapter* 29. Email and Text Marketing, Spam and the Law of Unsolicited Commercial Email and Text Messaging

30. Online Gambling

Part V. The Conduct and Regulation of Internet Commerce

31. Online Financial Transactions and Payment Mechanisms
 32. Online Securities Law
 33. Taxation of Electronic Commerce
 34. Antitrust Restrictions on Technology Companies and Electronic Commerce
 35. State and Local Regulation of the Internet
 36. Best Practices for U.S. Companies in Evaluating Global E-Commerce Regulations and Operating Internationally

Part VI. Internet Speech, Defamation, Online Torts and the Good Samaritan Exemption

37. Defamation, Torts and the Good Samaritan Exemption (47 U.S.C.A. § 230)
 38. Tort and Related Liability for Hacking, Cracking, Computer Viruses, Disabling Devices and Other Network Disruptions
 39. E-Commerce and the Rights of Free Speech, Press and Expression In Cyberspace

Part VII. Obscenity, Pornography, Adult Entertainment and the Protection of Children

40. Child Pornography and Obscenity
 41. Laws Regulating Non-Obscene Adult Content Directed at Children
 42. U.S. Jurisdiction, Venue and Procedure in Obscenity and Other Internet Crime Cases

Part VIII. Theft of Digital Information and Related Internet Crimes

43. Detecting and Retrieving Stolen Corporate Data
 44. Criminal and Related Civil Remedies for Software and Digital Information Theft
 45. Crimes Directed at Computer Networks and Users: Viruses and Malicious Code, Service Disabling Attacks and Threats Transmitted by Email

Volume 5

- Chapter* 46. Identity Theft

47. Civil Remedies for Unlawful Seizures

Part IX. Liability of Internet Sites and Service (Including Social Networks and Blogs)

48. Assessing and Limiting Liability Through Policies, Procedures and Website Audits
 49. The Liability of Platforms (including Website Owners, App Providers, eCommerce Vendors, Cloud Storage and Other Internet and Mobile Service Providers) for User Generated Content and Misconduct
 50. Cloud, Mobile and Internet Service Provider Liability and Compliance with Subpoenas and Court Orders
 51. Web 2.0 Applications: Social Networks, Blogs, Wiki and UGC Sites

Part X. Civil Jurisdiction and Litigation

52. General Overview of Cyberspace Jurisdiction
 53. Personal Jurisdiction in Cyberspace
 54. Venue and the Doctrine of Forum Non Conveniens
 55. Choice of Law in Cyberspace
 56. Internet ADR
 57. Internet Litigation Strategy and Practice
 58. Electronic Business and Social Network Communications in the Workplace, in Litigation and in Corporate and Employer Policies
 59. Use of Email in Attorney-Client Communications

“Should be on the desk of every lawyer who deals with cutting edge legal issues involving computers or the Internet.”

Jay Monahan

General Counsel, ResearchGate

ABOUT THE AUTHOR

IAN C. BALLON

Ian Ballon is Co-Chair of Greenberg Traurig LLP's Global Intellectual Property and Technology Practice Group and is a litigator based in the firm's Silicon Valley and Los Angeles offices. He defends data privacy, cybersecurity breach, TCPA, and other Internet and mobile class action suits and litigates copyright, trademark, patent, trade secret, right of publicity, database and other intellectual property matters, including disputes involving Internet-related safe harbors and exemptions and platform liability.



Mr. Ballon was the recipient of the 2010 Vanguard Award from the State Bar of California's Intellectual Property Law Section. He also has been recognized by *The Los Angeles and San Francisco Daily Journal* as one of the Top 75 Intellectual Property litigators, Top Cybersecurity and Artificial Intelligence (AI) lawyers, and Top 100 lawyers in California.

In 2017 Mr. Ballon was named a "Groundbreaker" by *The Recorder* at its 2017 Bay Area Litigation Departments of the Year awards ceremony and was selected as an "Intellectual Property Trailblazer" by the *National Law Journal*.

Mr. Ballon was named as the Lawyer of the Year for information technology law in the 2019, 2018, 2016 and 2013 editions of *The Best Lawyers in America* and is listed in Legal 500 U.S., *The Best Lawyers in America* (in the areas of information technology and intellectual property) and Chambers and Partners USA Guide in the areas of privacy and data security and information technology. He also serves as Executive Director of Stanford University Law School's Center for E-Commerce in Palo Alto.

Mr. Ballon received his B.A. *magna cum laude* from Tufts University, his J.D. *with honors* from George Washington University Law School and an LLM in international and comparative law from Georgetown University Law Center. He also holds the C.I.P.P./U.S. certification from the International Association of Privacy Professionals (IAPP).

In addition to *E-Commerce and Internet Law: Treatise with Forms 2d edition*, Mr. Ballon is the author of *The Complete CAN-SPAM Act Handbook* (West 2008) and *The Complete State Security Breach Notification Compliance Handbook* (West 2009), published by Thomson West (www.IanBallon.net).

He may be contacted at BALLON@GTLAW.COM and followed on Twitter and LinkedIn (@IanBallon).

Contributing authors: Parry Aftab, Ed Chansky, Francoise Gilbert, Tucker McCrady, Josh Raskin, Tom Smedinghoff and Emilio Varanini.

NEW AND IMPORTANT FEATURES FOR 2019

- > A comprehensive analysis of the **California Consumer Information Privacy Act**, **California's Internet of Things (IoT) security statute**, **Vermont's data broker registration law**, **Ohio's safe harbor** for companies with written information security programs, and other new state laws governing cybersecurity (chapter 27) and data privacy (chapter 26)
- > An exhaustive analysis of **FOSTA-SESTA** and what companies should do to maximize CDA protection in light of these new laws (chapter 37)
- > The **CLOUD Act** (chapter 50)
- > Understanding **the TCPA after ACA Int'l** and significant new cases & circuit splits (chapter 29)
- > Fully updated **50-state compendium** of security breach notification laws, with a **strategic approach** to handling notice to consumers and state agencies (chapter 27)
- > **Platform liability and statutory exemptions and immunities** (including a comparison of "but for" liability under the CDA and DMCA, and the latest law on secondary trademark and patent liability) (chapter 49)
- > Applying **the single publication rule** to websites, links and uses on social media (chapter 37)
- > The complex array of potential liability risks from, and remedies for, **screen scraping, database protection and use of AI to gather data and information online** (chapter 5)
- > State online dating and revenge porn laws (chapter 51)
- > **Circuit splits on Article III standing in cybersecurity litigation** (chapter 27)
- > Revisiting **sponsored link, SEO and SEM practices and liability** (chapter 9)
- > **Website and mobile accessibility** (chapter 48)
- > **The Music Modernization Act's impact on copyright preemption and DMCA protection for pre-1972 musical works** (chapter 4)
- > **Compelling the disclosure of passwords and biometric information to unlock a mobile phone, tablet or storage device** (chapter 50)
- > Cutting through the jargon to make sense of **clickwrap, browsewrap, scrollwrap and sign-in wrap agreements (and what many courts and lawyers get wrong about online contract formation)** (chapter 21)
- > The latest case law, trends and strategy for **defending cybersecurity and data privacy class action suits** (chapters 25, 26, 27)
- > **Click fraud** (chapter 28)
- > Updated **Defend Trade Secrets Act** and **UTSA** case law (chapter 10)
- > **Drafting enforceable arbitration clauses and class action waivers** (with new sample provisions) (chapter 22)
- > **Applying the First Sale Doctrine to the sale of digital goods and information** (chapter 16)
- > **The GDPR, ePrivacy Directive and transferring data from the EU/EEA** (by Francoise Gilbert) (chapter 26)
- > **Patent law** (updated by Josh Raskin) (chapter 8)
- > **Music licensing** (updated by Tucker McCrady) (chapter 17)
- > **Mobile, Internet and Social Media contests & promotions** (updated by Ed Chansky) (chapter 28)
- > **Conducting a risk assessment and creating a Written Information Security Assessment Plan (WISP)** (by Thomas J. Smedinghoff) (chapter 27)

SAVE 20% NOW!!

To order call **1-888-728-7677**
or visit legalsolutions.thomsonreuters.com,
enter promo code **WPD20** at checkout

List Price: \$2,567.50
Discounted Price: \$2,054

Chapter 4

Lawyers' Legal Obligations to Provide Data Security

Thomas J. Smedinghoff and Ruth Hill Bro

Virtually all of the daily transactions and key records of a business (whether a law firm, corporation, public interest entity, or the like) are created, used, communicated, and stored in electronic form using networked computer technology. Although such technology provides the business with tremendous economic benefits, including reduced costs and increased productivity, it also creates significant potential vulnerabilities that can adversely affect the business, its clients and customers, and other entities with whom it interacts.

Creating, using, communicating, and storing information in electronic form greatly increases the potential for unauthorized access, use, disclosure, alteration, loss, or destruction of the information. Front-page news stories about data security missteps made by companies, government agencies, and other businesses (including law firms) are a testament to the growing significance of this problem and should serve as a wake-up call for lawyers in all practice settings. Insert your firm's name, or your client's name, in the most recent data breach headline, and the risk of not taking sufficient security steps (especially those that are legally required) becomes all too real.

I. Overview

A. What Is Data Security?

The concept of "security" refers to an entity's implementation and maintenance of *security controls* to protect one or more of its *assets* (such as

62 THE ABA CYBERSECURITY HANDBOOK

buildings, equipment, cargo, inventory, and people) from *threats*. Information security (also referred to as “cybersecurity” or “data security”) involves the implementation of security controls to protect a business’s *digital assets*. It has been generally described as “the protection of *information and information systems* from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability.”¹ Thus, information security involves the protection of both (1) *information systems*—that is, computer systems, networks, and software—and (2) the *electronic records, data, messages, and other information* that are typically recorded on, processed by, communicated via, stored in, shared by, or received from such information systems.

The *objectives* of using security measures can be defined in terms of either the positive results to be achieved or the negative consequences to be avoided. The positive results to be achieved are typically described as ensuring the *confidentiality, integrity, and availability* of information.² The harms to be avoided, as noted above, are often described as unauthorized access, use, disclosure, disruption, modification, or destruction.³

Achieving these objectives involves implementing security measures designed to protect systems and information from the various threats they face. The kinds of threats, where they come from, what is at risk, and the seriousness of the consequences will, of course, vary greatly from case to case. But responding to those threats with appropriate security measures is the focus of the duty to provide security.

Measures designed to protect the security of information systems and data are generally grouped into the following three categories (based on the nature of the control):

1. NIST, NISTIR 7298, REV. 2, GLOSSARY OF KEY INFORMATION SECURITY TERMS 94 (May 2013) (definition of “information security”) (emphasis added). *See also* Federal Information Security Management Act (FISMA), 44 U.S.C. § 3542(b)(1) (definition of “information security”).

2. *See, e.g.*, FISMA, 44 U.S.C. § 3542(b)(1); Health Insurance Portability and Accountability Act of 1996 (HIPAA) Security Regulations, 45 C.F.R. § 164.306(a)(1).

3. *See supra* note 1.

- ***Physical security controls.*** These security measures are designed to protect the tangible items that comprise the physical computer systems, networks, and storage devices that process, communicate, and store the data, including servers, devices used to access the system, storage devices, and the like. Physical security controls are intended to prevent unauthorized persons from entering that environment and to help protect against natural disasters. One regulation defines physical safeguards as “physical measures, policies, and procedures to protect a covered entity’s or business associate’s electronic information systems and related buildings and equipment, from natural and environmental hazards, and unauthorized intrusion.”⁴ Examples of physical security controls include fences, walls, and other barriers; locks, safes, and vaults; armed guards; sensors; and alarm bells.
- ***Technical security controls.*** These security measures typically involve the use of software and data safeguards incorporated into computer hardware, software, and related devices. These measures are designed to ensure system availability, control access to systems and information, authenticate persons seeking access, protect the integrity of information communicated via and stored on the system, and ensure confidentiality where appropriate. Examples include firewalls, intrusion detection software, access control software, antivirus software, passwords, PIN numbers, smart cards, biometric tokens, and encryption processes.
- ***Administrative security controls.*** Sometimes referred to as “procedural” or “organizational” controls, these security measures consist of written policies, procedures, standards, guidelines, and supplemental administrative controls to guide conduct, prevent unauthorized access, and provide an acceptable level of protection for computing resources and data. Administrative security measures frequently include personnel management, employee use policies, training, discipline, and informing people how to conduct day-to-day operations.

4. HIPAA Security Regulations, 45 C.F.R. § 164.304.

64 THE ABA CYBERSECURITY HANDBOOK

Within each of these three categories, security measures are further classified into the following three separate categories (based on their timing regarding the risks and threats they are designed to address):

- **Preventive** security measures are designed to prevent the occurrence of events that compromise security. Examples include a lock on a door (to prevent access to a room containing computer equipment) or a firewall (to prevent unauthorized online access to a computer system).
- **Detective** security measures are designed to identify security breaches after they have occurred. Examples include a smoke alarm (to detect a fire) or intrusion detection software (to detect and track unauthorized online access to a computer system).
- **Reactive** security measures are designed to respond to a security breach and typically include efforts to stop or contain the breach, identify the party or parties involved, and allow recovery of information that is lost or damaged. Examples include calling the police (after an alarm detects that a burglary is in process) or shutting down a computer system (after intrusion detection software determines that an unauthorized user has obtained access to the system).

B. Security Law: The Basic Security Obligations

Concerns about individual privacy, accountability for financial information, the authenticity and integrity of transaction data, and the need to protect the confidentiality and security of sensitive business and client data are driving the enactment of new laws and regulations designed to ensure that all businesses adequately address the security of the data in their possession or under their control. Taken as a group, those laws and regulations impose two fundamental obligations on most businesses:

- The duty to provide security for their data; and
- The duty to warn of security breaches that occur.

The thesis of this chapter is that all businesses (including law firms), whether regulated or not, are generally subject to these legal duties regarding the security of the data in their possession or under their control. The following sections explain the source and scope of those duties.

II. The Duty to Provide Data Security

A. What Is the Duty?

The law often simply refers to the basic legal duty to provide data security as an obligation to implement “reasonable” or “appropriate” security measures designed to ensure the *confidentiality, integrity, and availability* of information. For example, several state security laws, such as in California, generally impose a duty to implement “*reasonable* security procedures and practices.”⁵ At the federal level, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requires “*reasonable and appropriate*” security,⁶ and the Gramm-Leach-Bliley (GLB) security regulations require security “*appropriate* to the size and complexity of the bank and the nature and scope of its activities.”⁷

The focus on the reasonableness or appropriateness of security makes clear that the law recognizes that security is a relative concept: what qualifies as reasonable or appropriate security varies with the situation. Thus, the law typically provides little or no guidance on what specific security measures are required or on how much security a business should implement to satisfy those legal obligations. Although some laws include specific requirements for particular security measures that must be implemented,⁸ the laws generally provide no safe harbors. Accordingly, the choice of security measures and technology can vary depending on the situation.

B. To Whom Does the Duty Apply?

Generally, the duty to provide security applies to all businesses, including law firms.

Certain sectors of the U.S. economy are, of course, subject to extensive regulations regarding data security. The most obvious examples are the

5. CAL. CIV. CODE § 1798.81.5(b) (emphasis added).

6. 42 U.S.C. § 1320d-2(d)(2) (emphasis added).

7. 12 C.F.R. pt. 208, app. D-2, pt. II.A (Federal Reserve System) (emphasis added). *See also* other GLB-implementing security regulations: 12 C.F.R. pt. 30, app. B, pt. II.A (OCC); 12 C.F.R. pt. 364, app. B, pt. II.A. (FDIC); and 16 C.F.R. § 314.3(a) (FTC) (adding “sensitivity of any customer information at issue” to the other factors in determining what is “appropriate”).

8. For example, the Massachusetts security regulations require implementation of firewalls, the use of virus software, and, in certain cases, the use of encryption. *See* 201 MASS. CODE REGS. 17.00.

66 THE ABA CYBERSECURITY HANDBOOK

financial sector,⁹ the healthcare sector,¹⁰ the federal government sector,¹¹ and other critical infrastructure sectors.¹² But there also is no doubt that unregulated businesses are subject to data security obligations.

One need look no further than the last 15 years of Federal Trade Commission (FTC) enforcement actions, as well as recent state attorney general enforcement actions, to see that numerous nonregulated businesses have been targeted for failing to provide appropriate security for their own data. Examples include software vendors (Oracle, Microsoft, Guidance Software), consumer electronics companies (ASUS, TRENDnet, HTC America, Genica/Computer Geeks), mobile app developers (Snapchat, Fandango, Credit Karma), clothing/shoe retailers (Guess, Life is Good, DSW), music retailers (Tower Records), animal supply retailers (Petco), general merchandise stores (Target, BJ's Wholesale, TJX Companies), restaurant and entertainment establishments (Dave & Busters, Briar Group), social media and networking sites (Twitter, Facebook, and Ashley Madison), transcription services (GMR), bookstores (Barnes & Noble), property management firms (Maloney Properties, Inc.), and hotels (Wyndham).¹³

In addition to the federal- and state-level unfair or deceptive trade practice statutes that often support these enforcement actions, many state security laws and regulations expressly apply to “any business” or “any person” that maintains certain types of data. Of course, this includes law firms.

Moreover, as discussed below, many sector-specific security regulations may be imposed on law firms through their client relationships. For example, the HIPAA regulations in the healthcare sector and the GLB

9. Subject to GLB, Pub. L. No. 106-102, §§ 501 and 505(b), 15 U.S.C. §§ 6801, 6805, and implementing security regulations; *see supra* note 7.

10. Subject to HIPAA, 42 U.S.C. § 1320d-2, and HIPAA Security Regulations, 45 C.F.R. pt. 164.

11. Subject to FISMA, 44 U.S.C. §§ 3541–3549.

12. *See* Exec. Order No. 13,636, Improving Critical Infrastructure Cybersecurity, 78 Fed. Reg. 11,739 (Feb. 19, 2013), *available at* <https://www.whitehouse.gov/the-press-office/2013/02/12/executive-order-improving-critical-infrastructure-cybersecurity>.

13. *See, e.g.*, FTC, Data Security, <https://www.ftc.gov/datasecurity> (for list of all FTC data security cases and enforcement actions); May 2017 state attorneys general settlement agreement with Target Corp., http://www.illinoisattorneygeneral.gov/pressroom/2017_05/17-AVC-0008TargetCorporation.pdf.

regulations in the financial sector both require that entities governed by those regulations push down certain security obligations to their service providers (which includes law firms) who access the protected data. In addition, the HIPAA regulations have been revised to impose security obligations directly on “covered entities” providing services to health-care companies.

C. What Is the Source of the Duty?

There is no single law, statute, or regulation that governs the obligations of a business or law firm to provide security for the information in its possession or under its control. Instead, legal obligations to implement data security measures are found in an ever-expanding patchwork of state, federal, and international laws, regulations, and enforcement actions, as well as in common-law duties and other express and implied obligations to provide “reasonable” or “appropriate” security for business data.

Some laws seek to protect the business and its owners, shareholders, investors, and business partners. Other laws focus on the interests of employees, customers, and prospects. In some cases, governmental regulatory interests or evidentiary requirements are at stake. Many of the requirements are industry-specific (e.g., focused on the financial sector or the healthcare sector) or data-specific (e.g., focused on personal information or financial data). Some laws focus only on public companies.

When viewed as a group, however, such laws and regulations provide ever-expanding coverage of most business activity. The most common sources of obligations to provide security include the following:

Statutes and Regulations.¹⁴ Numerous statutes and regulations impose obligations to provide data security. Sometimes these statutes and regulations use recognizable terms such as “security” or “safeguards,” but in many cases they are subtler by using attributes of security, such as

14. See Appendices A (Federal Statutes), B (State Statutes), C (Federal Regulations), and D (State Regulations) of this Handbook for examples of such statutes and regulations. See also Appendices H (CFPB Decision and Consent Decree), I (FTC Decisions and Consent Decrees), and J (SEC Decision and Consent Decree) for government enforcement actions under certain statutes and regulations.

68 THE ABA CYBERSECURITY HANDBOOK

“authenticate,” “integrity,” “confidentiality,” “availability of data,” and the like. Such statutes and regulations include the following:

- *Privacy laws and regulations*, which typically include provisions governing the security of the personal data covered by the applicable law.
- *Security laws and regulations*, such as the state-level security laws that impose a general obligation on businesses to protect the security of certain personal data they maintain about individuals and/or that regulate the communication or destruction of certain data;
- *E-transaction laws*, which are designed to ensure the enforceability and compliance of electronic documents generally;
- *Corporate governance legislation and regulations*, which are designed to protect public companies and their shareholders, investors, and business partners;
- *Unfair business practice laws*, at both the federal and state level, and precedent set by related government enforcement actions; and
- *Sector-specific regulations*, such as the HIPAA security regulations and the GLB Safeguard Rules, which impose security obligations regarding specific data in the healthcare and financial sectors, respectively.

Common-Law Obligations.¹⁵ For years, commentators have argued that there is a common-law duty to provide appropriate security for corporate and personal data, the breach of which constitutes a tort. Courts are beginning to accept that view. In one case, for example, the court held that “defendant did owe plaintiffs a duty to protect them from identity theft by providing some safeguards to ensure the security of their most essential confidential identifying information.”¹⁶ In another case of particular significance to lawyers, the court allowed plaintiffs to proceed on a “negligent misrepresentation” claim based on the theory that the defendants made implied representations that they had implemented the security measures required by industry practice to safeguard personal and financial information.¹⁷

15. See, e.g., selected cases listed in Appendix G (Court Decisions re Duty to Provide Data Security) of this Handbook.

16. *Bell v. Mich. Council*, 205 Mich. App. LEXIS 353, at *16 (Mich. App. Feb. 15, 2005).

17. *In re TJX Cos. Retail Sec. Breach Litig.*, 524 F. Supp. 2d 83 (D. Mass. 2007).

Rules of Evidence. Providing appropriate security to ensure the integrity of electronic records (and the identity of the creator, sender, or signer of the record) can be critical to securing the admission of an electronic record in evidence in a dispute. This conclusion is supported by the form requirement for an “original” in electronic transaction laws,¹⁸ the evidence rules regarding authentication,¹⁹ and case law addressing evidentiary authentication requirements.²⁰

Rules of Professional Responsibility. Lawyers are, of course, subject to rules of professional responsibility. Such rules generally are patterned after the ABA Model Rules of Professional Conduct, which were modified in August 2012 by the ABA Commission on Ethics 20/20 to provide updated guidance regarding lawyers’ use of technology and confidentiality obligations.²¹

Contractual Obligations. Businesses frequently try to satisfy (at least in part) their obligation to protect data by entering contracts with third parties who will possess, or have access to, their business data. This is particularly common in outsourcing agreements where the data will be processed by a third party. Several laws, such as the generally applicable Massachusetts data security regulations²² or the financial sector’s GLB Safeguard Rules, mandate that the business impose appropriate security obligations on the third party with access to its data. In other cases, businesses must comply with the requirements of certain technical security standards. Examples include the Payment Card Industry Data Security Standard (PCI Standard),²³ to which merchants must agree as a condition of accepting credit cards.

18. See, e.g., Unif. Electronic Transactions Act (UETA) § 12(d), http://www.uniformlaws.org/shared/docs/electronic%20transactions/ueta_final_99.pdf; Electronic Signatures in Global and National Commerce Act (E-SIGN) 15 U.S.C. § 7001(d)(3), available at <https://www.gpo.gov/fdsys/pkg/PLAW-106publ229/pdf/PLAW-106publ229.pdf>.

19. See, e.g., FED. R. EVID. 901(a).

20. See, e.g., *Am. Express v. Vinhnee*, 336 B.R. 437 (B.A.P. 9th Cir. 2005); *Lorraine v. Markel*, 241 F.R.D. 534 (D. Md. May 4, 2007).

21. These rules and other law applicable specifically to lawyers are covered in Chapter 6 of this Handbook.

22. Standards for the Protection of Personal Information of Residents of the Commonwealth, 201 MASS. CODE REGS. 17.00 *et seq.* (2012) [hereinafter *Mass. Standards for the Protection of Personal Info*], available at <http://www.mass.gov/ocabr/docs/idtheft/201cmr1700reg.pdf>.

23. See PCI Sec. Standards Council, <https://www.pcisecuritystandards.org>.

70 THE ABA CYBERSECURITY HANDBOOK

Self-Imposed Obligations. In many cases, security obligations are self-imposed. Through statements in privacy notices, on websites, in advertising materials, or elsewhere, businesses often make representations regarding the level of security they provide for their data (particularly personal data collected from persons to whom the statements are made). By making such statements, businesses impose on themselves an obligation to comply with the standard they have told the public that they meet. If those statements are not true, or are misleading, they may become deceptive trade practices under section 5 of the FTC Act or equivalent state laws.

Obligations Pushed Down from Clients. In some cases, data security laws and regulations do not apply directly to law firms, but might apply indirectly (e.g., because of law firm clients who themselves are subject to certain sector-specific security regulations). Such regulations frequently impose on covered businesses an obligation to push down certain security requirements to third parties with whom they do business or who otherwise are involved in processing their data. This approach is increasingly becoming a source of data security obligations for law firms. For example, a law firm must comply with security requirements imposed on its financial or healthcare sector clients where those requirements must be passed down to the law firm. In many such cases, client requirements to satisfy certain security regulations motivate client audits of law firm security measures.

Thus, the duty of any business (and any law firm) to provide security may come from several different sources and several different jurisdictions—each perhaps regulating a different aspect of the business’s information—but the net result is a general obligation to provide security for all business data and information systems. In other words, information security is not just good business practice; it is a legal obligation.

D. What Data Is Covered?

All types of law firm and client business information should be protected by appropriate security; such information includes financial information, personal information, tax-related records, employee information, transaction information, information obtained from or produced for clients, litigation information (including what is obtained in discovery), and other confidential information.

When examining particular security laws that may apply to a business or a law firm, it is important to note that such laws will frequently focus on a certain category of information. Commonly addressed categories include the following:

- ***Attorney-client data.*** Any client-related data held by the law firm is likely to be subject to numerous legal obligations to protect the security of that data. In addition to the legal obligations discussed here (which may be imposed directly on the law firm, or indirectly by clients obligated to push down their own imposed obligations), lawyers also have ethical obligations to protect client data.²⁴
- ***Personal data.*** The obligation to provide adequate security for personal data collected, used, communicated, or stored by a business is a critical component of all privacy laws as well as sector-specific privacy regulations, such as those governing healthcare or personal financial records.
- ***Financial data.*** Corporate governance laws designed to protect the company and its shareholders, investors, and business partners (such as Sarbanes-Oxley and implementing regulations) require public companies to ensure that they have implemented appropriate information security controls for their financial information.²⁵ Similarly, Securities and Exchange Commission (SEC) regulations impose various requirements for internal controls over information systems.
- ***Transaction records.*** Both the federal and state electronic transaction statutes—Electronic Signatures in Global and National Commerce Act (E-Sign) and the Uniform Electronic Transactions Act (UETA), now enacted in 47 states, the District of Columbia, and the U.S. Virgin Islands—require security for storage of electronic records relating to online transactions.

24. See Chapter 6 of this Handbook.

25. See generally Bruce H. Nearon et al., *Life after Sarbanes-Oxley: The Merger of Information Security and Accountability*, 45 JURIMETRICS: J.L., SCI. & TECH. 379–412 (Summer 2005).

72 THE ABA CYBERSECURITY HANDBOOK

- **Tax records.** Internal Revenue Service (IRS) regulations require businesses to implement information security to protect electronic tax records and as a condition of engaging in certain electronic transactions.
- **E-mail.** SEC regulations address security in a variety of contexts, and Food and Drug Administration (FDA) regulations require security for certain records.

Most laws do not differentiate based on the format of the data involved. Data kept in databases, e-mails, text documents, spreadsheets, voicemail messages, pictures, video, sound recordings, and other formats is typically treated the same.

In some cases, however, statutes and regulations governing data security differ based upon the media on which the data resides. Many laws focus only on electronic forms of data. Some, however, also address paper-based information (e.g., including those regulating proper data destruction). Some rules also can become very media-specific. For example, under some regulations, data kept on “removable media” is subject to additional encryption requirements that do not apply to data stored on other forms of electronic media.

E. What Level of Security Is Required?

Defining the scope of a lawyer’s security obligations begins with understanding that the law views security as a relative concept. Thus, as noted above, the basic standard for compliance is typically that security must be “reasonable”²⁶ or “appropriate.”²⁷

26. See, e.g., HIPAA, 42 U.S.C. § 1320d-2, and HIPAA Security Regulations, 45 C.F.R. § 164.306; COPPA, 15 U.S.C. § 6502(b)(1)(D), and COPPA regulations, 16 C.F.R. § 312.8; I.R.S. Rev. Proc. 97-22, sec. 4.01(2); SEC regulations, 17 C.F.R. § 257. See also UCC art. 4A, § 202 (“commercially reasonable” security procedure). Although HIPAA requires “reasonable and appropriate” security, 42 U.S.C. 1320d-2(d)(2) (emphasis added), some state personal information security laws require only that security procedures and practices be “reasonable”—e.g., CAL. CIV. CODE § 1798.81.5(b). See also Appendix B.1 (State Laws Imposing Obligations to Provide Security for Personal Information) of this Handbook.

27. HIPAA requires “reasonable and appropriate” security, 42 U.S.C. 1320d-2(d)(2). GLB requires covered financial institutions to “implement a comprehensive written information security program that includes administrative, technical, and physical safeguards *appropriate* to the size and complexity of the bank and the nature and scope of its activities.” 12 C.F.R. pt. 208, app. D-2, pt. II.A (Federal Reserve System) (emphasis added); see also GLB-implementing security regulations, *supra* note 7. The Massachusetts data security regulations require a comprehensive

In some cases, statutes and regulations define that standard in terms of positive results to be achieved, such as ensuring the *confidentiality*, *integrity*, and *availability* of systems and information.²⁸ In other cases, that standard is defined in terms of the harms to be avoided—for example, to protect systems and information against unauthorized access, use, disclosure, and so on. In some cases, the standard is not defined.

Regardless of approach, meeting this standard and achieving these objectives involves implementing appropriate physical, technical, and administrative security measures to protect both information systems and information from the various threats they face. Because those threats vary greatly from business to business, laws and regulations rarely specify or provide guidance about what specific security measures or technology a business should implement,²⁹ but instead require establishing and maintaining internal security “procedures,” “controls,” “safeguards,” or “measures”³⁰ designed to achieve the objectives identified above.

F. The Legal Requirements for “Reasonable Security”

Although security is relative, a legal standard for “reasonable” security is emerging. That standard rejects requirements for specific security measures (such as firewalls, passwords, or the like) and instead adopts a fact-specific approach to business security obligations that requires a “process” to assess risks, identify and implement appropriate security measures responsive to those risks, verify that the measures are effectively implemented, and ensure that they are continually updated in response to new developments.

written security program that contains safeguards that are “appropriate” to the size of the business, the resources available, the amount of stored data, and the need for security. Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.03(1).

28. See, e.g., FISMA and HIPAA Security Regulations, *supra* note 2. See also GLB Security Regulations (OCC), 12 C.F.R. pt. 30, app. B, pt. II.B; Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.00; N.Y. Dep’t of Fin. Servs., Cybersecurity Requirements for Financial Services Companies, N.Y. COMP. CODES R. & REGS. tit. 23, § 500.02.

29. Laws and regulations, however, do often focus on categories of security measures to address. See, e.g., HIPAA Security Regulations, 45 C.F.R. pt. 164. See also Appendices E (Best Practice Guidelines Issued by Federal Government Agencies) and F (Best Practices Guidelines Issued by State Government Agencies) of this Handbook.

30. See, e.g., FDA regulations at 21 C.F.R. pt. 11 (procedures and controls); SEC regulations at 17 C.F.R. 257.1(e)(3) (procedures); SEC regulations at 17 C.F.R. 240.17a-4 (controls); GLB regulations (FTC) 16 C.F.R. pt. 314 (safeguards).

74 THE ABA CYBERSECURITY HANDBOOK

This “process-oriented” legal standard for information security has been widely adopted:

- It was first outlined in a series of financial industry security regulations required under GLB titled Interagency Guidelines Establishing Standards for Safeguarding Consumer Information. They were issued by the Federal Reserve, the Office of the Comptroller of the Currency (OCC), the Federal Deposit Insurance Corporation (FDIC), and the Office of Thrift Supervision on February 1, 2001,³¹ and later were adopted by the FTC in its Safeguards Rule on May 23, 2002.³²
- The same approach was incorporated in the Federal Information Security Management Act of 2002³³ (FISMA) and in the HIPAA Security Standards issued by the Department of Health and Human Services on February 20, 2003.³⁴
- The FTC has since adopted the view that the “process-oriented” approach to information security outlined in these regulations is a “best practice” for legal compliance that should apply to all businesses in all industries. The FTC has, in effect, implemented this “process-oriented” approach in all of its decisions and consent decrees relating to alleged failures to provide appropriate information security.³⁵
- The National Association of Insurance Commissioners (NAIC) has recommended the same approach, and several state insurance regulators have adopted it.³⁶

31. 12 C.F.R. pt. 30, app. B (OCC), 12 C.F.R. pt. 208, app. D-2 and 12 C.F.R. pt. 25, app. F (Federal Reserve System), 12 C.F.R. pt. 364, app. B (FDIC), 12 C.F.R. pt. 568 and 570, app. B (Office of Thrift Supervision, which merged with the OCC as of July 21, 2011)).

32. FTC, Standards for Safeguarding Customer Information, 67 Fed. Reg. 36,484 (May 23, 2002) (FTC Safeguards Rule); 16 C.F.R. pt. 314.

33. 44 U.S.C. § 3544(b).

34. 45 C.F.R. pt. 164.

35. *See, e.g.*, FTC, Data Security, <https://www.ftc.gov/datasecurity> (listing FTC data security cases and corresponding decisions and consent decrees implementing this approach).

36. *See, e.g.*, NAT'L ASS'N OF INS. COMM'RS, ST-673-1, STANDARDS FOR SAFEGUARDING CUSTOMER INFORMATION MODEL REGULATION (Apr. 2002), <http://www.naic.org/store/free/MDL-673.pdf>. *See* other NAIC cybersecurity resources at http://www.naic.org/cipr_topics/topic_cyber_risk.htm.

- The Illinois Attorney General³⁷ endorsed this approach in 2012; the California Attorney General³⁸ did likewise in 2016.
- Some courts are taking the same approach.³⁹
- The Massachusetts Office of Consumer Affairs and Business Regulation adopted the approach in 2008 when it released its “Standards for Protection of Personal Information of Residents of the Commonwealth”⁴⁰ (Massachusetts Regulations), as required by the 2007 Massachusetts security breach and data destruction law.⁴¹ By specifically requiring businesses to implement a risk-based, process-oriented, comprehensive written information security program in accordance with a detailed list of requirements, the Massachusetts Regulations created one of the most comprehensive sets of general data security obligations imposed on businesses by a state.
- The 2017 cybersecurity regulations released by the New York State Department of Financial Services also adopted a similar approach.⁴²

The trend in the law is to recognize what security consultants have been saying for some time: “security is a process, not a product.”⁴³ Legal compliance with security obligations involves applying a “process” to the facts of each case to achieve an objective (i.e., to identify and implement the security measures appropriate for that situation), rather than implementing specific security measures in all cases. Thus, law firms cannot simply implement a

37. ILL. ATT’Y GEN., INFORMATION SECURITY AND SECURITY BREACH NOTIFICATION GUIDE 5 (Jan. 2012), http://www.illinoisattorneygeneral.gov/consumers/Security_Breach_Notification_Guidance.pdf.

38. CAL. ATT’Y GEN., CALIFORNIA DATA BREACH REPORT 2016, at 29 (Feb. 2016), <https://oag.ca.gov/breachreport2016>.

39. See, e.g., *Guin v. Brazos Higher Educ. Serv.*, 2006 U.S. Dist. LEXIS 4846 (D. Minn. Feb. 7, 2006).

40. Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.00 *et seq.*

41. MASS. GEN. LAWS ch. 93H, § 2(a).

42. N.Y. Dep’t of Fin. Servs., Cybersecurity Requirements for Financial Services Companies, N.Y. COMP. CODES R. & REGS. tit. 23, § 500.02.

43. BRUCE SCHNEIER, SECRETS & LIES: DIGITAL SECURITY IN A NETWORKED WORLD, at xii (2000). See also Appendices E (Best Practice Guidelines Issued by Federal Government Agencies) and F (Best Practices Guidelines Issued by State Government Agencies) of this Handbook.

76 THE ABA CYBERSECURITY HANDBOOK

standard set of security controls but rather must look more closely at their own individual situation.

This process-oriented approach to security compliance generally requires all businesses (including law firms) to take several steps, as outlined below.

1. Identify Information Assets

To protect something, you must know what it is, where it is, how it is used, how valuable it is, and so forth. Thus, when addressing information security, the first step is to identify the information assets to be protected and define the scope of the effort. This involves taking an inventory of the data that the business creates, collects, receives, uses, processes, stores, and communicates to others. It also requires examining the systems, networks, and processes by which such data is created, collected, received, used, processed, stored, and communicated.

Sensitive data files are often found in a variety of places within the firm. Data also is often in the possession and control of a third party, such as an outsource service provider or cloud provider. Yet the business (or law firm) is still responsible for the security of its (or its clients') data in the possession of third parties.

Identifying information assets also will help to determine the data security laws and regulations applicable to specific assets that must be addressed. This includes, for example, protected health information regulated under HIPAA, personally identifiable financial information regulated under GLB, information about children regulated under the Children's Online Privacy Protection Act (COPPA), and other types of personal information regulated under state security laws, the Fair Credit Reporting Act, or section 5 of the FTC Act.

Many security laws, regulations, and guidance documents expressly require identification of information assets. Examples include the following:

- An FTC business guidance document states what should be obvious but is often overlooked: "Effective data security starts with assessing what information you have and identifying who has access to it. Understanding how personal information moves into, through, and out of your business and who has—or could have—access to it is essential to

assessing security vulnerabilities. You can determine the best ways to secure the information only after you've traced how it flows.”⁴⁴

- A California attorney general guidance document states that organizations should “[i]dentify information assets and data to be secured.”⁴⁵
- Identification of information assets is a key component of the Cybersecurity Framework of the National Institute of Standards and Technology (NIST) and is included in the “Identify” function and “Asset Management” category of the Framework Core.⁴⁶

2. Conduct Periodic Risk Assessments

Just as you cannot implement security until you identify what you have that needs to be protected, you also cannot implement security until you know what risks you need to protect against. Thus, implementing reasonable security to protect the information assets of a business requires a thorough assessment of the potential risks to the entity's information systems and data.

A *risk assessment* is the process of identifying vulnerabilities and threats to the information assets used by the business or firm and assessing the potential impact and harm that would result if a threat materializes. This forms the basis for determining what countermeasures (i.e., security controls), if any, should be implemented to reduce risk to an acceptable level. Thus, a risk assessment requires:

- Conducting a threat assessment to identify all reasonably foreseeable internal and external *threats* to the information and system assets to be protected;⁴⁷

44. FTC, PROTECTING PERSONAL INFORMATION: A GUIDE FOR BUSINESS 2 (Oct. 2016), <https://www.ftc.gov/tips-advice/business-center/guidance/protecting-personal-information-guide-business>.

45. See CAL. ATT'Y GEN., CALIFORNIA DATA BREACH REPORT 2016, at 29 (Feb. 2016), <https://oag.ca.gov/breachreport2016>.

46. NIST FRAMEWORK FOR IMPROVING CRITICAL INFRASTRUCTURE CYBERSECURITY (Feb. 12, 2014) [hereinafter CYBERSECURITY FRAMEWORK], Framework Core at app. A, <https://www.nist.gov/cyberframework>. The Cybersecurity Framework is discussed in section II.H below.

47. See, e.g., GLB Security Regulations, 12 C.F.R. pt. 30, app. B, pt. III.B.1.

78 THE ABA CYBERSECURITY HANDBOOK

- Conducting a *vulnerability* assessment to identify the organization's vulnerabilities;
- Assessing the *likelihood* that each of the threats will materialize and, if so, the probability that one or more of the vulnerabilities will be exploited to cause harm—that is, identifying the likelihood that threat sources with the potential to exploit weaknesses or vulnerabilities in the system will actually do so;
- Evaluating the potential *damage* that will result; and
- Assessing the sufficiency of the security controls in place to guard against the threat.⁴⁸

A *threat* is anything that has the potential to cause harm. It can be an act of nature (such as a fire, flood, or tornado) or man-made, such as a computer virus, a hacker's actions, or an employee's negligent mistake. Threats should be considered in each area of relevant operation, including information systems; network and software design; information processing, storage, and disposal; prevention, detection, and response to attacks, intrusions, and system failures; and employee training and management.

Assessing risks also requires consideration of vulnerabilities. A *vulnerability* is a flaw or weakness that can be accidentally triggered or intentionally exploited by the threat to endanger or cause harm to an information asset. A vulnerability might be a hole in the roof, a system with easy-to-guess passwords, unencrypted data on a laptop computer, disgruntled employees, or employees who simply do not understand what steps they need to take to protect the security of the firm's data.

The likelihood that a threat will exploit a vulnerability to cause harm creates a *risk*. In other words, *risk* is the likelihood that something bad will happen that causes harm to an information asset. Somewhat more precisely, "[r]isk is a measure of the extent to which an entity is threatened by a potential circumstance or event, and is typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the

48. See, e.g., FISMA, 44 U.S.C. § 3544(a)(2)(A) and 3544(b)(1); GLB Security Regulations, 12 C.F.R. pt. 30, app. B, pt. III.B.2.

likelihood of occurrence.”⁴⁹ Risk is present wherever a threat intersects with a vulnerability. For example, if the threat is rain, and the vulnerability is a hole in the roof, risk is the likelihood that it will rain, causing water to enter the building through the hole in the roof and doing damage to the building and/or its contents. Similarly, if the threat is a hacker, and the vulnerability is open Internet access to a server containing sensitive data, risk is the likelihood that a hacker will enter the system and view, copy, alter, or destroy the sensitive data.

This process will be the baseline against which security controls can be selected, implemented, measured, and validated. The goal is to understand the risks that the firm faces, to determine which risks are acceptable, and to identify appropriate and cost-effective safeguards to combat the risks that are unacceptable. Thus, such risks should be evaluated in light of the nature of the business or law firm and its clients, its transactional capabilities, the sensitivity and value of the stored information to the business and its trading partners, and the size and volume of its transactions.⁵⁰

Numerous security laws and regulations expressly require a risk assessment as part of a comprehensive security program. Laws and regulations that do not expressly include such a requirement often do so implicitly.

- Various federal security statutes and regulations, including GLB,⁵¹ HIPAA,⁵² and FISMA,⁵³ expressly require a risk assessment.
- The consent decrees entered in FTC enforcement actions have expressly required “the identification of material internal and external risks to the security, confidentiality, and integrity of covered information that could result in the unauthorized disclosure, misuse, loss, alteration, destruction, or other compromise of such information.”⁵⁴

49. NIST SPEC. PUBL’N 800-30, REV. 1, GUIDE FOR CONDUCTING RISK ASSESSMENTS 8 (Sept. 2012).

50. *See, e.g.*, Fed. Fin. Insts. Examination Council, Authentication in an Electronic Banking Environment 3 (July 30, 2001), https://www.ffiec.gov/pdf/authentication_guidance.pdf.

51. 16 C.F.R. § 314.4(b).

52. 45 C.F.R. § 164.308(a)(1)(ii)(A).

53. 44 U.S.C. 3554(b)(1).

54. *See* FTC, Data Security, <https://www.ftc.gov/datasecurity> (listing FTC cases and enforcement actions alleging failure to provide reasonable security).

80 THE ABA CYBERSECURITY HANDBOOK

- State security laws (e.g., in Oregon⁵⁵) and regulations in Massachusetts⁵⁶ and New York⁵⁷ expressly require a risk assessment.
- Risk assessment is a key component of the NIST Cybersecurity Framework and is included in the “Identify” function and “Risk Assessment” category of the Framework Core.⁵⁸
- The California attorney general’s office released a report stating that “[i]nformation security laws and regulations generally require a risk management approach. In essence, this means organizations must develop, implement, monitor, and regularly update a comprehensive information security program [under which organizations must] assess risks to the assets and data.”⁵⁹
- Similarly, guidance issued by the Illinois attorney general recommends that businesses and government agencies should “[i]dentify reasonably foreseeable internal and external risks to the security, confidentiality, and integrity of customer information that could result in the unauthorized disclosure, misuse, alteration, destruction, or other compromise of such information, and assess the sufficiency of any safeguards in place to control these risks.”⁶⁰
- In addition, several U.S. courts have held that a risk assessment plays a key role in determining whether a duty will be imposed and liability found. Where injury is foreseeable and preventable, a business has a duty to provide appropriate security to address the potential harm.⁶¹ On the other hand, where a proper risk assessment was done, but a

55. OR. REV. STAT. § 646A.622(2)(d)(A).

56. Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.03(2)(b).

57. N.Y. Dep’t of Fin. Servs., Cybersecurity Requirements for Financial Services Companies, N.Y. COMP. CODES R. & REGS. tit. 23, § 500.02.

58. See CYBERSECURITY FRAMEWORK, *supra* note 46, § 1.2 and Framework Core at app. A.

59. CAL. ATT’Y GEN., CALIFORNIA DATA BREACH REPORT 2016, at 29 (Feb. 2016) <https://oag.ca.gov/breachreport2016>.

60. ILL. ATT’Y GEN., INFORMATION SECURITY AND SECURITY BREACH NOTIFICATION GUIDE 5 (Jan. 2012), http://www.illinoisattorneygeneral.gov/consumers/Security_Breach_Notification_Guidance.pdf.

61. See, e.g., *Wolfe v. MBNA Am. Bank*, 485 F. Supp. 2d 874, 882 (W.D. Tenn. 2007); *Bell v. Mich. Council*, 2005 Mich. App. LEXIS 353 (Mich. App. Feb. 15, 2005).

particular harm was not reasonably foreseeable, the defendant would not be liable for failure to defend against it.⁶²

The following publications provide general information and guidance on conducting a risk assessment:

- NIST Special Publication 800-30, Rev. 1, *Guide for Conducting Risk Assessments*⁶³
- Massachusetts's *A Small Business Guide: Formulating A Comprehensive Written Information Security Program*⁶⁴
- *Interagency Guidelines Establishing Information Security Standards: Small-Entity Compliance Guide*⁶⁵
- Federal Financial Institutions Examination Council (FFIEC) *IT Examination Handbook and Information Security Booklet*⁶⁶

3. Develop and Implement an Appropriate Security Program

Based on the results of the risk assessment, the law requires a business to design and implement a security program consisting of reasonable physical, technical, and administrative security measures to manage and control the risks identified during the risk assessment.⁶⁷ The security program should be

62. See *Guin v. Brazos Higher Educ. Serv.*, 2006 U.S. Dist. LEXIS 4846, at *13 (D. Minn. Feb. 7, 2006) (finding that where a proper risk assessment was done, the inability to foresee and deter a specific burglary of a laptop was not a breach of a duty of reasonable care).

63. See NIST SPEC. PUBL'N NO. 800-30, REV. 1, GUIDE FOR CONDUCTING RISK ASSESSMENTS (Sept. 2012), <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf>.

64. See MASS. OFFICE OF CONSUMER AFFAIRS, A SMALL BUSINESS GUIDE: FORMULATING A COMPREHENSIVE WRITTEN INFORMATION SECURITY PROGRAM, <http://www.mass.gov/ocabr/docs/idtheft/sec-plan-smallbiz-guide.pdf>. See also ILL. ATT'Y GEN., INFORMATION SECURITY AND SECURITY BREACH NOTIFICATION GUIDANCE (2012), [HTTP://ILLINOISATTORNEYGENERAL.GOV/CONSUMERS/SECURITY_BREACH_NOTIFICATION_GUIDEANCE.PDF](http://illinoisattorneygeneral.gov/CONSUMERS/SECURITY_BREACH_NOTIFICATION_GUIDEANCE.PDF).

65. FED. RESERVE BD. ET AL., INTERAGENCY GUIDELINES ESTABLISHING INFORMATION SECURITY STANDARDS: SMALL-ENTITY COMPLIANCE GUIDE (Dec. 14, 2005), <https://www.federalreserve.gov/boarddocs/press/bcreg/2005/20051214/attachment.pdf>.

66. FFIEC, IT Examination Handbook InfoBase, IT Booklets, <http://ithandbook.ffiec.gov/it-booklets.aspx> (links to booklets).

67. See, e.g., FTC, Data Security, <https://www.ftc.gov/datasecurity> (listing FTC data security decisions and consent decrees imposing such requirements); GLB Security Regulations (OCC), 12 C.F.R. pt. 30, app. B, pt. II.A; HIPAA Security Regulations, 45 C.F.R. § 164.308(a)(1)(i); FISMA, 44 U.S.C. § 3544(b).

82 THE ABA CYBERSECURITY HANDBOOK

designed to provide reasonable safeguards to control the identified risks⁶⁸—that is, to reduce them to a reasonable and appropriate level.

The presence or absence of specific security measures says little about the status of a business's legal compliance with its information security obligations. The security measures implemented by a business must respond to the particular threats it faces and address its specific vulnerabilities. Posting armed guards around a building sounds impressive as a security measure, but it is of little value if the primary threat the business faces is unauthorized remote access to its data via the Internet. Likewise, firewalls and intrusion detection software are often effective ways to stop hackers and protect sensitive databases, but if a business's major vulnerability is careless (or malicious) employees who inadvertently (or intentionally) disclose passwords or protected information, then even those sophisticated and important technical security measures will not adequately address the problem.

a. Relevant Factors to Consider

Virtually all of the existing precedent recognizes that there is no “one size fits all” approach when determining what security measures to implement within a particular business. Such a determination will depend upon a variety of factors.

Traditional negligence law suggests that the relevant factors are (1) the probability of the identified harm occurring (i.e., the likelihood that a foreseeable threat will materialize), (2) the gravity of the resulting injury if the threat does materialize, and (3) the burden of implementing adequate precautions.⁶⁹ In other words, the standard of care to be exercised in any particular case depends upon the circumstances of that case and the extent of foreseeable danger.⁷⁰

68. See, e.g., FTC, Data Security, <https://www.ftc.gov/datasecurity> (listing FTC data security decisions and consent decrees imposing such requirements); GLB Security Regulations, 12 C.F.R. pt. 30, app. B, pt. II.B.

69. See, e.g., *United States v. Carroll Towing*, 159 F.2d 169, 173 (2d Cir. 1947).

70. See, e.g., *DCR Inc. v. Peak Alarm Co.*, 663 P.2d 433, 435 (Utah 1983); see also *Glatt v. Feist*, 156 N.W.2d 819, 829 (N.D. 1968) (the amount or degree of diligence necessary to constitute ordinary care varies with the facts and circumstances of each case).

Security regulations take a similar approach and indicate that the following factors are relevant in determining what security measures should be implemented:

- The probability and criticality of potential risks;
- The size, complexity, and capabilities of the business;
- The nature and scope of the business activities;
- The nature and sensitivity of the information to be protected;
- The organization's technical infrastructure, hardware, and software security capabilities;
- The state of the art of technology and security; and
- The cost of the security measures (cost was the factor mentioned most often, which suggests that businesses are not required to do everything theoretically possible).

b. Categories of Security Measures to Consider

Most laws do not require businesses to implement specific security measures or use a particular technology, but instead provide flexibility to use measures reasonably designed to achieve the objectives specified in the regulations.⁷¹ This focus on flexibility means that, like the obligation to use “reasonable care” under tort law, determining compliance may ultimately become more difficult, as there are unlikely to be any safe harbors for security.

Nonetheless, statutes and regulations⁷² consistently focus on physical, technical, and administrative security measures and, within those areas, often mention certain *categories* of security measures that businesses should consider (although how a business must address the categories is typically not specified). Those categories of security measures include the following:

- ***Physical facility and device security controls.*** Measures to safeguard the facility; measures to protect against destruction, loss, or damage of information due to potential environmental hazards (such as fire and

71. See, e.g., HIPAA Security Regulations, 45 C.F.R. § 164.306(b)(1).

72. See, e.g., Appendix C.2 (Federal Regulations Imposing Authentication Requirements) of this Handbook.

84 THE ABA CYBERSECURITY HANDBOOK

water damage or technological failures); procedures that govern the receipt and removal of hardware and electronic media into and out of a facility; and procedures that govern the use and security of physical workstations;

- ***Physical access controls.*** Access restrictions at buildings, computer facilities, and records storage facilities to permit access only to authorized individuals;
- ***Technical access controls.*** Software, policies, and procedures to ensure that authorized persons who need access to the system have appropriate access and that those who should not have access are prevented from getting it, including procedures to determine access authorization, grant and control access, verify that a person or entity seeking access is the one claimed (i.e., authentication), and terminate access;
- ***Intrusion detection procedures.*** Software, policies, and procedures to monitor log-in attempts and report discrepancies; system monitoring and intrusion detection systems and procedures to detect actual and attempted attacks on, or intrusions into, the organization's information systems; and procedures for preventing, detecting, and reporting malicious software (e.g., virus software);
- ***Employee procedures.*** Job control procedures, segregation of duties, and background checks for employees with responsibility for or access to protected information, and controls to prevent employees from providing information to unauthorized individuals who may seek to obtain this information through fraudulent means;
- ***System modification procedures.*** Procedures designed to ensure that system modifications are consistent with the business's security program;
- ***Data integrity, confidentiality, and storage.*** Procedures to protect information from unauthorized access, alteration, disclosure, or destruction during storage or transmission, including storage of data in a format that cannot be meaningfully interpreted if accessed (e.g., encrypted), or in a location that is inaccessible to unauthorized persons and/or protected by a firewall;
- ***Data destruction and hardware and media disposal.*** Procedures regarding final disposition of information and/or hardware on which it resides,

and procedures for removal of data from media before reuse of the media;

- **Audit controls.** Maintenance of records to document repairs and modifications to the physical components of the facility related to security (walls, doors, locks, etc.), and hardware, software, and/or procedural audit control mechanisms that record and examine activity in the systems;
- **Contingency plan.** Procedures designed to ensure the ability to continue operations in an emergency, such as a data backup plan, disaster recovery plan, and emergency mode operation plan;
- **Incident response plan.** A plan for taking responsive steps if the business suspects or detects that a security breach has occurred; such steps include ensuring that appropriate persons within the organization are promptly notified of the breach, that prompt action is taken in responding to the breach (e.g., stopping further information compromise and working with law enforcement), and that persons who may be injured by the breach are appropriately notified.

4. *Provide Training and Education*⁷³

Training and education for employees is a critical component of any security program. Even the very best physical, technical, and administrative security measures are of little value if employees do not understand their roles and responsibilities regarding security. For example, installing heavy-duty doors with state-of-the-art locks (whether physical or virtual) will not provide the intended protection if the employees authorized to have access leave the doors open or unlocked for unauthorized persons to pass through.

Security education begins with communicating applicable security policies, procedures, standards, and guidelines to employees. It also includes implementing a security awareness program, providing periodic security reminders, and developing and maintaining relevant employee training, such as user education on virus protection, password management, and how to

73. Training and education are discussed in more detail in Chapter 13 of this Handbook.

86 THE ABA CYBERSECURITY HANDBOOK

report discrepancies. It is also important to impose appropriate sanctions against employees who fail to comply with security policies and procedures.

5. Monitor and Test the Security Controls

Merely implementing security measures is not sufficient. A business also must ensure that the security measures have been properly implemented and are effective. This includes conducting an assessment of the sufficiency of the security measures in place to control the identified risks and conducting regular testing or monitoring of the effectiveness of those measures. Existing precedent also suggests that a business must monitor compliance with its security program. To that end, a regular review of records of system activity, such as audit logs, access reports, and security incident tracking reports, is also important.

6. Review and Adjust the Security Program

The legal standard for information security recognizes that security is a moving target. Law firms and other businesses must continually keep up with ever-changing threats, risks, and vulnerabilities as well as the security measures available to respond to them. This requires conducting periodic internal reviews to evaluate and adjust the information security program in light of:

- The results of the testing and monitoring;
- Any material changes to the business or client arrangements;
- Any changes in technology;
- Any changes in internal or external threats;
- Any environmental or operational changes; and
- Any other circumstances that may have a material impact.⁷⁴

In addition to conducting periodic internal reviews, it also may be appropriate to obtain a periodic review and assessment (audit) by qualified

⁷⁴ See, e.g., GLB Security Regulations, 12 C.F.R. pt. 30, app. B, pt. II.E; HIPAA Security Regulations, 45 C.F.R. § 164.308(a)(8).

independent third-party professionals. Such professionals would use procedures and standards generally accepted in the profession to certify that the security program meets or exceeds applicable requirements and that the program is operating with sufficient effectiveness to provide reasonable assurances that the security, confidentiality, integrity, and availability of information are protected.

The business should then adjust the security program in light of the findings or recommendations that come from such reviews.

7. Oversee Third-Party Service Provider Arrangements

In today's business environment, companies and law firms often rely on third parties, such as outsource providers and cloud providers, to handle much of their data. When firm or client data is in the possession or under the control of a third party, this presents special security challenges. Thus, it is important to address the security of the firm's data in the possession of such third parties.

To that end, laws and regulations imposing information security obligations on businesses often expressly address requirements regarding the use of third-party outsource providers. Such rules and regulations make clear that regardless of who performs the work, the legal obligation to provide the security itself remains with the business. As it is often said, "you can outsource the work, but not the responsibility." Thus, third-party relationships should be subject to the same risk management, security, privacy, and other protection policies that would be expected if a company or firm were conducting the activities directly.⁷⁵

Generally, the legal standard for security imposes three basic requirements on businesses that outsource: (1) they must exercise due diligence in selecting service providers, (2) they must contractually require outsource providers to implement appropriate security measures, and (3) they must monitor the performance of the outsource providers.⁷⁶

75. See, e.g., Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.02(2)(f).

76. *Id.*

88 THE ABA CYBERSECURITY HANDBOOK

G. Rules Governing Specific Data Elements and Controls

1. *Special Rules for Specific Data Elements or Activities*

In addition to imposing a general obligation to provide data security, some laws and regulations require protection of specific data elements, such as Social Security numbers, credit card transaction data, and other sensitive data.

The various state security breach notification laws (discussed in section III below) have created a de facto category of sensitive information in the United States. These laws require special action (i.e., disclosure) upon a breach of security for a subcategory of personal data generally considered to be sensitive because it can facilitate identity theft.

The security of Social Security numbers has been the particular focus of numerous state laws enacted in recent years.⁷⁷ The scope of these laws ranges from restrictions on the manner in which Social Security numbers can be used to requirements for security when communicating and/or storing such numbers. For example, several states have enacted laws that prohibit requiring an individual to transmit his or her Social Security number over the Internet unless the connection is secure or the number is encrypted.⁷⁸

For businesses that accept credit card transactions, the PCI Standard⁷⁹ imposes significant security obligations for credit card data captured as part of any credit card transaction. The PCI Standard, jointly created by the major credit card associations, requires businesses that accept MasterCard, Visa, American Express, Discover, and Diners Club cards to comply. State law obligations also may apply.⁸⁰

77. See, e.g., Appendix B.5 (State SSN Laws), and Appendix B.6 (State Laws Requiring SSN Policies) of this Handbook.

78. See U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-05-1016T, SOCIAL SECURITY NUMBERS: FEDERAL AND STATE LAWS RESTRICT USE OF SSNs, YET GAPS REMAIN, at app. III (Sept. 15, 2005), <http://www.gao.gov/assets/120/112174.pdf> (list of state laws). Many federal agencies are making efforts to reduce collection, use, and display of SSNs, but have had mixed success given a number of factors (including statutes and regulations that mandate collection of SSNs); see U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-17-655T, SOCIAL SECURITY NUMBERS: OMB AND FEDERAL EFFORTS TO REDUCE COLLECTION, USE, AND DISPLAY (May 23, 2017), <https://www.gao.gov/products/GAO-17-655T>.

79. See PCI Sec. Standards Council, Document Library, https://www.pcisecuritystandards.org/document_library.

80. See, e.g., Appendix B.2 (State Laws Imposing Obligations to Provide Security for Credit Card Information) of this Handbook.

2. *Duty to Encrypt Data*

Some laws and regulations impose obligations to use encryption in certain situations. Initially this included state laws that mandate encryption of Social Security numbers for communication over the Internet.⁸¹ More recently, however, some state laws prohibit the electronic transmission of any personal information to a person outside of the secure system of the business unless the information is encrypted. Most notable are the Massachusetts Regulations, which require businesses to encrypt personal information if it is stored on “laptops or other portable devices,” “will travel across public networks,” or will “be transmitted wirelessly.”⁸²

3. *Duty to Destroy Data Properly*

Several laws and regulations impose security requirements regarding the way that data is destroyed.⁸³ Such statutes and regulations generally require businesses to properly dispose of personal information by taking reasonable measures to protect against unauthorized access to, or use of, the information in connection with its disposal. For information in paper form, this typically requires implementing and monitoring compliance with policies and procedures that require the burning, pulverizing, or shredding of papers containing personal information so that it cannot be read or reconstructed. For information in electronic form, such regulations typically require implementing and monitoring compliance with policies and procedures that require the destruction or erasure of electronic media containing consumer personal information so that it cannot be read or reconstructed.

H. Frameworks for Reasonable Security

The Cybersecurity Framework is one of the deliverables contemplated by President Barack Obama’s Executive Order 13,636, Improving Critical

81. See, e.g., ARIZ. REV. STAT. § 44-1373; CAL. CIV. CODE § 1798.85; CONN. GEN. STAT. § 42-470; MD. CODE ANN., COM. LAW § 14-3402(4). See also Appendix B.5 (State SSN Laws) of this Handbook; many state SSN laws mandate use of encryption when transmitting Social Security numbers.

82. Mass. Standards for the Protection of Personal Info., 201 MASS. CODE REGS. 17.04(3) and (5).

83. See, e.g., Appendix B.3 (State Data Disposal/Destruction Laws) and Appendix C.3 (Federal Data Disposal/Destruction Regulations) of this Handbook.

90 THE ABA CYBERSECURITY HANDBOOK

Infrastructure Cybersecurity, which was issued on February 12, 2013.⁸⁴ Recognizing that the national and economic security of the United States depends on the reliable functioning of the nation's critical infrastructure, the executive order directed NIST to work with the private sector to develop a voluntary framework—based on existing standards, guidelines, and practices—for reducing cybersecurity risks to critical infrastructure.

Consistent with the requirements of the executive order, the Cybersecurity Framework was created through collaboration between industry and government⁸⁵ and “provides a consensus description of what’s needed for a comprehensive cybersecurity program.” “It reflects the efforts of a broad range of industries that see the value of and need for improving cybersecurity and lowering risk.”⁸⁶ According to NIST, the Cybersecurity Framework “allows organizations—regardless of size, degree of cyber risk or cybersecurity sophistication—to apply the principles and best practices of risk management to improve the security and resilience of critical infrastructure.”⁸⁷

The Cybersecurity Framework references several generally accepted domestic and international security standards, and the participants generally agree that it constitutes a best practice for cybersecurity.⁸⁸ It might be argued that the Cybersecurity Framework is little more than a compilation of established industry security practices, but even so it collates such practices into a framework of activities that arguably establishes a set of requirements for the development of “reasonable” security practices. Moreover, it carries

84. Exec. Order No. 13,636, *supra* note 12.

85. The “framework is the culmination of a year-long effort that brought together thousands of individuals and organizations from industry, academia and government.” Press Release, NIST, NIST Releases Cybersecurity Framework Version 1.0 (Feb. 12, 2014), <https://www.nist.gov/itl/csd/launch-cybersecurity-framework-021214.cfm>.

86. *Id.* (quoting Under Secretary of Commerce for Standards and Technology and NIST Director Patrick D. Gallagher).

87. *Id.*

88. “Over the past year, individuals and organizations throughout the country and across the globe have provided their thoughts on the kinds of standards, best practices, and guidelines that would meaningfully improve critical infrastructure cybersecurity. The Department of Commerce’s National Institute of Standards and Technology (NIST) consolidated that input into the voluntary Cybersecurity Framework that we are releasing today.” Press Release, White House, Launch of the Cybersecurity Framework (Feb. 12, 2014), <https://www.whitehouse.gov/the-press-office/2014/02/12/launch-cybersecurity-framework>.

the weight of being a government-issued framework that was the result of a year-long collaboration between industry and government to develop a voluntary “how to” guide for organizations to enhance their cybersecurity.⁸⁹

Technically, the Cybersecurity Framework was written only for businesses in the 16 critical infrastructure sectors.⁹⁰ But the practical reality goes much further. The Cybersecurity Framework is written as a generally applicable document that is in no way unique to critical infrastructure industries. It is not industry-specific, nor is it country-specific. Consistent with existing law, the Cybersecurity Framework adopts a risk-based approach to managing cybersecurity risk. As such, it appears to fit quite well with the approach of existing legal requirements for cybersecurity obligations. It provides general approaches and activities to address cybersecurity for all businesses.

“The Framework is designed to complement existing business and cybersecurity operations. It can serve as the foundation for a new cybersecurity program or a mechanism for improving an existing program.”⁹¹ The drafters of the Cybersecurity Framework contemplated that “[o]rganizations can use the framework to determine their current level of cybersecurity, set goals for cybersecurity that are in sync with their business environment, and establish a plan for improving or maintaining their cybersecurity. It also offers a methodology to protect privacy and civil liberties to help organizations incorporate those protections into a comprehensive cybersecurity program.”⁹²

NIST noted that “an organization without an existing cybersecurity program can use the Framework as a reference to establish one”⁹³ or to improve an existing program.⁹⁴

89. *Id.*

90. According to Presidential Policy Directive 21 (PPD-21), the 16 critical infrastructure sectors are chemical, commercial facilities, communications, critical manufacturing, dams, defense industrial base, emergency services, energy, financial services, food and agriculture, government facilities, healthcare and public health, information technology, nuclear reactors, materials and waste, transportation systems, and water and wastewater systems.

91. CYBERSECURITY FRAMEWORK, *supra* note 46, at 13. *See generally id.* § 3.2, Establishing or Improving a Cybersecurity Program, at 13–15.

92. *See* Press Release, NIST, NIST Releases Cybersecurity Framework Version 1.0 (Feb. 12, 2014), <https://www.nist.gov/itl/csd/launch-cybersecurity-framework-021214.cfm>.

93. CYBERSECURITY FRAMEWORK, *supra* note 46, at 4.

94. *See id.* § 3.2.

92 THE ABA CYBERSECURITY HANDBOOK

Other cybersecurity frameworks are often used as a methodology to implement reasonable security in an organization. Two of the more commonly used are (1) the ISO/IEC 27001 framework, which was developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC),⁹⁵ and (2) the Control Objectives for Information and Related Technologies (COBIT) framework, which was created by ISACA (previously known as the Information Systems Audit and Control Association).⁹⁶

III. The Duty to Notify of Security Breaches

Legal requirements do not stop at obligations to *implement* security measures to protect data. Now there is a global trend to enact laws and regulations that impose an obligation to *disclose* security breaches to the persons affected.

A. What Is the Source of the Duty?

Today, almost all U.S. states have enacted security breach notification laws, generally based on a 2003 California law, and such obligations can also be triggered at the federal level.⁹⁷ The HIPAA regulations require breach notification,⁹⁸ as do the requirements of the federal banking regulatory agencies.⁹⁹ The IRS also has imposed a disclosure requirement with respect to taxpayers whose electronic tax records are the subject of a security breach.¹⁰⁰

95. ISO/IEC 27001, Information Technology—Security Techniques—Information Security Management Systems—Requirements (2013), available for purchase at <https://www.iso.org/isoiec-27001-information-security.html>.

96. See ISACA, <http://www.isaca.org/cobit/pages/default.aspx>.

97. See, e.g., Appendix B.4 (State Security Breach Notification Laws) and Appendix C.4 (Federal Security Breach Notification Regulations) of this Handbook.

98. 45 C.F.R. § 164.314(a)(2)(1)(C) & 45 C.F.R. § 164.410.

99. Interagency Guidance on Response Programs for Unauthorized Access to Customer Information and Customer Notice, 12 C.F.R. pt. 30, app., supp. A, pt. III (OCC), 12 C.F.R. pt. 208 (Federal Reserve System), and 12 C.F.R. pt. 364 (FDIC), and 12 C.F.R. pt. 568 (Office of Thrift Supervision, which merged with the OCC as of July 21, 2011), 70 Fed. Reg. No. 59, Mar. 29, 2005, at 15,736 [hereinafter Interagency Guidance].

100. See I.R.S. Rev. Proc. 98-25, § 8.01. See Appendix C.4 (Federal Security Breach Notification Regulations) of this Handbook.

These laws impose an obligation similar to the common law “duty to warn” of dangers, which is often based on the view that a party who has superior knowledge of a danger of injury or damage to another posed by a specific hazard must warn those who lack such knowledge. By requiring notice to persons who might be adversely affected (e.g., those whose compromised personal information may be used to facilitate identity theft), such laws seek to warn such persons that personal information has been compromised and provide an opportunity to take steps to self-protect against the consequences of identity theft.

Lawyers may have additional notification obligations under the rules of professional responsibility, other law applicable specifically to lawyers, or contractual obligations to clients.

B. What Is the Statutory Duty?

The statutory duty, as embodied in the state and federal security breach notification laws, generally requires that any business that possesses or controls certain sensitive personal information about a covered individual must disclose any breach of such information to the affected person.¹⁰¹ Several statutes also require notification to the state attorney general or other regulatory agency. In some cases, notification requirements also extend to informing credit reporting agencies and the press.

The key elements of the breach notification statutes can be summarized as follows:

Type of Information. The breach notification statutes generally apply to unencrypted sensitive personally identified information—for example, information consisting of first name or initial and last name, plus one of the following: Social Security number, driver’s license or state ID number, or financial account number or credit or debit card number (along with any PIN or other access code where required to access the account). In some states, this list is longer and may also include, for example, medical information, insurance policy numbers, passwords by themselves, biometric

101. Exception: Where the business maintains computerized personal information that the business does not own, the laws require the business to notify the owner or licensee of the information, rather than the individuals themselves, of any breach of the security of the system.

94 THE ABA CYBERSECURITY HANDBOOK

information, professional license or permit numbers, telecommunication access codes, mother's maiden name, employer ID number, electronic signatures, and descriptions of an individual's personal characteristics.¹⁰²

Triggering Event. The event that triggers the obligation to provide individuals with notice of a breach involving their personal information is typically referred to in the breach statutes as a "breach of the security of the system." This term is often defined as: "unauthorized acquisition of unencrypted computerized data that compromises the security, confidentiality or integrity of personal information maintained by the person or business."¹⁰³

The requirements of this definition, in combination with certain other exclusions available in many states (e.g., an exclusion for security breaches that the custodian of the exposed data determines will not likely cause harm),¹⁰⁴ allow for more than one approach to determining when factors are present that impose an obligation to notify under the breach notification statutes.

Who Must Be Notified. Notice must be given to, at a minimum, any residents of the state whose unencrypted personal information was the subject of the breach. In some cases, the state's attorney general (or other enforcement agency) and/or the media must also be notified.

When Notice Must Be Provided. Generally, persons must be notified in the most expedient time possible and without unreasonable delay (although some states specify a certain number of days). In most states, the time for notice may be extended:

102. See, e.g., ARK. CODE §§ 4-110-101 *et seq.*; LA. REV. STAT. §§ 51:3071 *et seq.*; MD. CODE ANN., COM. LAW §§ 14-3501 *et seq.*; NEB. REV. STAT. §§ 87-801 *et seq.*; N.J. STAT. 56:8-163; N.C. GEN. STAT. § 75-65; N.D. CENT. CODE §§ 51-30-01 *et seq.*; OR. REV. STAT. § 646A.600-.628. The Federal Banking Interagency Guidance, *see supra* note 99, also includes any combination of components of customer information that would allow someone to log onto or access the customer's account, such as user name and password, or account number and password.

103. See, e.g., CAL. CIV. CODE § 1798.82(d).

104. For example, Iowa's Breach Notification Statute stipulates that notification is not required if "after an appropriate investigation or after consultation with the relevant federal, state, or local agencies responsible for law enforcement, the person determined that no reasonable likelihood of financial harm to the consumers whose personal information has been acquired has resulted or will result from the breach. Such a determination must be documented in writing and the documentation must be maintained for five years." See IOWA CODE § 715C.2(6).

- For the legitimate needs of law enforcement, if notification would impede a criminal investigation.
- To take necessary measures to determine the scope of the breach and restore reasonable integrity to the system.

Form of Notice. Notice may be provided in writing (e.g., on paper and sent by mail), in electronic form (e.g., by e-mail, but only in compliance with E-SIGN¹⁰⁵), or by substitute notice. If the cost of providing individual notice is greater than a certain amount (e.g., \$250,000), or if more than a certain number of people would have to be notified (e.g., 500,000), the business may use substitute notice, consisting of:

- E-mail, when the e-mail address is available, and
- Conspicuous posting on the entity's website, and
- Publishing notice in all major statewide media.

Requirements vary from state to state, however, and some requirements have become controversial. One of the biggest issues concerns the nature of the triggering event. In California, for example, notification is required whenever there has been unauthorized access that compromises the security, confidentiality, or integrity of electronic personal data. In other states, unauthorized access does not trigger the notification requirement unless there is a reasonable likelihood of harm to the individuals whose personal information is involved or unless the breach is material.

C. When Does a Contract-Based Duty Arise?

It is increasingly common for contracts with business partners of all types to require the recipient or processor of a business's data to notify that party in the event of a breach. This trend is also being extended to law firms. Clients (particularly in regulated industries such as financial or healthcare) are requiring that their law firms provide prompt notice of any security breach. For example, breach reporting is a key requirement of the Model

105. 15 U.S.C. §§ 7001 *et seq.* This generally requires that entities comply with the requisite consumer consent provisions of E-SIGN at 15 U.S.C. § 7001(c).

96 THE ABA CYBERSECURITY HANDBOOK

Information Protection and Security Controls for Outside Counsel Processing Company Confidential Information, released in 2017 by the Association of Corporate Counsel.¹⁰⁶

IV. Practical Considerations: A Top Ten List

Lawyers have many legal obligations to provide data security. Below is a list of practical tips regarding compliance with those obligations:

1. Identify the data you have (yours, your clients', data obtained during due diligence or discovery) and understand where it is stored, how it can be accessed, and how it is used.
2. Evaluate the risks to the data you have.
3. Develop a written security program to protect that data against the identified risks.
4. If you use third parties (e.g., providers of cloud services or outsourcing services) to store or process the data, take appropriate steps to make sure that they adequately protect the security of the data you entrust to them.
5. On a regular basis, reevaluate the risks you face and the adequacy of your security program, and adjust the program as necessary.
6. Determine which data (yours, your clients', data obtained during due diligence or discovery) is subject to which laws and regulations (including special sector-specific regulations such as GLB or HIPAA), and be sure you handle it in accordance with any special requirements in those laws and regulations.
7. Recognize that other lawyers and staff within the firm can be a weak link, and provide appropriate training and awareness-raising reminders for all lawyers and staff.

106. ASS'N OF CORPORATE COUNSEL, MODEL INFORMATION PROTECTION AND SECURITY CONTROLS FOR OUTSIDE COUNSEL PROCESSING COMPANY CONFIDENTIAL INFORMATION (Jan. 2017), <http://www.acc.com/advocacy/upload/Model-Information-Protection-and-Security-Controls-for-Outside-Counsel-Jan2017.pdf>.

Lawyers' Legal Obligations to Provide Data Security **97**

8. Develop an incident response plan that covers the data you have.
9. Keep in mind that laws and regulations governing data security may apply to all of the data in your possession, independent of ethical obligations specifically applicable to attorneys.
10. Remember that security is a process and is never complete, so you must remain vigilant for new threats.

Chapter 13

Get SMART on Data Protection Training and How to Create a Culture of Awareness

Ruth Hill Bro and Jill D. Rhodes

I. Data Protection Training Basics and Core Principles

Any business that works with personal and sensitive data must develop a strategy for protecting that data. When assessing how to do so, organizations, including law firms, often mistakenly rely on technology as the solution. In fact, four factors are key to implementing a proper information security and data protection program in any setting:

- Establishing the appropriate *governance* for the data, such as policies and the oversight of an executive level committee tasked with reducing data protection risk;
- Ensuring that the *people* working with the data know how best to protect it;
- Assessing data protection and usage *processes*; and
- Employing appropriate *technology* to protect the network.

These four factors work together to develop an overarching and effective program.

272 THE ABA CYBERSECURITY HANDBOOK

This chapter focuses on the people aspect of that equation, but other factors (e.g., governance and processes) also come into play. Most data missteps in law firms and other businesses are directly linked to something an employee or contractor did, whether intentionally or unintentionally. The easiest way to address this risk is to educate employees and others about the risk and their role in protecting personal and sensitive data.

Education and training can be provided by many facets of the organization, whether human resources (HR), the chief privacy officer (CPO), the chief information security officer (CISO), or others. Regardless of which groups do the training, it is critical that they work together to produce a common vision and message that is then disseminated across the organization.

A. Why Train on Data Protection?

All organizations, including law firms, are increasingly recognizing that data underpins virtually everything that they do and—like other valuable business assets—should be protected.

The trend is to adopt a reasoned and comprehensive strategy that makes data protection a part of the corporate culture and the job of every individual working for the business (partners, associates, paralegals, interns/law students, information technology (IT), HR, executives, administrators, administrative assistants, and other staff).¹ Such an approach is designed to:

- Minimize missteps that can hit the bottom line (costly litigation; time and resources consumed in responding to government, press, or attorney disciplinary commission inquiries or investigations; adverse media coverage; damage to client or customer relationships; and so on), and
- Help businesses achieve a competitive advantage, enhance their profile and image, and enrich their relationship with clients and customers.

1. This trend is in keeping with the “Privacy by Design” (PbD) and “Security by Design” (SbD) movements that are transforming the way that businesses protect data in an information-driven age. See, e.g., *Privacy by Design: The 7 Foundational Principles*, by Ann Cavoukian, Ph.D., Distinguished Expert-in-Residence, Privacy by Design Centre of Excellence, Ryerson University and former Ontario Privacy Commissioner, at <http://www.ryerson.ca/pbdce>; see also FED. TRADE COMM’N, *Start with Security: A Guide for Business*, <http://www.ftc.gov/startwithsecurity> for insights and guidance on SbD gleaned from over 50 FTC data security settlements.

Yet making data protection a part of the corporate culture is easier said than done:

- Properly addressing data protection issues can require a comprehensive understanding of rapidly changing applicable law in 50 states and territories, at the federal level, and globally (where client or customer data might originate, where third parties might be providing U.S.-based businesses with 24/7 services, etc.). Many laws (particularly for government entities and regulated industries) and lawyers' professional rules of responsibility expressly or by implication require appropriate data protection training for employees and sometimes contractors as well.²
- Command of the law is not enough, as businesses are often tried in the court of public opinion or are challenged by third-party watchdog groups, regardless of the current legality of the entity's practices.
- Likewise, technological innovation is occurring at a startling and accelerating pace. The Internet, mobile devices, and ever-more-sophisticated computer technology (all connected to each other and always on) make it easy to collect, analyze, combine, reproduce, and disseminate data, thereby enhancing efficiency and cost-effectiveness but also escalating the risk of making catastrophic mistakes at the speed of light. Yet employees often do not really understand that the latest "smart" technology at work or home (TVs, appliances, toys or gadgets, automated fish tanks, security cameras, digital assistants, voice-controlled smart home hubs, etc.) could be invisibly eavesdropping on confidential discussions using connected microphones, spying via built-in cameras, or providing a new attack vector for accessing the organization's digital assets.

Change is the watchword, and businesses and their cultures must be nimble in spotting trends and addressing issues that were not even on the radar screen months before.

Business leaders often breathe a sigh of relief once the state-of-the-art security system is installed and comprehensive data protection policies and

2. Please see Chapters 4 and 6 of this Handbook for further discussion about the types of legal and professional responsibility requirements placed on lawyers and law firms, which often include education and training.

274 THE ABA CYBERSECURITY HANDBOOK

procedures have been established. Yet notwithstanding adoption of the latest technology and sound data protection principles, businesses are only as strong as their weakest (human) link:

- The disgruntled or downsized “Gen X” employee who has it in for the organization and whose system access was not terminated on the last day of employment.
- The IT director who fails to install patches on a regular basis, thereby leaving networks vulnerable.
- The HR employee who leaves sensitive employee records unlocked or in electronic files with inadequate access restrictions.
- The associate who unwittingly compromises the firm’s client relationships through a lost laptop, phone, or unencrypted flash drive left on an airplane or in a taxi or rideshare vehicle.
- The super-connected, tech-savvy “Millennial” employee who overshares on social media and underestimates how that may sabotage the company’s confidential data.
- The road-warrior employee whose actions (or inaction) regarding the latest mobile technology (including “bring your own device”) may violate internal data security policies or rules of professional responsibility.
- The “Baby Boomer” senior partner who unleashes ransomware by clicking on a link that looks like it came from a colleague or board member.
- The administrative assistant who provides extensive client or firm information after receiving a fraudulent e-mail that appears to be coming from her supervisor or a firm or business executive requesting information.
- The third-party vendor who stores data overseas without appropriate security controls.

Countless studies, audit trails, and surveys over the years have repeatedly confirmed that the biggest data protection threats come from within one’s own organization. Most missteps are unintentional. Many mistakes can be avoided and risks can be minimized with appropriate training and awareness-raising. Yet this is often an overlooked component of data protection initiatives—the missing link when it comes to security.

B. What Does SMART Training Look Like?

What does training actually mean, and what are businesses doing to address data protection's weakest link? Over the years, this question has been posed to CPOs drawn from various industries, locations, and corporate cultures,³ and a consistent pattern of answers has emerged. In short, when conducting training, businesses need to be SMART:

Start training on hiring.

Measure what you do.

Annually provide training.

Raise awareness and provide updates continually.

Tailor training by role.

In considering these SMART training steps and what they mean for one's business, it is important to keep in mind that the particular data protection training that is right for one entity is not necessarily right for another, even if they are in the same industry or are law firms of similar size. Businesses differ in many ways—for example, the degree of centralization, their corporate cultures, the jurisdictions in which they operate, their objectives, their resources and budget, their existing data protection infrastructure, their buy-in from senior management, and so on.

1. *S—Start Training on Hiring*

Given the fundamental role of data in everything a business does, training on how to protect that data should start on day one. Data protection training should be provided to all new employees and, increasingly, to contractors as well. In cases where it is not feasible to do such training for all employees initially (due to bandwidth, budget, or other constraints), businesses might choose to focus training on selected employees (e.g., HR personnel and those in key business roles or units).

3. Chapter co-author Ruth Hill Bro has posed such training questions since 2005 in her recurring column called *CPO Corner: Interviews with Leading Chief Privacy Officers*, which features 17 questions designed to identify trends and best practices, showcase the diverse range of CPOs, and capture key benchmarking and practical implementation information regarding data protection issues; see interviews posted at ABA, Section of Science & Technology Law: E-Privacy Law Committee, <http://www.ambar.org/eprivacy>.

276 THE ABA CYBERSECURITY HANDBOOK

In the employee context, training is provided as a part of employee orientation. Such training can take different forms, using a variety of media:

- An initial in-person, instructor-led session (large group, small group, or one-on-one, as appropriate), which can encourage interaction (but may not always be scalable or practical for some organizations in all situations), and/or
- An intranet/computer-based training module.

Coverage can include a wide range of topics, including:

- High-level overviews applicable to all employees and contractors;
- Instruction on relevant data protection laws and regulations (and professional rules of responsibility, where applicable), internal policies and procedures, fundamentals of the relevant technology, and industry best practices;
- Protecting confidentiality and security of data; and
- Steps to take when addressing a suspected data breach.

Such training should be coordinated with other training (regarding records management, code of conduct, etc.) and should be reviewed to avoid contradictions and conflicts in approach and message. Consideration should be given to whether the time, format, and content are suitable across different parts of the organization. Issues of translation, local law, and local customs can come into play here as well.

2. *M—Measure What You Do*

Measurement and assessment are a core component of many of these initial training sessions as well as in follow-up training. Administering tests (e.g., a graded online quiz) can help to confirm understanding and gauge the overall effectiveness of the training; it can also help to ensure that the work has actually been done. For example, employees and contractors could be required to correctly answer four of five assessment questions at the end of each training section. Broader measurements—such

as comparisons of incidents and types of missteps before and after training—can also help businesses to make training more effective while demonstrating return on investment (which can be important in making the case for budget).

3. *A—Annually Provide Training*

It is prudent (and in some cases required under applicable law, rules, or policies) to ensure that employees annually receive a data protection training update, along with corresponding assessments or tests. Where relevant, certification or continuing legal education (CLE) or professional responsibility credit could be provided, thereby offering an additional incentive to do the training. Such follow-up instruction is often computer-based, so it can be deployed to diverse geographic locations and in a time frame that is convenient for the person receiving the training. Sometimes these annual updates are a part of annual recertification regarding business conduct guidelines.

4. *R—Raise Awareness and Provide Updates Continually*

It is impossible to integrate appropriate data security practices into a culture by using just introductory training on hiring and mandatory annual training. To address this, businesses need to look for ways to raise data awareness and update employees on data protection on an ongoing basis. This is due to a number of factors, including the speed with which the issues change, the different ways in which people learn, the need for reinforcement, and so on. With ongoing awareness-raising, law firms and other businesses can integrate information security practices in such a way that they become as commonplace as turning on a computer.

5. *T—Tailor Training by Role*

Going beyond high-level, one-size-fits-all training allows for training to be tailored to focus on specific roles of individuals, different generational challenges, and specific requirements for contractors and third parties. Tailoring of data protection training can take various forms, depending on the organization:

278 THE ABA CYBERSECURITY HANDBOOK

- Start with a Data Protection 101 online course that is available on demand (successfully completing it results in a certificate). The basic module can then be supplemented by training and awareness-raising specific to role (HR, those involved heavily in data handling, contractors, product design, engineering, sales, senior executives, lawyers, paralegals, administrative assistants, etc.), business unit, geographic location, and the like.
- Determine who should receive direct training from, or at least meet in person with, the CPO, CISO, CIO or IT director, legal counsel, or other qualified trainers. It is helpful for those tasked with training to meet with selected employees to learn about their data practices and then tailor training efforts accordingly. For example, some CPOs meet regularly with the company's engineering, product design, and sales teams to raise important issues in planning meetings and gain insights to develop appropriate training.
- Ask data protection officers (or other relevant individuals) associated with the business lines to develop training and tools to enable the application of data protection policies to their respective areas.
- Hire specialists, internally or externally, to refine and enhance training efforts.
- Not all training and awareness-raising comes from within. Small firms or solo practitioners, those who lack specialized staff, and others looking for cost-effective approaches should take advantage of online training modules, relevant CLE courses and conferences, resources offered by bar associations (the American Bar Association and the ABA's Sections, Divisions, and Forums; state and local bar associations; specialty bar associations; etc.), training publications, and the like.
- As noted above, training for some roles (e.g., lawyers) may be accompanied by certification or CLE or professional responsibility credit.

Businesses that use SMART training can provide the missing link that will help make data protection a part of the culture and turn their employees into one of their strongest links when it comes to protecting one of the most valuable assets of any business.

II. SMART Training in Action

Implementing a SMART training program does not have to be complicated or require significant budget. The program pays for itself by reducing the risk of data loss and increasing awareness about data protection.

A. Understanding the Basics of Employees: Role and Generational Differences

First, any training program should assess and understand the recipients of the training. As mentioned above, the role of the employee in the organization will make a difference in the type of training received. An associate working with e-discovery matters and technology every day will have different considerations than a mail clerk or even other associates and partners in the firm.

In addition, generational differences play a role in how training should be developed, the type of training and communication that a person prefers to receive, and how best to provide the training. Cam Marston, in his practical and often humorous book *Generational Insights: Practical Solutions for Understanding and Engaging a Generationally Disconnected Workforce*,⁴ discusses how each generation differs in its approach to learning:

- **Baby Boomers** (born 1946–1964) tend to continue to hold key leadership positions (e.g., partners) in the organization. They focus on work ethic and often measure it in terms of hours spent, rather than productivity. They value face time and relationships and seek loyalty. They look for those willing to put in whatever time is necessary to complete the task and support the team.⁵ When training Baby Boomers, it is critical to include preevaluation of technology skills and training that is participatory but not intimidating.⁶
- **Generation Xers** (born 1965–1979) often have a more entrepreneurial spirit and are focused on challenging or reinventing the status quo.

4. CAM MARSTON, *GENERATIONAL INSIGHTS: PRACTICAL SOLUTIONS FOR UNDERSTANDING AND ENGAGING A GENERATIONALLY DISCONNECTED WORKFORCE* (2010).

5. *Id.* at 33–34.

6. *Id.* at 39.

280 THE ABA CYBERSECURITY HANDBOOK

They tend to seek open communication, no matter their title or status. Unlike Baby Boomers, Gen Xers focus on productivity rather than time. They often seek a person, not a company, where their loyalty will lie.⁷ Training programs should address the Gen X employee's career goals and be flexible, providing options and choices. For Gen Xers to see training as valuable, senior-level management must demonstrate its commitment to the training as valuable.⁸

- **Millennials** (born 1980–2000) tend to be the most idealistic of the three groups. Unlike Gen Xers, who prefer to work independently with few checkpoints, Millennials want constant communication and positive reinforcement and prefer regular checkpoints at each phase of their work.⁹ Training programs for Millennials need to be group-oriented (where practical), interactive, and fun. They prefer that everyone be allowed to take a role in some part of the teaching as well as the learning.¹⁰

Given the diverse nature of the workforce and the different means by which people learn and absorb information, any training or education campaign must integrate a variety of employee perspectives and capabilities and incorporate a variety of approaches.

B. Building an Effective and Diverse Program

Leveraging the **SMART** principles described above, any organization can quickly and easily build an ongoing training and education campaign.

First, it is critical to make the campaign fun and creative. While the message of data protection is serious, the delivery does not need to be. People of all generations tend to learn more through consistent messaging that has a direct impact on their lives. Those working on the data protection training should develop easy, fun, and catchy slogans that employees will remember.

7. *Id.* at 35.

8. *Id.* at 41.

9. *Id.* at 35, 37.

10. *Id.* at 42.

One example is the SAFE program,¹¹ an information security awareness program that was developed for Option Care Enterprises, Inc. as a way to help employees remember how best to protect and secure sensitive information:

Secure the organization's data: Where are you storing client data? How are you deleting it?

Asset protection: Do you know where your computer/iPad/phone is?

Friend or Foe: Who is sending you an e-mail? Is it something you expected or phishing?

Encrypt: Are you encrypting sensitive e-mails before sending them out?

A program such as SAFE can be used throughout the year to educate staff; different themes within each of the four SAFE categories above can be featured.

Next, identify something, such as a mascot, that represents the organization and symbolizes data protection to help lighten the delivery of a serious message. For example, Option Care, which provides infusion services to patients in their homes, uses a mascot named "The Infuser."

The Infuser's motto is "Infusing Security into Everything We Do." Every time employees see this mascot and message, they are reminded about protecting sensitive information. It is a fun, easy, and quick cue that costs very little to the organization to develop and implement.

Third, ensure that training is continuous, and use various methods to implement it. In addition to mandatory training at specific times (when employees join the organization and subsequent annual training), continuous education is key to any successful cultural transformation. The following are some ideas to keep the momentum going:



© 2016 Option Care Enterprises, Inc. All Rights Reserved.

11. The SAFE program was developed by Option Care CISO Jill Rhodes, who is also an author of this chapter and co-editor of this Handbook. For further information on the program, contact Ms. Rhodes at jill.rhodes@optioncare.com.

282 THE ABA CYBERSECURITY HANDBOOK

- Leverage current newsletters, and place brief articles within them that discuss data protection.
- Conduct e-mail campaigns (monthly or as needed) with data protection guidelines, relevant media coverage, and so on to remind everyone (or otherwise make them aware) of relevant policies and practices.
- Offer periodic “Data Protection Awareness Weeks” or “Security Awareness Drives” with guest speakers and other special events.
- Strategically place wall posters and other communications that promote data protection.
- Publish monthly articles on the company and line-of-business intranet home pages to raise data protection awareness.
- Send periodic e-mails to highlight ongoing opportunities for online training and in-person sessions conducted by members of the data protection team or outside speakers.
- Develop white papers and other material related to relevant data protection topics (aiming for greater frequency and detail over time).
- Remind employees that data protection training is an important part of their job by including it as a factor in their annual performance evaluation; celebrate successes and reward those who meet the objectives (and, if needed, identify opportunities for growth and improvement).

Fourth, involve employees directly. Hold data protection competitions between divisions, offices, or floors in a building with the goal of identifying an employee/group activity that protected the organization’s information in a noteworthy way. Recognize the individual or group winners, name them in the monthly newsletter or blog, and provide a pizza lunch for the winner—the more recognition, the better.

Build an ambassador/liaison or similar program across the organization. Whether it is by office, region, subject matter expertise, or business unit, identify a way to have a data protection representative in each. Although senior leadership is important, the representative should be a mid-level employee who still has influence with peers, subordinates, and leadership. Meet with the data protection ambassadors/liaisons regularly to discuss data protection issues and (in line with the discussion about Millennials above) ensure that they are a part of the solution by having them serve as

the leaders who will train and educate the employees they work with on a regular basis.

Educate employees about how to protect data at home as well as in the workplace. Data protection does not start when a person logs into the network or end when she shuts down for the evening. Everyone's family members and friends are constantly touching sensitive and/or personal digital data. Whether it is through social media or new mobile apps, data is being collected. By educating employees to protect data in all facets of their lives, they will approach data protection more holistically in their daily work life.

All of these methods are easy, cheap, fun, and effective ways to communicate and educate employees about enhancing data protection in the organization. As noted earlier, it is critical to find ways to measure the success of these campaigns (the "M" in "SMART" training).

C. Measuring Success (Through Phishing Campaigns and Other Means)

One of the easiest ways to measure the success of a campaign is to test employees by phishing them directly. Phishing normally occurs when a malicious e-mail is sent either directly to an individual (spear phishing) or to many in the hope that the target will click on a link within the e-mail and then spread a virus that could infect the individual's computer, at a minimum, or the entire enterprise network. Ransomware, discussed throughout this Handbook, has often been caused by phishing.

As part of a SAFE campaign (Friend or Foe), organizations can implement their own company-wide phishing campaign, sending "malicious" e-mails to employees, as someone trying to harm the organization would do. When an employee clicks on the e-mail, instead of infecting the system, the employee receives an educational message about the phishing e-mail and the fact that had it been real, harm could have come to the organization. This type of campaign measures the click rate and, when conducted regularly, can be used to monitor those who are clicking regularly. As a result, specific training can be developed for those individuals or groups. Phishing programs provide a quantitative measurement related to security awareness.

Reporting numbers also can provide both quantitative and qualitative opportunities for measuring success. As more training and education occurs, the number of incidents reported to appropriate leadership will also increase.

284 THE ABA CYBERSECURITY HANDBOOK

Increased reporting could be anything from the reporting of a specific data breach or loss incident to reporting of phishing e-mails. As these incidents are tracked, greater information becomes available about employee knowledge and understanding of data protection.

In the end, a data breach or loss will most likely occur as a result of something an employee did or did not do. The best way to prevent such missteps is to educate the people in the organization about how they can better protect the information around them.

III. Ten Key Points

1. Make data protection a part of the corporate culture and the job of every individual.
2. Recognize that the biggest risks to data come from the people working for the organization and that training and raising awareness are essential to reducing those risks.
3. Be SMART in training: Start training on hiring. Measure what you do. Annually provide training. Raise awareness and provide updates continually. Tailor training by role.
4. Recognize that one size doesn't fit all; it is important to undertake training that fits a business's own needs.
5. Build a program that represents the organization's employees both from a role perspective and a generational one.
6. Make any training campaign fun and interesting—let the employees lead it through ambassador/liaison programs and in other ways.
7. Train employees on how to protect information in all facets of their lives, not just in the workplace. By helping them protect their family and friends at home, they will further integrate these practices at work.
8. Reward! Reward! Reward! Use competitions with prizes to further induce employees to become more aware and supportive of data protection across the organization.
9. Measure success through phishing programs and tracking of reporting of incidents and responses.

10. Know that training and awareness-raising is a never-ending journey (not a destination) that can require changes in direction in response to changes in the law, technology, media coverage, and one's own experiences and new business initiatives. Adapt accordingly, while keeping message delivery mechanisms light and easy to understand for all of the people who work for the organization.

Civil and Human Rights Implications of AI

H

Civil and Human Rights Implications of AI:

Presentation Outline

Steve Crown

Microsoft
Redmond, Washington

Jessica Fjeld

Berkman Klein Center for Internet
& Society, Harvard Law School
Cambridge, Massachusetts

Vivek Krishnamurthy

Samuelson-Glushko Canadian Internet
Policy and Public Interest Clinic,
University of Ottawa
Ottawa, Ontario

What are the key developments in human rights bearing on the impacts of artificial intelligence?

- Sources of law
 - Universal Declaration of Human Rights (“UDHR”) – widely considered binding on states that adhere to the UN Charter
 - International Covenant on Civil and Political Rights (“ICCPR”) and International Covenant on Economic, Social, and Cultural Rights (“ICESCR”) – international treaties binding on states that have ratified them
 - Corporations’ responsibilities to respect human rights laid out in United Nations Guiding Principles on Business and Human Rights
- Human rights impacted by AI
 - Dignity (UDHR Article 1) – respecting unique capacities of each human being, supporting human agency and free development of personhood
 - Freedom from discrimination (UDHR Article 2, ICCPR Article 26) – e.g. risk assessment tools, facial recognition, hiring and credit algorithms, AI-enabled content moderation systems
 - Equality before the law (UDHR Article 7, ICCPR Article 26) – e.g. risk assessment tools
 - Freedom from arbitrary arrest (UDHR Article 9, ICCPR Article 9) – e.g. risk assessment tools, facial recognition
 - Fair public hearing (UDHR Article 10, ICCPR Article 14.1)
 - Innocent until proven guilty (UDHR Article 11, ICCPR Article 14.2) – e.g. – e.g. risk assessment tools, facial recognition

- Privacy (UDHR Article 12, ICCPR Article 17) – e.g. facial recognition, hiring and credit algorithms, AI-enabled content moderation systems
- Freedom of information and expression (UDHR Article 19, ICCPR Article 19) – e.g. hiring and credit algorithms, AI-enabled content moderation systems

What are the key developments in Constitutional law bearing on the impacts of artificial intelligence?

- Sources of law
 - U.S. Constitution and Bill of Rights
 - State Constitutions (often more protective of individual rights than the U.S. Constitution)
 - Analogous case law, e.g. *Carpenter v. United States* (cell phone users have a reasonable expectation of privacy in location data held by providers)
- Constitutional rights impacted by AI
 - Freedom of speech (First Amendment to the U.S. Constitution) – e.g. hiring and credit algorithms, AI-enabled content moderation systems
 - Due process and equal protection (Fifth and Fourteenth Amendments to the U.S. Constitution) – e.g. risk assessment tools, facial recognition
 - Privacy (implied by the First, Third, Fourth, and Fifth Amendments to the U.S. Constitution) – e.g. facial recognition, hiring and credit algorithms, AI-enabled content moderation systems; chilling effect of State surveillance on freedom of speech and freedom of assembly
 - Confrontation Clause (Sixth Amendment) – e.g. algorithmic risk assessments used to determine bail or prison sentences

Further questions for discussion

- To what extent is AI just a force multiplier on existing institutions, or to what extent does it introduce entirely new challenges?
- Do developments in the public and private regulation of facial recognition technology illustrate a way forward?
- What is the role of private statements of ethical principles and codes of conduct, multistakeholder collaboration, academic research, government regulation, and good old fashioned litigation in addressing the impacts of AI, and how can these efforts be coordinated?

A Map of Ethical and Rights-Based Approaches to Principles for AI

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

Designers: Arushi Singh (arushisingh.net) and Melissa Axelrod (melissaaxelrod.com)

Date, Location
Document Title
Author

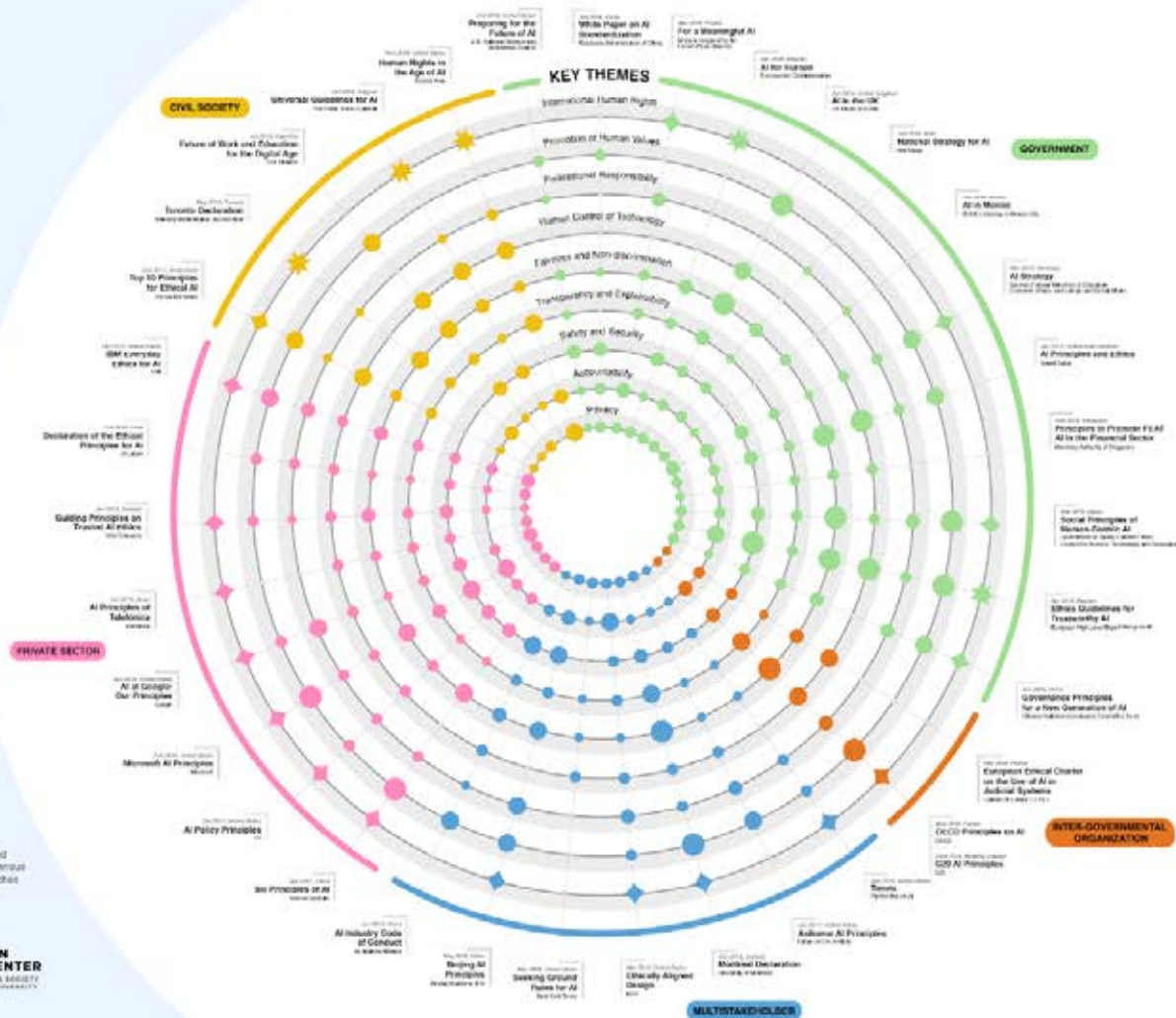
The principles within each theme are:

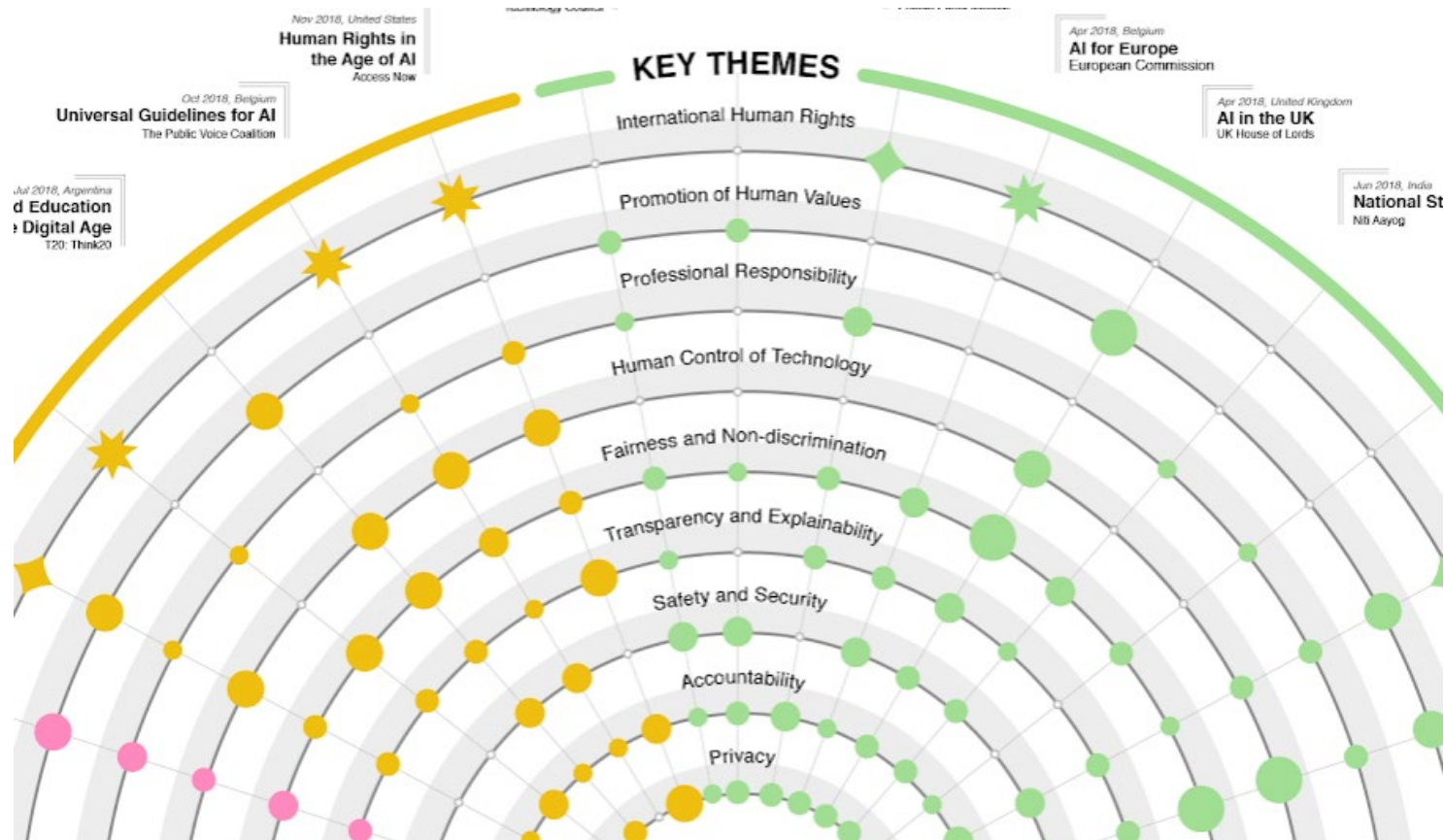
- Privacy
 - Control over Use of Data
 - Consent
 - Privacy by Design
 - Recommendation for Data Protection Laws
 - Ability to Restrict Processing
 - Right to Rectification
 - Right to Erasure
- Accountability
 - Accountability
 - Recommendation for New Regulations
 - Impact Assessment
 - Evaluation and Auditing Requirements
 - Verifiability and Replicability
 - Privacy and Legal Responsibility
 - Ability to Appeal
 - Environmental Responsibility
 - Creation of a Monitoring Body
 - Necessity for Automated Decision

- Security
- Safety and Reliability
- Predictability
- Security by Design

- Explainability
- Transparency
- Open Source Data and Algorithms
- Notification when Interacting with an AI
- Notification when AI Makes a Decision about an Individual
- Regular Flushing Requirement
- Right to Information
- Open Procurement (for Government)
- Fairness and Non-discrimination:**
- Non-discrimination and the Prevention of Bias
- Fairness
- Inclusiveness in Design
- Inclusiveness in Impact
- Representative and High Quality Data
- Equity
- Human Control of Technology:**
- Human Control of Technology
- Human Review of Automated Decision
- Ability to Opt out of Automated Decision
- Professional Interoperability:**
- Multistakeholder Collaboration
- Responsible Design
- Consideration of Long Term Effects
- Accuracy
- Scientific Integrity

Further information on findings and methodology is available in *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches* (Bentham Kline, 2020) available at cyber.harvard.edu

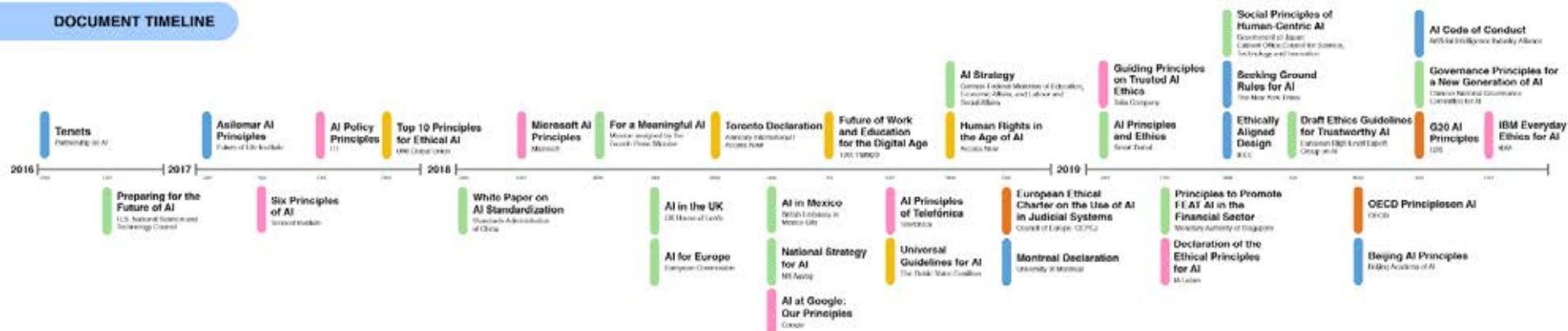




PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

DOCUMENT TIMELINE

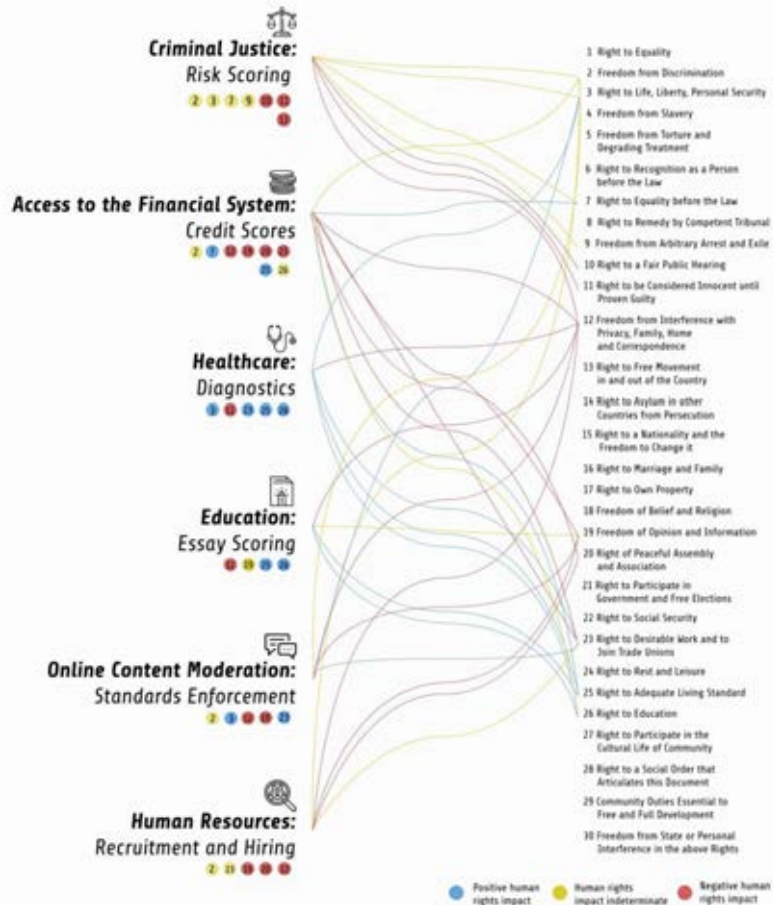


Nature of Actors

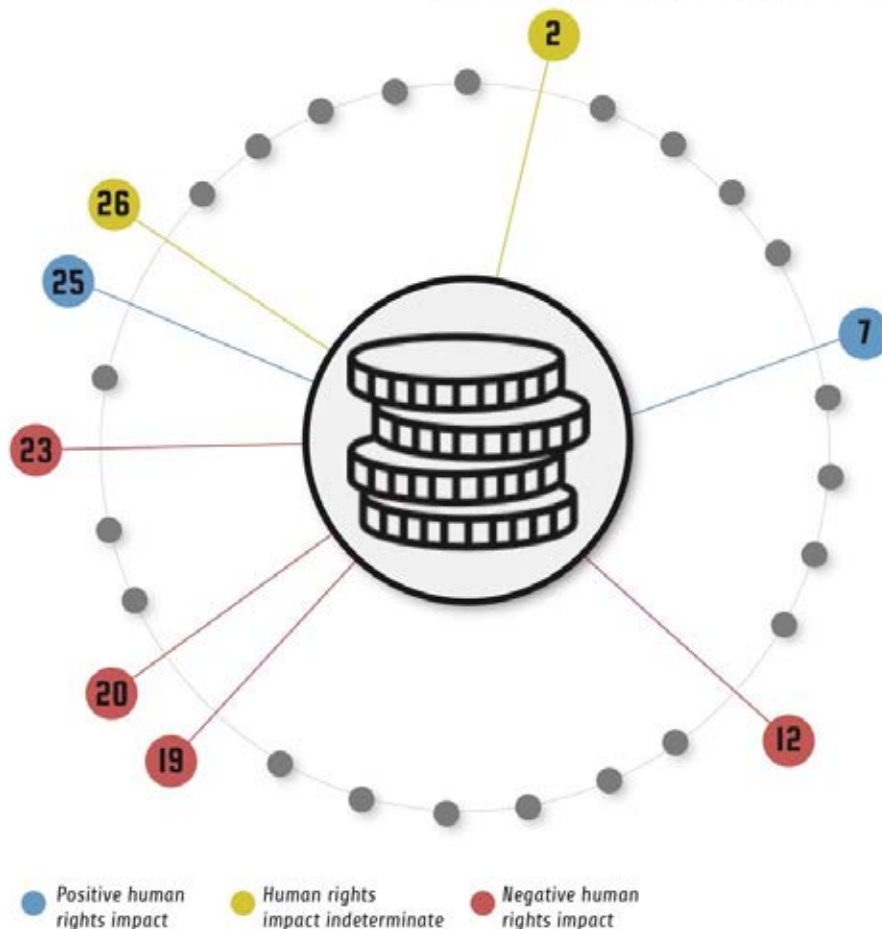
- Civil Society
- Government
- Inter-governmental Organization
- Multistakeholder
- Private Sector

Further information on findings and methodology is available in *Principled Artificial Intelligence: Mapping Consensus and Divergence in Ethical and Rights-Based Approaches* (Berkman Klein, 2020) available at cyber.harvard.edu

ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS



ACCESS TO THE FINANCIAL SYSTEM: CREDIT SCORES



2. Freedom from Discrimination and

7. Right to Equality Before the Law

AI may reduce discrimination in lending by providing more accurate determinations of the creditworthiness of marginalized groups, yet it may also discriminate against them in novel ways.

12. Right to Privacy

AI-based credit scoring systems are premised on the collection, storage, and analysis of vast amounts of personal data, which raises significant privacy concerns.

19. Freedom of Opinion, Expression, and Information, and 20. Right of Peaceful Assembly and Association

Since "all data is credit data" for AI-generated credit scores, people may be chilled from expressing themselves or associating with certain others for fear of how this might impact their ability to borrow.

23. Right to Desirable Work, 25. Right to Adequate Standard of Living, and 26. Right to Education

AI will likely be used to extend credit to people who have been passed over by lenders using traditional credit scores, who can then use this money to improve their economic well-being.

Artificial Intelligence & Human Rights: Opportunities & Risks

Filippo Raso
Hannah Hilligoss
Vivek Krishnamurthy
Christopher Bavitz
Levin Kim

Berkman Klein Center for Internet & Society at
Harvard University

September 25, 2018

Background & Context*

This report explores the human rights impacts of artificial intelligence (“AI”) technologies. It highlights the risks that AI, algorithms, machine learning, and related technologies may pose to human rights, while also recognizing the opportunities these technologies present to enhance the enjoyment of the rights enshrined in the Universal Declaration of Human Rights (“UDHR”). The report draws heavily on the United Nations Guiding Principles on Business and Human Rights (“Guiding Principles”) to propose a framework for identifying, mitigating, and remedying the human rights risks posed by AI.

Readers wishing to better understand the often-paradoxical human rights impacts of the six current AI applications that are detailed in this report are invited to explore a series of interactive visualizations that are available at <https://ai-hr.cyber.harvard.edu>.

* We would like to express our appreciation to our Berkman Klein colleagues Amar Ashar, Ryan Budish, Dan Jones, Rob Faris, Jessica Fjeld, Sarah Newman and Casey Tilton for their helpful suggestions throughout the course of this project; to our interns and research assistants Sam Bookman, Daniel Chase, Christina Chen, Areeba Jibril, Adam Nagy, and Marianne Strassle for their assistance in finalizing this report; and to Urs Gasser and Jonathan Zittrain, respectively the Executive Director and Faculty Chair of the Berkman Klein Center, for their visionary leadership of our Center and of the Ethics and Governance of Artificial Intelligence Fund.

We are grateful to Kim Albrecht, Solon Barocas, Dinah PoKempner, Mark Latonero, An Xiao Mina, Brian Root, Maria Sapignoli, Seamus Tuohy, and Andrew Zick for consulting with us at various stages of this project and helping us refine our analytical methodology.

Finally, we wish to thank the Government of Canada for its sponsorship of this project, and in particular thank Tara Denham, Salahuddin Rafiqhuddin, Philippe-André Rodriguez, Marketa Geislerova, Maroussia Levesque, Asha Mohidin Siad and Jennifer Jeppsson of Global Affairs Canada for their consistent support of this project.

The views expressed in this report are those of the authors alone and do not reflect those of the Government of Canada or of the Berkman Klein Center for Internet & Society at Harvard University.

Summary of Findings

A Human Rights-based Approach to AI's Impacts

The ongoing dialogue regarding the ethics of artificial intelligence (AI) should expand to consider the human rights implications of these technologies.

International human rights law provides a universally accepted framework for considering, evaluating, and ultimately redressing the impacts of artificial intelligence on individuals and society.

Since businesses are at the forefront of developing and implementing AI, the United Nations Guiding Principles on Business and Human Rights are especially salient in ensuring that AI is deployed in a rights-respecting manner.

Determining Impacts

We propose that the best way to understand the impact of AI on human rights is by examining the difference, both positive and negative, that the introduction of AI into a given social institution makes to its human rights impacts. We take this view for two reasons:

1. Determining the human rights impacts of AI is no easy feat, for these technologies are being introduced and incorporated into existing social institutions, which are not rights-neutral.
2. Each application of AI impacts a multitude of rights in complicated and, occasionally, contradictory ways. Exploring these relationships within use cases allows for more nuanced analysis.

Measuring Impacts

Current implementations of AI impact the full range of human rights guaranteed by international human rights instruments, including civil and political rights, as well as economic, cultural, and social rights.

Privacy is the single right that is most impacted by current implementations of AI. Other rights that are also significantly impacted by current AI implementations include the rights to equality, free expression, association, assembly, and work.

Regrettably, the impact of AI on these rights has been more negative than positive to date.

The positive and negative impacts of AI on human rights are not distributed equally throughout society. Some individuals and groups are affected more strongly than others, whether negatively or positively. And at times, certain AI implementations can positively impact the enjoyment of a human right by some while adversely impacting it for others.

Addressing Impacts

Addressing the human rights impacts of AI is challenging because these systems can be accurate and unfair at the same time. Accurate data can embed deep-seated injustices that, when fed into AI systems, produce unfair results. This problem can only be addressed through the conscious efforts of AI systems designers, end users, and ultimately of governments, too.

Many of the existing formal and informal institutions that govern various fields of social endeavor are ill-suited to addressing the challenges posed by AI. Institutional innovation is needed to ensure the appropriate governance of these technologies and to provide accountability for their inevitable adverse effects.

The Path Forward

Human rights due diligence by businesses can help avoid many of the adverse human rights impacts of AI.

Non-state grievance and remedy mechanisms can provide effective redress for some, but by no means all, of the inevitable adverse impacts that AI will produce.

Governments have an important role to play in creating effective mechanisms to remedy the adverse human rights impacts of AI.

The role of government is essential to addressing the distributive consequences of AI by means of the democratic process.

Table of Contents

1.	Introduction	1
2.	What is Artificial Intelligence?	5
3.	What are Human Rights?	8
4.	Identifying the Human Rights Consequences of AI	10
5.	AI's Multifaceted Human Rights Impacts.....	13
5.1.	Criminal Justice: Risk Assessments	17
5.2.	Access to the Financial System: Credit Scores	24
5.3.	Healthcare: Diagnostics.....	32
5.4.	Online Content Moderation: Standards Enforcement.....	37
5.5.	Human Resources: Recruitment and Hiring	44
5.6.	Education: Essay Scoring.....	50
6.	Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach	56
7.	Conclusion	65
8.	Further Reading.....	67
8.1.	Understanding AI.....	67
8.2.	Defining the Problem: What's at Stake?.....	68
8.3.	Approaches to Regulating AI	69
8.4.	Business and AI.....	70

1. Introduction

Artificial intelligence (“AI”) is changing the world before our eyes. Once the province of science fiction, we now carry systems powered by AI in our pockets and wear them on our wrists. Vehicles on the market can now drive themselves, diagnostic systems determine what is ailing us, and risk assessment algorithms increasingly decide whether we are jailed or set free after being charged with a crime.

The promise of AI to improve our lives is enormous. AI-based systems are already outperforming medical specialists in diagnosing certain diseases, while the use of AI in the financial system is expanding access to credit to borrowers that were once passed by. Automated hiring systems promise to evaluate job candidates on the basis of their bona fide qualifications, rather than on qualities such as age or appearance that often lead human decision-makers astray. AI promises to allow institutions to do more while spending less, with concomitant benefits for the availability and accessibility of all kinds of services.

Yet AI also has downsides that dampen its considerable promise. Foremost among these is that AI systems depend on the generation, collection, storage, analysis, and use of vast quantities of data—with corresponding impacts on the right to privacy. AI techniques can be used to discover some of our most intimate secrets by drawing profound correlations out of seemingly innocuous bits of data. AI can easily perpetuate existing patterns of bias and discrimination, since the most common way to deploy these systems is to “train” them to replicate the outcomes achieved by human decision-makers. What is worse, the “veneer of objectivity” around high-tech systems in general can obscure the fact that they produce results that are no better, and sometimes much worse, than those hewn from the “crooked timber of humanity.”

These dystopian possibilities have given rise to a chorus of voices calling for the need for Fairness, Accountability, and Transparency in Machine Learning (“FAT” or “FAT/ML”). Advocates of this approach view the response to AI’s potential problems in terms of ethics. For example, the Institute of Electrical and Electronics Engineers—the world’s largest technical professional body that plays an important role in setting technology standards—has published an influential treatise on *Ethically Aligned Design* that suggests that “the full benefit of these technologies will be attained only if they are aligned with our defined values and ethical principles.”¹ In a similar vein, the

¹ IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically

governments of France and India have recently released discussion papers to frame their national strategies on AI that embrace an ethics-based approach to addressing the social impacts of these technologies.²

During the pendency of this project, however, several influential actors have come to recognize the value of examining the challenges around AI from a human rights perspective.³ This incipient conversation on AI and human rights has already produced two significant documents. One is the Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems (“Toronto Declaration”), which was opened for signatures on May 16, 2018.⁴ As its full title suggests, the Toronto Declaration highlights the potential adverse effects of machine learning on rights to equality and non-discrimination and calls for the development of effective remedial mechanisms for all those who are adversely affected by these systems.⁵ The other is Global Affairs Canada’s Draft Strategy Paper on the Human Rights and Foreign Policy Implications of AI, which examines how AI can impact the rights to equality, privacy, free expression, association, and assembly, and suggests ways that these impacts can be redressed.⁶

Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.” Version 2.

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

² For France’s strategy, see: Cédric Villani, “For a Meaningful Artificial Intelligence: Towards a French and European Strategy” (AI For Humanity), accessed June 22, 2018, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. For India’s, see Amitabh Kant, “National Strategy for Artificial Intelligence” (NITI Aayog, June 2018),

www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. Although both papers are substantial works that are each over 100 pages long, they barely mention the concept of human rights.

³ For example, Amnesty International launched a structured initiative on Artificial Intelligence and Human Rights in 2017, while the New York-based Data & Society Research Institute hosted a workshop on Artificial Intelligence and Human Rights in April, 2018. See Sherif Elsayed-Ali, “Artificial Intelligence and the Future of Human Rights,” Oct. 19, 2017. <https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5>. Mark Latonero, “Artificial Intelligence & Human Rights: A Workshop at Data & Society.” May 11, 2018. <https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149>.

⁴ Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems, May 16, 2018. <https://www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-DoV2.pdf>.

⁵ Ibid.

⁶ Digital Inclusion Lab, Global Affairs Canada, “Artificial Intelligence: Human Rights & Foreign Policy Implications.” Accessed June 1, 2018. https://docs.google.com/document/d/1fhIJYznWSI7oD3TVJ5CgLGHJMj2HouEZiQ9a_qKbLGo/edit (“GAC Strategy Paper”).

This project is rooted in the belief that there is considerable value in adopting a human rights perspective to evaluating and addressing the complex impacts of AI on society. The value lies in the ability of human rights to provide an agreed set of norms for assessing and addressing the impacts of the many applications of this technology, while also providing a shared language and global infrastructure around which different stakeholders can engage.⁷

While there are many different conceptions of human rights, from the philosophical to the moral, we in this project take a legal approach. We view human rights in terms of the binding legal commitments the international community has articulated in the three landmark instruments that make up the International Bill of Rights.⁸ This body of law has developed over time with the ratification of new treaties, the publication of General Comments that authoritatively interpret the provisions of these treaties, and through the work of international and domestic courts and tribunals, which have applied the provisions of these treaties to specific cases.

Our project seeks to advance the burgeoning conversation on AI and human rights by mapping the human rights impacts of the current deployment of AI systems in six different fields of endeavor. We strive to move beyond the predominant focus on AI's impact on select civil and political rights, to consider how these technologies are impacting other rights guaranteed by international law—especially economic, social, and cultural rights.

In so doing, we suggest what we believe to be the optimal method for identifying the human rights impacts of introducing a particular AI system into a given field of endeavor. Simply put, we believe it is important to recognize that AI systems are not being deployed against a blank slate, but rather against the backdrop of social conditions that have complex pre-existing human rights impacts of their own. This may well appear to be a self-evident truth, but in our view, the existing literature does not adequately consider the impact of these background conditions on the consequences of introducing AI. As a result, human rights impacts, both positive and negative,

⁷ Jason Pielemeier, “The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence,” Data & Society: Points, June 6, 2018, <https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a>.

⁸ The “International Bill of Rights” is a term to describe the three most important international human rights instruments, namely the Universal Declaration of Human Rights (“UDHR”), the International Covenant on Civil and Political Rights (“ICCPR”), and the International Covenant on Economic, Social, and Cultural Rights (“ICESCR”).

may be misattributed to AI, contributing to the extreme claims of optimists and pessimists alike about the extent to which AI is changing our lives.

Our report and the accompanying visualizations make clear that AI is already impacting the enjoyment of the full range of human rights—sometimes in paradoxical ways. In the final section, we examine and evaluate how international human rights law generally, and the growing field of business and human rights specifically, can help the developers, users, and regulators of AI systems to address many of these impacts.

2. What is Artificial Intelligence?

Despite its expanding presence across many aspects of our lives, there is no widely accepted definition of “artificial intelligence.”⁹ Instead, it is an umbrella term that includes a variety of computational techniques and associated processes dedicated to improving the ability of machines to do things requiring intelligence, such as pattern recognition, computer vision, and language processing.¹⁰ With such a loose conceptualization and given the rapid growth of technology, it is no surprise that what is considered artificial intelligence changes over time. This is known as the “AI effect” or the “odd paradox”: formerly cutting-edge innovations become mundane and routine, losing the privilege of being categorized as AI, while new technologies with more impressive capabilities are labelled as AI instead.¹¹

The impossibly large set of technologies, techniques, and applications that fall under the AI umbrella can be usefully classified into two buckets. The first is comprised of *knowledge-based systems*, which are “committed to the notion of generating behavior by means of deduction from a set of axioms.”¹² These include “expert systems” which use formal logic and coded rules to engage in reasoning. Such systems, which are sometimes also called “closed-rule algorithms,” include everything from commercial tax preparation software to the first generation of healthcare diagnostic decision support algorithms. These systems are good at taking concrete situations and reasoning optimal decisions based on defined rules within a specific domain. They cannot, however, learn or automatically leverage the information they have accumulated

⁹ National Science and Technology Council: Committee on Technology, “Preparing for the Future of Artificial Intelligence,” Government Report (Washington, D.C.: Executive Office of the President, October 2016).

¹⁰ One seminal textbook categorizes AI into (1) systems that think like humans (e.g., cognitive architectures and neural networks); (2) systems that act like humans (e.g., pass the Turing test, knowledge representation, automated reasoning, and learning), (3) systems that think rationally (e.g., logic solvers, inference, and optimization); and (4) systems that act rationally (e.g., intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting), Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence (Englewood Cliffs, N.J.: Prentice Hall, 1995).

¹¹ Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).

¹² Nello Cristianini, “On the Current Paradigm in Artificial Intelligence,” *AI Communications* 27, no. 1 (January 1, 2014): 37–43, <https://doi.org/10.3233/AIC-130582>.

over time to improve the quality of their decision-making (unless they are paired up with some of the techniques described below).¹³

The second bucket of technologies uses statistical learning to continuously improve their decision-making performance. This new wave of technology, which encompasses the widely-discussed techniques known as “machine learning” and “deep learning,” has been made possible by the exponential growth of computer processing power, the massive decline in the cost of digital storage, and the resulting acceleration of data collection efforts.¹⁴ Systems in this category include self-driving vehicles, facial recognition systems used in policing, natural language processing techniques that are used to automate translation and content moderation, and even algorithms that tell you what to watch next on video streaming services. While these systems are impressive in their aggregate capacities, they are probabilistic and can thus be unreliable at the individual level. For example, deep learning computer vision systems can classify an image almost as accurately as a human; however, they will occasionally make mistakes that no human would make—such as mistaking a photo of a turtle for a gun.¹⁵ They are also susceptible to being misled by “adversarial examples,” which are inputs that are tampered with in a way that leads an algorithm to output an incorrect answer with high confidence.¹⁶

In this report, we focus on AI systems from both of these conceptual buckets that “perceive[] and act[]”¹⁷ upon the external environment by “tak[ing] the best possible action in a situation.”¹⁸ Simply put, the scope of our report is limited to analyzing those AI systems that automate the making of decisions that were formerly the exclusive province of human intelligence. This view of AI embraces everything from medical diagnostic software that determine what is ailing a patient based on the available evidence, to self-driving vehicles that “decide” whether to steer, accelerate, or brake, millisecond by

¹³ Bruce G. Buchanan, “Can Machine Learning Offer Anything to Expert Systems?,” *Machine Learning* 4, no. 3–4 (December 1, 1989): 251–54, <https://doi.org/10.1023/A:1022646520981>.

¹⁴ Gheorghe Tecuci, “Artificial Intelligence,” *Wiley Interdisciplinary Reviews: Computational Statistics* 4, no. 2 (2012): 168–80, <https://doi.org/10.1002/wics.200>.

¹⁵ Adam Conner-Simons, “Fooling Neural Networks w/3D-Printed Objects,” MIT Computer Science & Artificial Intelligence Lab (blog), November 2, 2017, <https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects>.

¹⁶ Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and Harnessing Adversarial Examples,” *ArXiv:1412.6572 [Cs, Stat]*, December 19, 2014, <http://arxiv.org/abs/1412.6572>; A Nguyen, J Yosinski, and J Clune, “Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” *CVPR, IEEE*, 15 (2015)

¹⁷ Russell and Norvig, *Artificial Intelligence*, 7.

¹⁸ *Ibid.*, 27.

millisecond. The crucial factor for us is that the system must function and impact the external environment, rather than simply be a theoretical construct that remains under development, to be considered within the scope of our report. Furthermore, we limit our scope to AI technologies that are either currently in use or are far along in the development process; therefore, we do not delve into the realm of artificial general intelligence.¹⁹ We restrict our consideration of AI in this report to those technologies that are being used to make decisions with real-world consequences for the simple reason that these are the technologies that are most likely to have discernable human rights impacts. By contrast, many other strains of AI research remain conceptual for now, and are thus yet to impact human rights.

¹⁹ Broadly speaking, an AGI system is one that can perform any task as well as a human can, or a “synthetic intelligence that has a general scope and is good at generalization across various goals and contexts,” Ben Goertzel, “Artificial General Intelligence: Concept, State of the Art, and Future Prospects,” *Journal of Artificial General Intelligence* 5, no. 1 (December 1, 2014): 1–48, <https://doi.org/10.2478/jagi-2014-0001>.

3. What are Human Rights?

As noted in the introduction, in this report we adopt a legal conception of human rights. We use the term human rights to refer to those individual and collective rights that have been enshrined first and foremost in the Universal Declaration of Human Rights (“UDHR”), and then further detailed in the International Covenant on Civil and Political Rights (“ICCPR”) and the International Covenant on Economic, Social and Cultural Rights (“ICESCR”).

The UDHR is the leading statement of the rights that every human being enjoys by virtue of their birth. Although the UDHR was adopted by means of a non-binding U.N. General Assembly resolution,²⁰ Canada and many other states have long believed that there is an “obligation on states to observe the human rights and fundamental freedoms enunciated in the [UDHR] [that] derives from their adherence to the Charter of the United Nations,” which is binding international law.²¹

The ICCPR and the ICESCR, meanwhile, are international treaties that are binding upon those states that have ratified them. These treaties elaborate upon the human rights that were first articulated by the UDHR at the international level, and clarify the duties of states in relation to two categories of rights. Whereas the ICCPR’s protections of civil and political rights come into force immediately upon ratification,²² the ICESCR instead requires states to take measures to progressively realize the economic, social, and cultural rights it protects, having due regard for the state’s economic condition and resources.²³

States shoulder a binding obligation under international law to protect human rights. This includes a duty to respect human rights in their own conduct, and to prevent natural and juridical persons subject to their jurisdiction (including corporations) from committing human rights abuses. These obligations persist even

²⁰ Universal Declaration of Human Rights (10 Dec. 1948), U.N.G.A. Res. 217 A (III) (1948) [hereinafter “UDHR”].

²¹ Letter from the Legal Bureau, Jan. 9, 1979, reprinted in *Canadian Practice in International Law*, 1980 Can. Y.B. Int’l L. 326.

²² International Covenant on Civil and Political Rights (New York, 16 Dec. 1966) 999 U.N.T.S. 171 and 1057 U.N.T.S. 407, entered into force 23 Mar. 1976, art. 2 [hereinafter “ICCPR”].

²³ International Covenant on Economic, Social and Cultural Rights (New York, 16 Dec. 1966) 993 U.N.T.S. 3, entered into force 3 Jan. 1976, art. 2(1) [hereinafter “ICESCR”].

when privatizing the delivery of services that may impact human rights.²⁴

Especially since the end of the Cold War, businesses have come to be viewed as having their own responsibilities under international law to respect human rights.²⁵ The nature and scope of these responsibilities have been articulated most authoritatively in the United Nations Guiding Principles on Business and Human Rights (“UNGP” or “Guiding Principles”). Specifically, the responsibility to respect human rights requires enterprises to avoid causing or contributing to adverse human rights impacts through their own activities, and to seek to prevent or mitigate such impacts when the enterprise is “directly linked” to them via a business relationship.²⁶ This, in turn, requires enterprises to engage in ongoing due diligence processes to identify, prevent, and mitigate salient human rights risks.²⁷ To the extent that adverse human rights impacts do occur, businesses should provide remediation for those impacts through legitimate mechanisms²⁸—although it is emphatically the duty of the state to provide effective remedies through judicial and other mechanisms to those who have suffered business-related human rights abuses.²⁹

Although the Guiding Principles do not themselves have the force of law, they clarify how pre-existing international human rights standards apply to business activities, and provide useful guidance on how businesses can operate in a rights-respecting manner.³⁰ In any event, since businesses are at the forefront of developing and deploying AI, the Guiding Principles are of immense importance to ensuring that the human rights impacts of these powerful new technologies are positive. Consequently, the Guiding Principles will feature prominently in the discussion that follows of the human rights impacts of AI systems that are currently in use, and in our suggestions regarding how they should be addressed.

²⁴ Human Rights Council Res. 17/4, Rep. of the Hum. Rts. Council, 17th Sess., June 16, 2011, U.N. Doc. A/HRC/RES/17/4 (July 6, 2011); Special Rep. of the Sec’y Gen., Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework, Hum. Rts. Council, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter “Guiding Principles”], Principle 5.

²⁵ Guiding Principles, Principle 11.

²⁶ *Ibid.*, Principle 13.

²⁷ *Ibid.*, Principle 17.

²⁸ *Ibid.*, Principle 22.

²⁹ *Ibid.*, Principle 25.

³⁰ Justine Nolan, “The Corporate Responsibility to Respect Human Rights: Soft Law or Not Law?,” in *Human Rights Obligations of Business*, ed. Surya Deva and David Bilchitz (Cambridge: Cambridge University Press, 2013), 138–61, <https://doi.org/10.1017/CBO9781139568333.010>.

4. Identifying the Human Rights Consequences of AI

AI is not being developed in a vacuum or deployed against a blank slate. Rather, specific actors in society are deploying AI to automate decision-making in particular fields of endeavor. They are doing so to achieve outcomes that they view as desirable, against the backdrop of social institutions that have their own, pre-existing human rights implications.

Consider, for example, the deployment of AI in the criminal justice system, which is discussed in more detail in the first case study below. Over the course of the last several hundred years, criminal defendants have been endowed with various rights to ensure the fairness of criminal proceedings. These include the presumption of innocence,³¹ the principle of legality,³² the right to a fair trial,³³ and many others. Even so, no existing criminal justice system comes close to perfectly respecting the rights of defendants and other relevant rights-holders: every such system has at least some negative impacts on rights-holders that predate the introduction of AI.³⁴

It is only by embracing a comparative approach, that accounts for background conditions from the pre-AI world, that we can properly understand the human rights impacts of introducing AI into the criminal justice system or any other human institution. Unless the human rights implications, both positive and negative, of pre-existing institutional structures are identified and accounted for, the human rights impacts of introducing AI will be conflated with the ongoing impacts of whatever was there before. Below, we propose a two-step methodology for avoiding such difficulties.

Step 1: Establish the Baseline

As noted, the first step is to simply consider the existing human rights implications, both positive and negative, of whatever field of endeavor AI is being introduced into. This evaluation properly involves consideration of the availability and effectiveness of institutional mechanisms that are currently in place to regulate and redress the negative human rights implications arising from that

³¹ UDHR art. 11.

³² UDHR art. 11(2).

³³ ICCPR art. 14(1).

³⁴ For the last two years, the Berkman Klein Center has been conducting extensive research on the use of algorithms in the criminal justice system, in its capacity as one of the two anchor institutions for the Ethics and Governance of Artificial Intelligence initiative. The research outputs of this ongoing work can be found at <https://cyber.harvard.edu/research/ai>.

field. When human decision-making in the field in question has already been supplanted by a first-generation automated decision-making technology, such as a closed-rule diagnostic algorithm, the first step consists of evaluating the human rights implications of the pre-AI status quo.

Step 2: Identify the Impacts of AI

The second step involves identifying how the introduction of AI changes the human rights impacts of the field into which the technology is introduced. If the introduction of AI improves the human rights performance of the field, AI can be said to have a positive impact on human rights. That is true even if the field of endeavor continues to produce adverse human rights impacts after the introduction of AI. Conversely, if the human rights performance of the field of endeavor deteriorates with the introduction of AI, then it is clear that the technology has produced adverse human rights impacts. To a significant extent, the outcome of this evaluation will depend on whether the mechanisms currently in place to regulate and remedy the adverse human rights consequences of the field in question continue to be effective following the introduction of AI.

The human rights impacts of AI stem from at least three sources, two of which can be considered by conducting a human rights impact assessment before a particular system is deployed. The third source, meanwhile, can be hard to identify even after an AI system is in operation, due to the complexity of the technology:

1. *Quality of Training Data:* To the extent that the data used to “train” an AI system is biased, the resulting system will reflect, or perhaps even exacerbate, those biases.³⁵ This is a version of what is known as the “garbage in, garbage out” problem, and it can have profound consequences for a wide variety of human rights—depending on what the system is intended to do.
2. *System Design:* Decisions made by an AI system’s human designers can have significant human rights consequences. Human designers can, for example, prioritize the variables they would like the AI system to optimize and decide what variables the AI should take into consideration as it operates. Such design decisions can have both positive and negative human rights

³⁵ Osonde Osoba & William Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. (Santa Monica: Rand Corporation, 2017). https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

impacts, which will be informed by the individual life experiences and biases of the designers. Some of these impacts will be foreseeable, while others will not be.

3. *Complex Interactions*: Once an AI system is introduced, it will interact with the environment in ways that produce outcomes that might not have been foreseen. These complex interactions can have significant human rights impacts. In some cases, the impacts of these interactions may be detectable through the use of certain analytical techniques, but the possibility exists that certain human rights impacts resulting from the deployment of an AI system will escape detection. This is not an issue that is unique to AI: pre-digital societies are staggeringly complex, and the human rights impacts of the actions of individuals and institutions are not always knowable at the time they are made or for some time thereafter.

Limitations of our approach

Our two-step methodology provides a useful, generalizable approach to identifying the positive and negative implications of introducing AI into an extant field of endeavor. This methodology, which we have validated in consultations with stakeholders from the technology and human rights communities, undergirds our assessment of the human rights impacts of AI across six different use cases below.

Our framework has its limitations, especially due to the scarcity of available information into the design and operation of any given AI system. This is due in part to the novelty of AI, but also because so much AI technology is proprietary, which results in information about the design, operation, and impact of the systems being treated by their creators as commercially sensitive information.³⁶

Consequently, the analysis we undertake in our six case studies, below, is at the level of detail that one would find in a sectoral human rights impact assessment. Based on our desktop research, we have drawn reasonable inferences as to the likely human rights impacts of introducing particular AI systems into the prevailing social and institutional context.

³⁶ Rebecca Wexler, “Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System,” *Stanford Law Review* 70, no. 5 (2017): 1343–1429, <https://doi.org/10.2139/ssrn.2920883>.

5. AI's Multifaceted Human Rights Impacts

Applying the two-step framework from the previous section, we now explore the wide-ranging human rights consequences of introducing AI decision-making into six fields:

- Criminal Justice (risk assessments)
- Finance (credit scores)
- Healthcare (diagnostics)
- Content Moderation (standards enforcement)
- Human Resources (recruitment and hiring)
- Education (essay scoring)

We chose these six fields out of many possibilities because they illustrate the promise and the perils of this technology across a range of human rights. What is more, AI decision-making technologies are already in use in all of these fields, which allows our analysis to be grounded in the here and now, rather than speculating about future developments.

In choosing these six use cases, we consciously decided not to include two AI applications that have generated a great deal of debate and controversy: namely, self-driving vehicles and autonomous weapons systems. We excluded these applications from our analysis because both are much better studied than the use of AI in the other six fields that we have chosen. Furthermore, the issues surrounding autonomous weapons systems are more appropriately answered with reference to international humanitarian law rather than international human rights law, since such systems are meant to be used in times of conflict.

In undertaking this analysis, it quickly became apparent to us that each AI deployment had the potential to impact a large number of rights via their first- and second-order effects. In the interest of clarity and analytical efficiency, however, we have focused our analysis on those rights that we believe to be most impacted by the deployment in question. This is an exercise in line-drawing that is subjective by its very nature, but is part and parcel of the approach embraced by the Guiding Principles to identifying human rights impacts so that they may be appropriately addressed.³⁷

There are five main points that emerge from our analysis.

³⁷ Guiding Principles, Principles 17 and 24.

First and foremost, the six use cases we explore in detail reveal how AI-based decision-making technologies impact the full spectrum of political, civil, economic, social, and cultural rights secured by the UDHR and further expounded upon in the ICCPR and the ICESCR.

Second, the positive and negative human rights impacts caused by AI are not evenly distributed across society. Some individuals and groups experience positive impacts from the very same applications that adversely impact other rights-holders. In some cases, a particular AI application can positively impact the enjoyment of a given human right for a particular class of individuals, while adversely affecting the enjoyment of the very same human right by others. For example, the use of automated risk scoring systems in the criminal justice system may reduce the number of individuals from the majority group who are needlessly incarcerated, at the very same time that flaws in the system serve to increase the rate of mistaken incarcerations for those belonging to marginalized groups.³⁸

Third, AI carries the serious risk of perpetuating, amplifying, and ultimately ossifying existing social biases and prejudices, with attendant consequences for the right to equality. This problem, which has been termed by one analyst as “counter-serendipity,” results from the fact that AI systems are trained to replicate patterns of decision-making they learn from training data that reflects the social status quo—existing human biases, entrenched power dynamics and all.³⁹ But therein lies the problem: to the extent that an AI accurately replicates past patterns of human decision-making, it will necessarily perpetuate existing social biases as well.⁴⁰ What is worse, unlike human decision-makers, who have the agency and the free will to change their moral perspective over time, for the foreseeable future AI systems will not have any such capabilities of their own. Instead, they require constant attention by those who are responsible for the design and operation of such systems to ensure that their outputs are consistent with evolving notions of fairness.

To be sure, the automation of decision-making through AI offers the possibility of righting significant social wrongs by designing the

³⁸ Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, “Machine Bias,” *ProPublica*, May 23, 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

³⁹ Edward Tenner, *The Efficiency Paradox: What Big Data Can't Do* (New York: Alfred A. Knopf, 2018); Berkman Klein Center for Internet and Society, “Artificial Intelligence and Inclusion,” accessed June 22, 2018, <https://aiandinclusion.org/>.

⁴⁰ Anupam Chander, “The Racist Algorithm?,” *Mich. Law Review* 115, no. 6 (2017): 1023, http://michiganlawreview.org/wp-content/uploads/2017/04/115MichLRev1023_Chander.pdf.

systems to have ameliorative effects. Such effects could be achieved by seeking to correct for biases in human decision-making, or more controversially, through “algorithmic affirmative action”—that is, by designing algorithms to counter the historical disadvantages that marginalized groups have faced.⁴¹ The larger point, however, is that unless AI systems are consciously designed and consistently evaluated for their differential impacts on different populations, they have the very real potential to hinder rather than help progress towards greater equity.

Fourth, as is likely expected, most AI technologies have a deleterious impact on the right to privacy. AIs are data-hungry by their nature; they are fundamentally premised on algorithms automatically poring over vast datasets to generate answers, predictions, and insights. Accordingly, AI systems rely on the collection, storage, consolidation, and analysis of vast quantities of data. They also create powerful incentives to gather and store as much additional data as can be, in view of the possibility that new data streams will allow for AI systems to generate powerful new insights. Much of the data that fuels AI systems will either be personally identifiable, or rife with the possibility of being re-identified using an algorithm in the event that it was anonymized. Moreover, even if techniques such as differential privacy⁴² are used to protect the privacy of particular individuals, AI technologies may generate insights from such data that are then used to make predictions about, and act upon, the intimate characteristics of a particular person—all while refraining from identifying the natural person. For example, a retailer might train an AI-based marketing system using sales data that has been de-identified and subjected to differential privacy techniques. But even assuming that the training data is discarded once the system is in operation, the insights generated by the system from the data it is tasked with analyzing can nevertheless have a significant impact on an individual’s privacy.⁴³ Given that most extant AI applications have very significant privacy implications, we focus our analysis in the case studies below on the other rights that are impacted by these systems. This is a pragmatic choice made in the interests advancing the AI and human rights conversation beyond privacy.

⁴¹ Ibid.

⁴² Cynthia Dwork et al., “Calibrating Noise to Sensitivity in Private Data Analysis,” in *Theory of Cryptography*, ed. Shai Halevi and Tal Rabin (Springer Berlin Heidelberg, 2006), 265–84.

⁴³ Charles DuHigg, “How Companies Learn Your Secrets,” *The New York Times*, February 16, 2012, sec. Magazine, <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>.

Fifth, the rise of artificial intelligence poses a challenge for many of the existing mechanisms that currently exist to right wrongs. In the United States, for example, individuals have a right to request a copy of their credit report and to require credit reporting agencies to investigate and correct any errors appearing on their report.⁴⁴ By contrast, there is currently no law in the United States that would provide an individual with recourse if a lender using an algorithm that crunches through thousands of variables from thousands of sources does so on the basis of erroneous data. Even in Canada⁴⁵ and the European Union,⁴⁶ where privacy laws currently in force allow individuals to demand the correction of errors in their data, the sheer volume of information that AI systems use as they make a decision makes it difficult to exercise this right effectively. Moreover, even if one aggrieved individual corrects errors in their own data, significant harms can occur due to the presence of systematic errors in a data set and ubiquitous data sharing, which can lead to unfair outcomes for potentially vast numbers of people.

⁴⁴ *Fair Credit Reporting Act*, 15 U.S.C. § 1681i (2012) (“...if the completeness or accuracy of any item of information contained in a consumer’s file . . . is disputed by the consumer . . . the agency shall, free of charge, conduct a reasonable reinvestigation to determine whether the disputed information is inaccurate and record the current status of the disputed information, or delete the item from the file[.]”); *Fair Credit Reporting Act*, 15 U.S.C. § 1681j (2012) (free annual copy of one’s credit report).

⁴⁵ *Personal Information Protection and Electronic Documents Act*, S.C. 2000, c. 5 (as amended June 23, 2015), Schedule 1 Principle 4.9 (“Upon request, an individual shall be informed of the existence, use, and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.”)

⁴⁶ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. (L. 119) [henceforth “GDPR”], art. 19 (“The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.”).

5.1. Criminal Justice: Risk Assessments

The criminal justice system is the most potent and fearsome institution through which democratic societies may restrict an individual's enjoyment of their fundamental human rights. In view of the severity of its impacts on human rights, society has evolved a system of procedural rights to protect criminal defendants and convicts from the vagaries of human decision-making, from intentional abuse of power to unconscious influences ranging from racism to fatigue.⁴⁷

In search of both fairness and efficiency, justice systems are increasingly employing automated decision-making tools at every procedural stage. This is especially true of risk assessments, which are used to inform decisions about pretrial detention, sentencing, and parole. To the extent that they are fair and accurate, risk assessment tools can have a significant positive impact on the rights of individuals accused and convicted of crimes. The corollary, however, is that flaws or unknown limitations in the operation of such systems can have deleterious effects on a wide range of rights.

5.1.1. *Traditional Approach to Risk Assessments*

The first efforts to formalize the process of assessing an individual's risk of recidivism date back to the 1920s, when statisticians began to identify objective factors that are predictive of this risk for parolees.⁴⁸ As with AI now, the force driving the development of these earlier tools was the desire to avoid unnecessary deprivations of liberty and reduce the incidence of discrimination in the criminal justice system attributable to human bias. Statisticians developed these tools by collecting and analyzing information about defendants to identify factors that distinguish those that reoffend from those who do not.

⁴⁷ Millicent H. Abel and Heather Watters, "Attributions of Guilt and Punishment as Functions of Physical Attractiveness and Smiling," *The Journal of Social Psychology* 145, no. 6 (December 2005): 687–702, <https://doi.org/10.3200/SOCP.145.6.687-703>; Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 17 (April 26, 2011): 6889–92, <https://doi.org/10.1073/pnas.1018033108>.

⁴⁸ Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. (Chicago: University of Chicago Press, 2006) 48–60; James Bonta, "Risk-Needs Assessment and Treatment," in *Choosing Correctional Options That Work: Defining the Demand and Evaluating the Supply*, ed. Alan T. Harland (Thousand Oaks, Calif: Sage Publications, 1996); Thomas Mathiesen, "Selective Incapacitation Revisited," *Law and Human Behavior* 22, no. 4 (1998): 455–69.

As these assessments became more sophisticated, statisticians began to consider both static factors, such as a defendant's age and gender, as well as dynamic factors, such as a defendant's skill set or psychological profile.⁴⁹ Over time, these efforts led to the development of risk assessment inventories such as the Level of Service Inventory-Revised ("LSI-R")⁵⁰ that, while developed and validated by statisticians, are deployed in the field by individuals without much if any statistical expertise. Especially when such tools require their operators to make subjective determinations, such as whether an individual is engaging in antisocial behavior,⁵¹ these tools may suffer from low inter-rater reliability ("IRR"), calling into question the validity of the predictions generated by such tools for any given individual.⁵² Furthermore, the data available to actuarial risk assessment systems to identify who is truly at a high risk of re-offending is systematically skewed by the fact that the pre-existing system has sentenced those it believes to pose the highest risk to long prison sentences, during which time those inmates cannot reoffend.⁵³

Risk assessment tools in the U.S. criminal justice system have been critiqued as inherently unfair due to the disproportionate targeting of minority individuals and communities by the police.⁵⁴ This, in turn, raises the risk that such tools will miscalculate the risk of recidivism for individuals from minority versus majority communities. Moreover, as the Supreme Court of Canada recently noted in *Ewert v. Canada*, risk assessment tools that are developed and validated based on data from majority groups may lack validity in predicting the same traits in minority groups.⁵⁵ This may have

⁴⁹ James Bonta, "Risk-Needs Assessment and Treatment."

⁵⁰ Ibid.

⁵¹ Thomas H. Cohen, "Automating Risk Assessment Instruments and Reliability: Examining an Important but Neglected Area in Risk Assessment Research," *Criminology & Public Policy* 16, no. 1 (February 2017): 271-79, <https://doi.org/10.1111/1745-9133.12272>.

⁵² In the risk assessment context, inter-rater reliability refers to the degree of agreement between distinct raters applying an assessment tool. A high IRR means raters apply the tool in the same manner as others; in other words, a high IRR means a particular defendant would receive the same score regardless of who conducted the assessment. A low IRR, in turn, would indicate raters may score the same defendant differently. Grant Duwe and Michael Rocque, "Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment (ROI): Recidivism Risk Assessment," *Criminology & Public Policy* 16, no. 1 (February 2017): 235-69, <https://doi.org/10.1111/1745-9133.12270>.

⁵³ Shawn Bushway and Jeffrey Smith, "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us," *Journal of Quantitative Criminology* 23, no. 4 (December 1, 2007): 377-87, <https://doi.org/10.1007/s10940-007-9035-1>.

⁵⁴ Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (2016): 671-732, <https://doi.org/10.15779/z38bg31>.

⁵⁵ *Ewert v. Canada*, 2018 SCC 30. Note, however, that other tools—such as the Ontario Domestic Assault Risk Assessment Tool ("ODARA")—are in widespread

deleterious effects on the rehabilitation of offenders from minority communities by impacting their access to cultural programming and their opportunities for parole, among other things.⁵⁶

The answer to the question of whether earlier generations of risk assessment tools have a positive or negative impact on the rights of criminal defendants and convicts to life, liberty, and security of person⁵⁷ is unclear. On one hand, they may represent an improvement over the situation where judges had essentially unfettered discretion regarding bail and sentencing decisions. On the other hand, the possibility of negative impacts exists due to the potential for the misclassification of some number of defendants as “high risk,” which results in their being sentenced more harshly than they otherwise would, or should, have been. Such tools also adversely impact criminal defendants’ rights to a fair public trial, to a defense, and to an appeal,⁵⁸ because their predictions are not subject to meaningful review by courts. Not only do courts lack the institutional capacity to review the operation of such tools, but the objective veneer that coats the outputs of these tools obscures the subjective determinations that are baked into them.

Furthermore, these tools raise fundamental questions as to whether it is fair to treat a particular individual more harshly simply because they share characteristics with others who have reoffended. This is a particularly serious difficulty when it comes to individuals who are classified as “high risk” yet for whatever reason do not reoffend. While statistical techniques can determine with a high degree of accuracy the characteristics of individuals in a population who are likely to behave in a certain way, they cannot generate accurate predictions as to how any particular individual in that population will behave. This raises some truly vexing legal, moral, and philosophical questions that are common to all the case studies that follow.

5.1.2. AI-Generated Risk Assessments

In recent years, criminal justice systems in many different countries have begun to use algorithmic risk assessment tools. All such tools automate the analysis of whatever data has been inputted into the

use in Canada and have been adopted by courts in several provinces and territories. For examples of courts relying on these tools, see *R v. Beharri*, 2015 ONSC 5900; *R v. Primmer* 2017 ONSC 2953; *R v. Sassie*, 2015 NWTCA 7; *R. v. Robertson*, 2006 ABPC 88.

⁵⁶ *Ewert v. Canada*, 2018 SCC 30.

⁵⁷ UDHR art. 3.

⁵⁸ UDHR arts. 10 and 11(1); ICCPR art. 14(5).

system. Most of these tools still rely on manually-inputted data from questionnaires similar to those that were part and parcel of the last generation of risk-assessment tools, while newer tools are fully automated and rely on information that already exists in various government databases.⁵⁹

Full automation improves the predictive accuracy and validity of risk assessment tools because the software interprets every piece of data consistently.⁶⁰ Automation also obviates the need for manual data collection, entry, and scoring, which carries with it the possibility of improving the accuracy of these systems by, for example, allowing additional variables to be considered.⁶¹

Beyond full automation, the latest generation of risk assessment tools leverages machine learning techniques to continually rebalance risk factors in response to new inputs. In theory, the predictive power and accuracy of such systems should improve over time. This was the finding of a proof-of-concept study in New York City, where researchers used machine learning techniques to determine which criminal defendants should receive bail.⁶² The study's results suggest that New York could reduce the number of people held in pretrial detention by 40% without any corresponding increase in the crime rate. Alternately, the city could reduce its crime rate by 25% by incarcerating the same number of people, but changing the criteria for who gets bail. In so doing, the number of African-Americans and Hispanics housed in the city's jails would be significantly reduced, with concomitant positive effects on the right to equality and non-discrimination.⁶³

For all of these potential positives, the single most widely-used algorithmic risk assessment system in the United States has been accused of perpetuating racial bias. An investigation by ProPublica found that COMPAS, a proprietary risk-assessment system that certain U.S. state courts use in making bail and sentencing

⁵⁹ The Minnesota Screening Tool Assessing Recidivism Risk 2.0 ("MnSTARR 2.0") under development by the government of the U.S. State of Minnesota is a leading example of a fully-automated risk assessment tool in the criminal justice context. Kenneth C. Land, "Automating Recidivism Risk Assessment: Should We Stay or Should We Go?," *Criminology & Public Policy* 16, no. 1 (February 2017): 231–33, <https://doi.org/10.1111/1745-9133.12271>.

⁶⁰ Ibid.

⁶¹ According to Barocas and Selbst, one source of bias is inaccuracies in the selected features. Additional features should, in theory, allow for more accurate generalizations to be developed. Barocas and Selbst, "Big Data's Disparate Impact."

⁶² Jon Kleinberg et al., "Human Decisions and Machine Predictions" (Cambridge, MA: National Bureau of Economic Research, February 2017), <https://doi.org/10.3386/w23180>.

⁶³ UDHR art. 2.

decisions, misclassified African-American offenders as “high-risk” at twice the rate of Caucasians, even though the system had nearly the same accuracy rate (63% vs. 59%) in predicting when individuals from both racial groups would reoffend.⁶⁴ In other words, COMPAS classified 45% of those African-American convicts who ultimately did not reoffend as “high risk,” as compared to just 23% for similarly-situated Caucasians. Questions have been raised about the accuracy and the methodological validity of the ProPublica report,⁶⁵ but more fundamentally, an important paper published in the aftermath of the COMPAS controversy suggests that it may be well-nigh impossible to design algorithms that treat individuals belonging to different groups equally fairly across multiple different dimensions of fairness.⁶⁶ Assuming, however, that the issues ProPublica identified with COMPAS are well-founded and are true of other risk assessment algorithms, then there is a substantial risk that the rights of minority groups to equality and non-discrimination will be adversely affected by such tools.⁶⁷

Furthermore, there is a serious issue relating to the existence of systematic patterns of bias against minorities in the data being used to train these algorithmic risk-assessment tools, arising from the disproportionate police scrutiny that minority community members receive. Consequently, minority communities are over-represented in the training data, which results in variables that are close proxies for race being over-weighted by these algorithms in assessing the risk that any particular individual poses.⁶⁸ This, too, raises concerns about algorithmic risk assessment tools having negative impacts on the rights of minority groups to equality and non-discrimination.⁶⁹

There are also issues that arise from the development of these risk assessment tools by private companies who, for commercial reasons, guard their algorithms and the data that is used to train them as trade secrets.⁷⁰ The secrecy that often surrounds the operation of

⁶⁴ Jeff Larson and Julia Angwin, “How We Analyzed the COMPAS Recidivism Algorithm,” *ProPublica*, May 23, 2016, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

⁶⁵ For example, Anthony W Flores, Kristin Bechtel, and Christopher T. Lowenkamp, “False Positives, False Negatives, and False Analyses: A Rejoinder to ‘Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.’,” *Federal Probation* 80, no. 2 (2016): 9.

⁶⁶ Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, “Inherent Trade-Offs in the Fair Determination of Risk Scores,” *ArXiv:1609.05807 [Cs, Stat]*, September 19, 2016, <http://arxiv.org/abs/1609.05807>.

⁶⁷ UDHR art. 2.

⁶⁸ Barocas and Selbst, “Big Data’s Disparate Impact.”

⁶⁹ UDHR art. 2.

⁷⁰ Rebecca Wexler, “Life, Liberty, and Trade Secrets,” *Stanford Law Review* 70, no. 5 (2018): 1343.

these risk assessment tools can have adverse impacts on the rights of criminal defendants to defend themselves against criminal charges⁷¹ and to appeal a conviction.⁷² The situation is further complicated when risk assessment algorithms rely upon machine learning techniques to adapt their performance over time, as the results generated by such techniques are oftentimes neither reproducible nor explainable in any meaningful way.

5.1.3. *Summary of Impacts*

The current generation of automated risk-assessment tools has the potential to positively impact the rights of “low-risk” criminal defendants and offenders to life, liberty, and security of the person.⁷³ If indeed such tools are more accurate than humans at predicting the risk of recidivism,⁷⁴ low-risk offenders will end up being incarcerated at a lower rate and for shorter periods of time than under the status quo. Members of society at large will also be more secure in the enjoyment of their right to security of the person should these tools result in a lower rate of crime.

It is hard to know, however, whether the current generation of automated risk assessment tools is having a negative or positive impact on the equality and non-discrimination rights of criminal defendants from groups that have historically been discriminated against, such as ethnic minorities and the mentally ill.⁷⁵ While the existence of systemic biases in the training data may result in the automation of existing social biases against individuals from these groups, the results of the New York City proof-of-concept study suggest that such systems may nevertheless ameliorate the over-representation of individuals from these groups in jail and prison populations.

Finally, in view of the inscrutability of the latest generation of automated risk assessment tools, and the secrecy surrounding these tools when they are developed by the private sector, we believe that these tools are likely to adversely impact the rights of criminal defendants to a fair and public hearing before an independent and

⁷¹ UDHR art. 11(1).

⁷² ICCPR art. 14(5).

⁷³ UDHR art. 2.

⁷⁴ This assumption has been questioned. Julia Dressel and Hany Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism,” *Science Advances* 4, no. 1 (January 1, 2018), <https://doi.org/10.1126/sciadv.aao5580>.

⁷⁵ UDHR art. 3.

impartial tribunal,⁷⁶ and to enjoy all of the guarantees needed for their defense.⁷⁷

⁷⁶ UDHR art. 10.

⁷⁷ UDHR art. 11(1). Relatedly, the right of criminal convicts under ICCPR art. 14(5) is similarly impacted.

5.2. Access to the Financial System: Credit Scores

Access to financial services such as banking and lending are an important means of promoting social and economic well-being. Access to credit in particular can help disadvantaged and marginalized individuals better enjoy their economic, social, and cultural rights by, for example, providing them with the means to pursue higher education,⁷⁸ access health care,⁷⁹ purchase property,⁸⁰ or start a business through which they can be gainfully employed.⁸¹ In view of the role that credit can play in advancing the achievement of a wide range of human rights, the Nobel laureate Muhammad Yunus has suggested that access to credit itself ought to be considered a human right.⁸²

5.2.1. *Traditional Approach to Credit Scoring*

In deciding whether to extend someone credit, lenders have long sought to ascertain the prospective borrower's risk of defaulting on the debt. Such determinations have historically been of dubious accuracy and rife with the possibility of discrimination, as lenders based them on their personal impressions of the borrower coupled with references from the community.⁸³ Nor were these determinations improved much by the development of the first credit reports around the turn of the 20th century, which consisted of compilations of information about an individual's personal affairs that were subject to the discretionary review of the lender.⁸⁴

In the United States, the legislative efforts of the 1970s to outlaw discrimination in lending based on race, religion, gender, age, and other similar traits roughly coincided with the development of the first credit scores, which attempted to reduce all of the information

⁷⁸ UDHR art. 26.

⁷⁹ UDHR art. 25.

⁸⁰ UDHR art. 17.

⁸¹ UDHR art. 23.

⁸² Matt Wade, "Access to Credit a 'Human Right', Says the Father of Microfinance," *Sydney Morning Herald*, October 9, 2014, <https://www.smh.com.au/national/access-to-credit-a-human-right-says-the-father-of-microfinance-20141009-113j3x.html>. For more on the debate, see the following two resources: Marek Hudon, "Should Access to Credit Be a Right?," *Journal of Business Ethics* 84, no. 1 (2009): 17–28 and John Gershman and Jonathan Morduch, "Credit Is Not a Right," in *Microfinance, Rights and Global Justice*, ed. Tom Sorell and Luis Cabrera (Cambridge: Cambridge University Press, 2015), 14–26, <https://doi.org/10.1017/CBO9781316275634.002>.

⁸³ Matthew A. Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data," *Chicago-Kent Law Review* 93, no. 1 (March 9, 2018): 2–60.

⁸⁴ Sean Trainor, "The Long, Twisted History of Your Credit Score," *Time*, accessed June 10, 2018, <http://time.com/3961676/history-credit-scores/>.

contained in an individual's credit score into a simple, numerical indication of that person's credit-worthiness.⁸⁵ Different companies use different approaches to calculate credit scores. FICO Scores, which are used by 90% of lenders in the United States, are generated based on a combination of an individual's payment history, the amount that they owe, the age of their accounts, their sources of credit, and how much additional credit they have sought recently.⁸⁶

Despite their objective veneer, traditional credit scores suffer from several limitations that can adversely impact human rights. Since traditional credit scores rely on information gathered by credit bureaus about an individual's past financial history, oftentimes individuals with a "thin" credit file are given a credit score that is not indicative of their true risk of defaulting or are denied a credit score entirely.⁸⁷ Such "thin-file" borrowers tend to belong to marginalized groups such as minorities, young adults, immigrants, and recently-divorced women.⁸⁸ Since financial institutions are less likely to lend to individuals from these groups, even when in reality they are just as credit-worthy as "thick-file" applicants from other groups, the right to equality may be adversely impacted.⁸⁹

There are also issues relating to the fairness and accuracy of the data being fed into credit-scoring algorithms. In the United States at least, credit bureaus rely on "furnishers"—banks, utilities, and other businesses—to voluntarily report relevant information, such as on-time payments, debt balances, and the like.⁹⁰ In view of the legal obligations that attach to furnishers when they provide information to a credit bureau, such businesses are more likely to report adverse events (such as a missed payment or foreclosure) that negatively impact its own bottom line, as opposed to routine, unremarkable

⁸⁵ Those laws are the Fair Credit Reporting Act ("FCRA") of 1970, the Equal Credit Opportunity Act ("ECOA") of 1974 and the Community Reinvestment Act of 1977. Willy E. Rice, "Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950-1995," *San Diego Law Review* 33 (1996): 583-700.

⁸⁶ Rob Kaufman, "5 Factors That Determine a FICO Score," *myFICO* (blog), September 23, 2016, <https://blog.myfico.com/5-factors-determine-fico-score/>.

⁸⁷ Kenneth P. Brevoort, Philip Grimm, and Michelle Kambara, "Data Point: Credit Invisibles" (Consumer Financial Protection Bureau Office of Research, May 2015).

⁸⁸ Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data." ("In other words, credit invisibles are generally either too young to have established a credit history, or have never been welcomed into the traditional banking system. As such, [algorithmic credit scores] would especially benefit the young, the low-income, and minorities." (internal citations omitted)).

⁸⁹ UDHR art. 2. Brevoort, Grimm, and Kambara, "Data Point: Credit Invisibles."

⁹⁰ "Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003" (Federal Trade Commission, January 2015).

positive events (such as timely payments).⁹¹ Since individuals from minority communities suffer from adverse financial events (such as evictions) at a higher rate than would be predicted by their actual financial circumstances,⁹² there is a significant risk that the information used to generate credit scores is systematically biased against minority communities.

Furthermore, even if all relevant information (both positive and negative) is reported to a credit agency, there is no guarantee that the credit scoring algorithm will consider it. For example, the FICO score in wide use in the United States considers only mortgage and credit-card payment history, but not rental or bill payment history.⁹³ In view of the long legacy of discriminatory lending policies in the U.S. and of housing policies that make it much more likely that individuals from minority groups will rent rather than own a home,⁹⁴ these practices can have significant discriminatory impacts.⁹⁵

The growing use of credit scores beyond the lending context amplifies these effects. It is increasingly common for employers, landlords, and insurers to review an individual's credit score before offering them a job, renting them an apartment, or selling them insurance.⁹⁶ Employers may think that credit scores are a proxy for an applicant's integrity and responsibility, even though they have not been validated for that purpose.⁹⁷ Insurers may similarly view

⁹¹ Jocelyn Baird, "What Gets Reported to Your Credit Reports (and What Doesn't)?," *NextAdvisor* (blog), accessed June 20, 2018, <https://www.nextadvisor.com/blog/what-gets-reported-to-your-credit-reports/>.

⁹² Deena Greenberg, Carl Gershenson, and Matthew Desmond, "Discrimination in Evictions: Empirical Evidence and Legal Challenges," *Harvard Civil Rights* 51, no. 1 (2016): 44.

⁹³ Preeti Vissa, "How Credit Scores Disproportionately Hurt Communities of Color," *Huffington Post* (blog), December 15, 2010, https://www.huffingtonpost.com/preeti-vissa/credit-scores-and-the-for_b_797148.html; "How Credit History Impacts Your FICO® Score," *myFICO* (blog), accessed June 20, 2018, <http://www.myfico.com/credit-education/credit-payment-history>.

⁹⁴ Christopher E. Hebert et al., "Homeownership Gaps Among Low-Income and Minority Borrowers and Neighborhoods" (U.S. Department of Housing and Urban Development, March 2005); Sarah Ludwig, "Credit Scores in America Perpetuate Racial Injustice. Here's How," *The Guardian*, October 13, 2015, sec. Opinion, <http://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why>.

⁹⁵ UDHR art. 3.

⁹⁶ "Past Imperfect: How Credit Scores and Other Analytics 'Bake In' and Perpetuate Past Discrimination" (National Consumer Law Center, May 2016) ("Credit history is used as a gatekeeper for many important necessities – employment, housing (both rental and homeownership), insurance, and of course, affordable credit.").

⁹⁷ Gary Rivlin, "Employers Pull Applicants' Credit Reports," *The New York Times*, May 11, 2013, sec. Business Day,

those with poor credit scores as posing a higher actuarial risk because “recklessness” in paying down one’s debts shows the individual to be a reckless person in general.⁹⁸ Yet again, such practices pose a grave risk of perpetuating and amplifying age-old patterns of inequality and discrimination that bear little resemblance to reality.

5.2.2. *AI-Generated Credit Scores*

In recent years, lenders have begun to use artificial intelligence to more accurately assess whether a potential borrower is a good credit risk. Unlike conventional credit scoring algorithms, the AI-based approach treats “all data as credit data” and analyzes vast amounts of data from many sources.⁹⁹ The resulting AI-generated credit scores are better than traditional scores at addressing some kinds of situations, at the same time as they create new challenges of their own.

The volume of data that AI-based credit scoring systems collect and analyze is so staggering as to be concerning. ZestFinance, one of the leading companies in this field in the US, considers over 3,000 variables in deciding whether to offer someone credit¹⁰⁰—including whether the applicant tends to type in all-caps, which apparently is correlated with a higher risk of default.¹⁰¹ Lenddo, another American company in this space, examines an applicant’s entire digital footprint—including social media use, geolocation, website browsing habits, phone use history (including text and call logs), purchasing behavior, and more in deciding whether to extend them credit.¹⁰²

AI-generated credit scores have particularly significant applications in emerging markets, where almost everyone is a “thin-file” borrower. For example, the MyBucks Haraka app in use in India, the Philippines, and several Sub-Saharan countries uses data gleaned from an applicant’s mobile phone (call logs, geolocation

<https://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html>.

⁹⁸ “Past Imperfect: How Credit Scores and Other Analytics ‘Bake In’ and Perpetuate Past Discrimination.”

⁹⁹ James Rufus Koren, “Some Lenders Are Judging You on Much More than Finances,” *Los Angeles Times*, December 19, 2015,

<http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html>.

¹⁰⁰ “Zest Automated Machine Learning Data Sheet” (ZestFinance), accessed June 20, 2018, <https://www.zestfinance.com/hubfs/Underwriting/Zest-Automated-Machine-Learning-Data-Sheet.pdf?hsLang=en>.

¹⁰¹ Koren, “Some Lenders Are Judging You on Much More than Finances.”

¹⁰² “Credit Scoring: The LenddoScore Fact Sheet” (Lenddo), accessed June 20, 2018, https://www.lenddo.com/pdfs/Lenddo_FS_CreditScoring_201705.pdf.

information, and the like) and their social media accounts to generate an alternative credit score that partner banks can use to inform their lending decision.¹⁰³ This AI-based approach has the potential to help members of historically marginalized groups, such as women and ethnic minorities, gain access to credit in the developed and developing world alike,¹⁰⁴ thereby fostering financial inclusion and advancing the right to equality.¹⁰⁵

Early results suggest that these technologies are succeeding in fostering financial inclusion. ZestFinance claims that its AI-based technology allowed it to reduce its default rate to less than half of the prevailing industry average,¹⁰⁶ while Lenddo claims to have increased its approval rate by 15% while slashing defaults by 12%.¹⁰⁷ If these early results are accurate and generalizable, the positive impact on all of the economic, cultural, and social rights that access to credit enables would be very significant indeed.

Yet there are considerable risks to this new approach as well. One arises from the quality and accuracy of the data used to train these systems, as well as the fairness and accuracy of the data these systems use to decide upon a particular individual's application for credit. The issues are similar in nature to those affecting traditional credit scoring algorithms, but they are different in degree due to the vast number of data sources that AI-based algorithms take into consideration.

Another arises from the subjective decisions that programmers make on how to code and categorize the data that they feed into their seemingly-objective algorithms.¹⁰⁸ For example, ZestFinance translates certain continuous variables (such as the length of time one spends reading their website's terms and conditions) into

¹⁰³ Penny Crosman, "This Lender Is Using AI to Make Loans through Social Media," *American Banker*, December 8, 2017, <https://www.americanbanker.com/news/this-lender-is-using-ai-to-make-loans-through-social-media>.

¹⁰⁴ Brevoort, Grimm, and Kambara, "Data Point: Credit Invisibles.," Geri Stengel, "How One Woman Is Changing Business Lending In Africa," *Forbes*, January 14, 2015, <https://www.forbes.com/sites/geristengel/2015/01/14/how-one-woman-is-changing-business-lending-in-africa>.

¹⁰⁵ UDHR art. 2.

¹⁰⁶ John Lippert, "ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed out Deadbeats," *The Washington Post*, October 11, 2014, sec. Business, https://www.washingtonpost.com/business/zestfinance-issues-small-high-rate-loans-uses-big-data-to-weed-out-deadbeats/2014/10/10/e34986b6-4d71-11e4-aa5e-7153e466a02d_story.html.

¹⁰⁷ "Credit Scoring: The LenddoScore Fact Sheet."

¹⁰⁸ Mikella Hurley and Julius Adebayo, "Credit Scoring in the Era of Big Data," *Yale Journal of Law & Technology* 18, no. 1 (2016): 148–216.

categorical values (like 0, 1, or 2).¹⁰⁹ This is an inherently subjective process that can introduce explicit or implicit bias into the data and consequently into the results generated by the algorithm.

Furthermore, AI-generated scores may perpetuate existing patterns of discrimination through “network discrimination,”¹¹⁰ whereby individuals are penalized (or rewarded) based on the characteristics of others who are in their personal network. For example, if there are two individuals in an identical financial position, yet the first individual’s friends live in “rich” neighborhoods while the second’s friends live in “poor” neighborhoods, an algorithm may well determine the first to be a better credit risk than the second.¹¹¹ To the extent that such network factors correlate with invidious classifications such as those based on race and gender, the potential for discriminatory impacts is quite serious indeed.¹¹²

The use of AI in financial decision-making may even burden individuals’ freedom of opinion, expression, and association by chilling individuals from engaging in activities that they believe will negatively affect their credit score. This is not a mere theoretical possibility. In 2009, American Express reduced the credit limit of an African-American businessman because “[o]ther customers who have used their card at establishments where [he] recently shopped . . . ha[d] a poor repayment history.”¹¹³ Another lender in the U.S. reduced the credit limit of its customers who had incurred expenses at “marriage counselors, tire retreading and repair shops, bars and nightclubs, pool halls, pawn shops, massage parlors, and others.”¹¹⁴

An even more extreme example of this phenomenon is China’s incipient “social credit score” system, which generates a numerical index of an individual’s “trustworthiness” based on a vast array of data points, including social media data, arrest and infraction records, volunteer activity, city and neighborhood records, and

¹⁰⁹ Ibid.

¹¹⁰ danah boyd, Karen Levy, and Alice Marwick, “The Networked Nature of Algorithmic Discrimination,” in *Data and Discrimination: Collected Essays*, ed. Seeta Peña Gangadharan (New America, 2014), 53–57.

¹¹¹ Kaveh Waddell, “How Algorithms Can Bring Down Minorities’ Credit Scores,” *The Atlantic*, December 2, 2016, <https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/>.

¹¹² UDHR arts. 2 and 20.

¹¹³ Ron Lieber, “American Express Watched Where You Shopped,” *The New York Times*, January 30, 2009, sec. Your Money, <https://www.nytimes.com/2009/01/31/your-money/credit-and-debit-cards/31money.html>.

¹¹⁴ Ibid.

more.¹¹⁵ Those with high social credit scores enjoy benefits such as lower utility rates and more favorable borrowing conditions, while those with unfavorable scores might be unable to purchase airline or high speed rail tickets.¹¹⁶ These systems are being piloted in several communities, but a national roll-out of the “social credit score” system is expected by 2020. While anecdotal reports suggest that the “social credit scoring” system has curbed corruption and incentivized certain forms of good behavior, such as stopping for pedestrians at crosswalks,¹¹⁷ it is not hard to imagine how this system could chill a great deal of expressive and associative activity.¹¹⁸

5.2.3. *Summary of Impacts*

Compared to the status quo credit scoring algorithms, the introduction of AI into the lending process is likely to have an overall positive impact on the ability of objectively low-risk borrowers to access credit. This is likely to have positive impacts on the enjoyment by these individuals of the right to an adequate standard of living,¹¹⁹ the right to work,¹²⁰ and the right to education,¹²¹ as access to credit is a powerful enabler of these economic and social rights.

The introduction of AI into the lending process is also likely to have a positive impact on the right to equality and non-discrimination for some individuals, while adversely affecting it for others. On the positive side, the fact that AI-based algorithms consider a wide variety of data sources may improve the ability of well-qualified individuals from marginalized communities to access credit by overcoming the “thin-file” problem. On the other hand, the specter of “network discrimination” having a negative impact on the ability of members of these very same communities to borrow money cannot be discounted.

¹¹⁵ Mara Hvistendahl, “In China, a Three-Digit Score Could Dictate Your Place in Society,” *WIRED*, Dec. 14, 2017, <https://www.wired.com/story/age-of-social-credit/>.

¹¹⁶ Simina Mistreanu, “Life Inside China’s Social Credit Laboratory,” *Foreign Policy*, April 3, 2018, <https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/>.

¹¹⁷ *Ibid.*

¹¹⁸ UDHR arts. 19 and 20. In this context, the U.N. Human Rights Committee has noted that it is “impermissible” for states to engage in conduct that create “chilling effects that may unduly restrict the exercise of freedom of expression....” United Nations Human Rights Committee, General Comment 34 (ICCPR Art. 19: Freedoms of opinion and expression) (2011), U.N. Doc. CCPR/C/GC/34, p. 12.

¹¹⁹ UDHR art. 17.

¹²⁰ UDHR art. 23.

¹²¹ UDHR art. 26.

Finally, it is likely that AI-based decision-making algorithms in the financial sector will adversely impact the freedoms of opinion, expression, and association. In an era where “all data is credit data,” individuals may feel chilled from expressing certain points of view or associating with others, out of fear that an algorithm may use their behavior against them in the financial context.

5.3. Healthcare: Diagnostics

In the course of the last century, modern medicine has produced astonishing improvements in the length and the quality of the lives of all those who can access it. Not only does the ICESCR recognize “the right of everyone to the enjoyment of the highest attainable standard of physical and mental health,”¹²² but good health is arguably a necessary condition for each and every one of us to enjoy the full range of human rights that we are guaranteed by law.

Recent advances in health outcomes are attributable to improvements in the three pillars of healthcare: prevention, diagnosis, and treatment. AI has applications across all three pillars, but its greatest impact to date has been on improving the accuracy of medical diagnosis.

5.3.1. *Traditional Approach to Diagnostics*

Physicians use a wide range of approaches to diagnose disease. Perhaps the simplest and most widespread is to identify the patient’s symptoms and correlate them to conditions or diseases that are characterized by the same pattern of symptoms.¹²³ The same basic approach can be applied to interpreting the results of diagnostic tests: a radiologist reviewing MRI imagery or a pathologist analyzing a biopsy sample compare what they are seeing to what they have learned in order to make a diagnosis.

Needless to say, it takes years of training and years more of experience to develop the knowledge and mastery required to accurately diagnose the wide range of maladies that afflict our species. To simplify matters, physicians often rely on “diagnostic criteria” in determining what ails someone. These are essentially statistically-validated rules of thumb that can be used to rule in or rule out a particular condition.¹²⁴ By contrast, experts in particular diseases engage in *gestalt* pattern recognition to recognize the

¹²² ICESCR art. 12.

¹²³ Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*, ed. Erin P. Balogh, Bryan T. Miller, and John R. Ball (Washington, D.C.: National Academies Press, 2015), <https://doi.org/10.17226/21794>.

¹²⁴ The “Centor Criteria” for diagnosing strep throat in adults is an example of this. Since roughly 50% of patients who have all five of the criteria (cough, tonsillar exudates, swollen lymphatic nodes, fever, neither young nor old) will turn out to have strep throat, the Centor criteria presents a quick and easy way to rule in or rule out strep throat as a possibility in a patient presenting with these symptoms. Robert M. Centor et al., “The Diagnosis of Strep Throat in Adults in the Emergency Room,” *Medical Decision Making* 1, no. 3 (August 1981): 239–46, <https://doi.org/10.1177/0272989X8100100304>.

characteristic indicators of a particular disease in a sea of information.¹²⁵

Unfortunately, errors in diagnostics are extremely common, and they can have life-and-death consequences. One recent study found that 5% of patients in the U.S. are misdiagnosed every year,¹²⁶ while another found that misdiagnosis is the cause of 10% of patient deaths.¹²⁷ The challenge for physicians is growing as the number of diagnostic tests and procedures multiplies with the advance of medical science. Since each of these procedures has its own unique operating parameters and error rates,¹²⁸ it is becoming increasingly difficult for the average medical practitioner to choose the right test for their patient—or to even refer their patients to the right sub-specialists in view of their symptoms.¹²⁹

5.3.2. *AI-Assisted Diagnostics*

Medical diagnostics is one of the fields in which AI-based technologies went into widespread use. Efforts began in the 1970s to start codifying the knowledge of human diagnostic experts into automated “expert systems.”¹³⁰ These systems, which are known as “diagnostic decision support systems” and are used in many healthcare settings today, require the human clinician to answer a series of questions about the patient’s condition that help to rule in or rule out certain specific diagnoses.

In the last several years, systems based on machine learning or deep learning have begun to be developed to facilitate and automate the diagnosis of illness across a range of medical specialties. Few of these technologies are currently in use, though early results suggest

¹²⁵ John P. Langlois, “Making a Diagnosis,” in *Fundamentals of Clinical Practice*, ed. Mark B Mengel, Warren Lee Holleman, and Scott A Fields, 2nd ed. (New York: Kluwer Academic/Plenum Publishers, 2005).

¹²⁶ Hardeep Singh, Ashley N D Meyer, and Eric J Thomas, “The Frequency of Diagnostic Errors in Outpatient Care: Estimations from Three Large Observational Studies Involving US Adult Populations,” *BMJ Quality & Safety* 23, no. 9 (September 2014): 727–31, <https://doi.org/10.1136/bmjqs-2013-002627>.

¹²⁷ Ibid.

¹²⁸ Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*.

¹²⁹ J. Hickner et al., “Primary Care Physicians’ Challenges in Ordering Clinical Laboratory Tests and Interpreting Results,” *The Journal of the American Board of Family Medicine* 27, no. 2 (March 1, 2014): 268–74, <https://doi.org/10.3122/jabfm.2014.02.130104>.

¹³⁰ MYCIN was one of the pioneer expert systems developed at Stanford starting in 1972. Nancy McCauley and Mohammad Ala, “The Use of Expert Systems in the Healthcare Industry,” *Information & Management* 22, no. 4 (April 1, 1992): 227–35, [https://doi.org/10.1016/0378-7206\(92\)90025-B](https://doi.org/10.1016/0378-7206(92)90025-B).

that they have great promise in improving the accuracy of medical diagnosis.

For example, an AI-powered image recognition system was able to detect cancerous skin lesions correctly 72% of the time, whereas human dermatologists correctly diagnosed the cancers 66% of the time.¹³¹ There are also anecdotes about AI-powered diagnostic systems quickly solving intractable mysteries. For example, a diagnostic system powered by IBM's Watson was able to diagnose a patient as possessing a rare form of leukemia within 10 minutes, even though her symptoms had stumped experts for several months.¹³² The system did so by comparing information in the patient's medical records with over 20 million oncology records held by the University of Tokyo. To be sure, physicians currently outperform AI systems on a wide variety of diagnostic tasks—from microscopy to general diagnosis. Yet it is impressive that AI systems rival or outperform human experts in diagnosing conditions ranging from brain cancer to autism and Alzheimer's disease.¹³³

AI-based diagnostic systems also have the potential to provide greater access to specialist-level treatment than is currently possible. One of the few AI-based diagnostic systems to be approved for clinical use by the U.S. Food and Drug Administration is able to detect and diagnose diabetic retinopathy (a disorder affecting the vision of individuals suffering from diabetes) autonomously.¹³⁴ Whereas this condition was previously one that could only be diagnosed by a specialist, AI makes it possible for anyone trained in using the machinery to do so.

5.3.3. *Summary of Impacts*

AI-based diagnostic systems, especially the latest generation of systems that leverage artificial intelligence, are very likely to positively impact the right each of us enjoys to the highest attainable standard of health.¹³⁵ Not only do AI-based diagnostic systems

¹³¹ Siddhartha Mukherjee, "A.I. Versus M.D.," *The New Yorker*, March 27, 2017, <https://www.newyorker.com/magazine/2017/04/03/ai-versus-md>.

¹³² James Billington, "IBM's Watson Cracks Medical Mystery with Life-Saving Diagnosis for Patient Who Baffled Doctors," *International Business Times UK*, August 8, 2016, <https://www.ibtimes.co.uk/ibms-watson-cracks-medical-mystery-life-saving-diagnosis-patient-who-baffled-doctors-1574963>.

¹³³ "AI vs Doctors," *IEEE Spectrum: Technology, Engineering, and Science News*, September 26, 2017, <https://spectrum.ieee.org/static/ai-vs-doctors>.

¹³⁴ Angela Chen, "AI Software That Helps Doctors Diagnose like Specialists Is Approved by FDA," *The Verge*, April 11, 2018, <https://www.theverge.com/2018/4/11/17224984/artificial-intelligence-idxdr-fda-eye-disease-diabetic-rethinopathy>.

¹³⁵ ICESCR art. 12.

appear to meet or exceed the performance of human experts in diagnosing disease, they have the potential to be much more accessible than specialized human experts, who require years of training and experience to rival the accuracy of an AI.

It is significant that in recognizing the right that each of us possesses “to a standard of living adequate for the health and well-being of himself and of his family,” Article 25 of the UDHR links access to medical care to the basic requisites of life, such as food, clothing, and housing. In view of this link between good health, access to health care, and the full range of economic, social, and cultural rights that each of us enjoys, the use of AI in medical diagnostics is likely to have positive impacts on the right of each of us to work and in so doing, ensure ourselves an existence worthy of human dignity.¹³⁶ Likewise, the better health outcomes that AI-based diagnostics are likely to produce will positively impact the enjoyment of the right to education by those who would otherwise be excluded by reasons of illness.¹³⁷

As with many other automated technologies, there is the possibility that AI-based diagnostic technologies will cause employment losses in the medical field. While the right to work does not entail the right to work in any particular position, occupation, or field, the state obligation to protect the right to work and progressively adopt measures to realize full employment could be burdened by the widespread adoption of AI-based technologies that displace workers.¹³⁸ Indeed, there is already evidence that the impressive performance of AI-based diagnostic systems is leading medical students to shy away from entering certain specialty fields, such as radiology, where AI systems routinely outperform humans.¹³⁹

Furthermore, the gathering of personal data necessary to create AI-powered tools creates particularly acute privacy risks in the healthcare context. In order to train the algorithms, healthcare providers must collect a vast range of intensely personal health and genetic data. The scope for the misuse of this data is vast—especially since an individual’s genetic and health characteristics are often immutable—with potential implications for privacy,¹⁴⁰ dignitary

¹³⁶ UDHR art. 23.

¹³⁷ UDHR art. 26.

¹³⁸ UN Committee on Economic, Social and Cultural Rights (“CESCR”), General Comment No. 18: The Right to Work (Art. 6 of the Covenant), 6 February 2006, UN Doc. E/C.12/GC/18.

¹³⁹ Thomas H. Davenport and D. O. Keith J. Dreyer, “AI Will Change Radiology, but It Won’t Replace Radiologists,” *Harvard Business Review*, March 27, 2018, <https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists>.

¹⁴⁰ UDHR art. 12.

rights,¹⁴¹ freedom from discrimination,¹⁴² and fair criminal procedure.¹⁴³ For example, such data could be used to deny a person health coverage on the basis of genetic factors that are beyond their control.¹⁴⁴ Or such data might be appropriated by the government for law enforcement purposes, as in the recent case from California of a 1970s-era serial killer who was identified based on the statistical analysis of DNA samples that his distant relatives submitted to a family ancestry website.¹⁴⁵

Going further, one can argue that the fundamental right to life may be positively impacted by the introduction of AI diagnostic systems, which hold the promise of not only reducing the rate of diagnostic errors, but making high quality diagnostic services cheaper or more widely available. Although the right to life is generally viewed as a protection against the arbitrary deprivation of life by the state, the Supreme Court of Canada has ruled that inadequate access to medical care can result in deprivations of the right to life.¹⁴⁶ Correspondingly, improvements in the availability of high-quality medical services can be viewed as enhancing the right to life.

¹⁴¹ UDHR art. 1.

¹⁴² UDHR art. 7.

¹⁴³ UDHR art. 10. In particular, it raises questions of self-incrimination, as protected by ICCPR art. 14(3)(g).

¹⁴⁴ For discussion on attempts to regulate genetic discrimination in the United States, see Louise Slaughter, “Genetic Information Non-Discrimination Act”, *Harvard J. on Legislation* 50, no. 1 (2013): 41.

¹⁴⁵ Thomas Fuller, “How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect,” *The New York Times*, April 26, 2018, sec. United States, <https://www.nytimes.com/2018/04/26/us/golden-state-killer.html>.

¹⁴⁶ *Chaoulli v. Quebec (Attorney General)*, 2005 SCC 35.

5.4. Online Content Moderation: Standards Enforcement

The sheer amount of information that is available online has mostly been a blessing for humanity, though sometimes it can be a curse. On the one hand, never has so much information about so many different topics been available to most anyone, anywhere, who is fortunate enough to have an internet connection. On the other hand, the dark side of humanity is also plain to view on the internet. By virtue of the volume of what is available online, there is a substantial amount of content that is racist, sexist, gruesome, or harmful in other ways—such as by fomenting violence against identifiable groups or targeting individuals for bullying or harassment.

Some of the objectionable content online is subject to regulation by governments in conformity with international human rights law. Article 19(3) of the ICCPR recognizes that the right to free expression may be subject to certain exceptions provided by law that are necessary to protect the rights and reputations of others, or to protect national security, public order, public health, and morals. Moreover, Article 20 of the ICCPR expressly requires states to prohibit “propaganda for war” and the advocacy of “national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.”

Beyond expression that is unlawful and therefore subject to bona fide government regulation, there is also the problem of content that may be lawful but is nonetheless undesirable—either because it is not being posted in an appropriate place online, or because it violates the standards of an established online community. This is precisely the context in which the private companies that operate the internet platforms that house so much of the world’s online content promulgate standards as to what is acceptable and what is not, and engage in the “virtue of moderating”¹⁴⁷ materials that are inconsistent with those standards.

5.4.1. *Traditional Approach to Content Standards Enforcement*

The setting and enforcement of content standards by private companies is a controversial topic. Some liken it to a form of censorship due to the burdens it places on the rights to free

¹⁴⁷ James Grimmelman, “The Virtues of Moderation,” *Yale Journal of Law & Technology* 17, no. 1 (2015): 68.

expression, thought, and association.¹⁴⁸ Indeed, some commentators have noted that the largest online platforms, such as Facebook and Google, exercise more power over our right to free expression than any court, king, or president ever has¹⁴⁹—in view of the very significant percentage of human discourse that occurs within the boundaries of these “walled gardens.”¹⁵⁰

By the same token, however, the failure of companies to adequately deal with online content that is harmful to others places its own burden on human rights. Companies therefore face the unenviable challenge of balancing between different rights belonging to different rights-holders, all the while remaining mindful of their responsibility to respect human rights in so doing.¹⁵¹

To discharge this difficult task with dispatch while avoiding discriminatory or speech-chilling outcomes, companies communicate their content guidelines to the public on their websites, at the same time as they have developed detailed internal guidance documents for their employees on when particular forms of content are subject to removal.¹⁵² Ideally, both the external and internal guidelines will be informed by the principles of international human rights law, both with regards to their substantive content and the process that they outline.¹⁵³

Until recently, the primary means by which questionable content was brought to the attention of a company was through the efforts of individual users of the platform, who flagged content as unlawful or inappropriate. Human reviewers working for the company then assess the content against the guidelines and determine whether it

¹⁴⁸ UDHR arts. 18, 19 and 20.

¹⁴⁹ Jeffrey Rosen, “The Deciders: Facebook, Google, and the Future of Privacy and Free Speech,” in *Constitution 3.0: Freedom and Technological Change*, ed. Jeffrey Rosen and Benjamin Wittes (Brookings Institution Press, 2013).

¹⁵⁰ Jonathan Zittrain, *The Future of the Internet and How to Stop It* (New Haven, [Conn.]: Yale University Press, 2008).

¹⁵¹ Perhaps counter-intuitively, content moderation may be among the AI applications where the distributive effects of these technologies are most apparent. Absent government-established policies, companies will be tasked with choosing which rights and rights-holders are prioritized over others.

¹⁵² Alexis C. Madrigal, “Inside Facebook’s Fast-Growing Content-Moderation Effort,” *The Atlantic*, February 7, 2018, <https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/>.

¹⁵³ Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) (by David Kaye), available at http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

should stay up or come down.¹⁵⁴ In view of the massive volume of content that the leading internet platforms host, these companies now each employ thousands if not tens of thousands of individuals whose sole job is to determine the fate of content that has been flagged.

In recent years, companies have been coming under increased scrutiny both as to the substance of the standards they employ in adjudging content, and also for their decisions in particular cases. For example, Facebook's broader policy against the display of nudity on its platform drew controversy when it removed images of breast-feeding women and the infamous "napalm girl" photograph from the Vietnam War from its platform.¹⁵⁵ Facebook ultimately relented in the face of public pressure in both incidents, but that too raises further questions about the consistency of its application of policies that burden the right to free expression.

There are also growing concerns that company policies on acceptable content may discriminate against certain viewpoints or perspectives, usually in a manner that favors the powerful over the marginalized.¹⁵⁶ For example, Facebook permitted a U.S. Congressman to state his view that all radicalized Muslims should be "hunted" or "killed," whereas it banned activists associated with the Black Lives Matter movement from stating that "all white people are racist."¹⁵⁷ While these anecdotes are not necessarily indicative of a larger pattern of bias or discrimination, they do raise troubling questions about how well these companies are meeting their responsibility to respect the full range of human rights in this important operational area.

Finally, there is also the issue of how companies are coming under growing pressure from governments to comply with their local laws on a global basis,¹⁵⁸ even when these laws are inconsistent with

¹⁵⁴ Brittan Heller, "What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech," *WIRED*, May 2, 2018, <https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/>.

¹⁵⁵ Kate Klonik, "Facebook Erred by Taking down the 'Napalm Girl' Photo. What Happens Next?," *Slate*, September 12, 2016, http://www.slate.com/articles/technology/future_tense/2016/09/facebook_erred_by_taking_down_the_napalm_girl_photo_what_happens_next.html.

¹⁵⁶ Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica*, June 28, 2017, <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

¹⁵⁷ *Ibid.*

¹⁵⁸ Alicia Solow-Niederman et al., "Here, There, or Everywhere?," Berkman Klein Center Working Paper, March 2017,

international guarantees of free expression, the right to association, and other human rights.¹⁵⁹ To date, companies have attempted to blunt these efforts by complying with local laws on a local basis, such as by prophylactically blocking content that is unlawful in a particular jurisdiction while leaving it available everywhere else. A string of recent court decisions has called into question the continuing viability of this technique, however.¹⁶⁰

Meanwhile, other governments are even beginning to use company terms of service as a way to act against content that is lawful under domestic and international law, yet undesirable in the eyes of public policymakers. Some in the human rights community have expressed concerns that this activity of “referring” content for removal constitutes an end run around well-established judicial procedures for removing unlawful content.¹⁶¹

5.4.2. *AI-Assisted Content Standards Enforcement*

As the volume of online content in need of moderation grows inexorably and exponentially, the major online platforms are making significant investments in developing AI systems to automate this task. One major impetus for so doing are recently-enacted laws that require companies to promptly remove content that violates national laws, at the risk of facing substantial penalties for noncompliance.¹⁶² These technologies are still in their infancy, and most simply work to identify potentially problematic content for a human reviewer to evaluate. That being said, fully automated content removal systems have been used against content that is suspected of violating copyright for a number of years,¹⁶³ and there are indications that

<http://blogs.harvard.edu/cyberlawclinic/files/2017/03/Here-There-or-Everywhere-2017-03-27.pdf>.

¹⁵⁹ For example, see Jens-Henrik Jeppesen & Laura Blanco, “European Policymakers Continue Problematic Crackdown on Undesirable Online Speech,” Center for Democracy and Technology (blog), Jan. 18, 2018, <https://cdt.org/blog/european-policymakers-continue-problematic-crackdown-on-undesirable-online-speech/>.

¹⁶⁰ For example, *Google Inc. v. Equustek Solutions Inc.* 2017 SCC 34.

¹⁶¹ Jens-Henrik Jeppesen, “First Report on the EU Hate Speech Code of Conduct shows need for transparency, judicial oversight, and appeals,” Center for Democracy and Technology (blog), Dec. 12, 2016, <https://cdt.org/blog/first-report-eu-hate-speech-code-of-conduct-shows-need-transparency-judicial-oversight-appeals/>.

¹⁶² Yascha Mounk, “Verboten,” *The New Republic*, April 3, 2018, <https://newrepublic.com/article/147364/verboten-germany-law-stopping-hate-speech-facebook-twitter>.

¹⁶³ “How Content ID Works - YouTube Help,” accessed June 21, 2018, <https://support.google.com/youtube/answer/2797370?hl=en>.

some internet platforms are employing fully automated content review and removal systems for at least some purposes.¹⁶⁴

The current generation of AI-based content review and removal systems is built on natural language processing (“NLP”) technology. As things stand right now, NLP technologies are domain-specific: that is to say, they were only built to identify the particular types of content on which they were trained and nothing else.¹⁶⁵ Hence an NLP system that is trained to detect say, racist speech, is incapable of detecting violent content. What is more, even within a particular domain, NLP technologies are not sophisticated enough to understand all of the nuances of human speech. A system that is able to detect racist content in a blog post might not accurately identify such content in a tweet, which results in these technologies having a very substantial error rate. This has led skeptics, including Facebook CEO Mark Zuckerberg, to conclude that AI systems are not yet sophisticated enough to replace human reviewers.¹⁶⁶

This, however, does not render AI technologies useless. The speed at which they can sift through content makes them a powerful tool to assist, rather than to replace, human reviewers by identifying content that appears to be suspect.¹⁶⁷ AI systems can also be used to study the evolution of hate speech to spot emerging trends, as the Anti-Defamation League is currently doing with its Online Hate Index.¹⁶⁸

5.4.3. *Summary of Impacts*

Due to the higher error rates of existing AI-based content flagging systems as compared to human reviewers,¹⁶⁹ the use of these systems to automatically remove content that is suspected to violate the law or an online platform’s community standards is likely to have a

¹⁶⁴ Lizzie Dearden, “New Technology Can Detect ISIS Videos before They Are Uploaded,” *The Independent*, February 12, 2018, <http://www.independent.co.uk/news/uk/home-news/isis-videos-artificial-intelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html>.

¹⁶⁵ Natasha Duarte, Emma Llanso, and Anna Loup, “Mixed Messages? The Limits of Automated Social Media Content Analysis,” Center for Democracy & Technology (blog), November 2017, <https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/>.

¹⁶⁶ Heller, “What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech.”

¹⁶⁷ Susan Wojcicki, “Expanding our work against abuse of our platform,” YouTube Official Blog, Dec. 4, 2017, <https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html>.

¹⁶⁸ “The Online Hate Index,” Anti-Defamation League, accessed June 18, 2018, <https://www.adl.org/resources/reports/the-online-hate-index>.

¹⁶⁹ Duarte, Llanso, and Loup, “Mixed Messages?”

negative impact on the rights to free expression, opinion, and information.¹⁷⁰ This is because such systems are likely to remove a significant volume of content that is lawful and consistent with the platform's own community standards. Such errors would deprive individuals of the opportunity to express themselves, and their audience from viewing the opinions those individuals have expressed, with few opportunities for recourse.¹⁷¹ That said, future developments in AI could well have a positive impact on these rights if they result in a lower error rate than the systems and procedures that are currently in use.

In their current state, these systems have the potential to positively impact the rights to life, liberty, and security of person¹⁷² by improving the detection and removal of content that incites terrorism or hatred or violence against vulnerable populations. For example, automated techniques have been quite effective at detecting child pornography with a low error rate, and some forms of terrorist content exhibit consistent patterns that facilitate their detection by training a machine learning algorithm, by contrast to the fast-evolving and always-subjective nature of hate speech.¹⁷³

The net impact of these systems on the right to be free from discrimination¹⁷⁴ is indeterminate. As with free expression, AI systems could be trained to avoid the biases exhibited by human reviewers, but on the other hand there is a considerable risk that machine learning techniques will result in the replication and scaling of existing human patterns of bias into new automated content review systems.¹⁷⁵

One additional right that merits discussion in this context is that of the individuals employed in content review and moderation

¹⁷⁰ UDHR art. 19.

¹⁷¹ For example, it is only in April of this year that Facebook announced that it was creating a system by which its users could appeal content removal decisions that they believe to be in error. Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Facebook (official blog), April 24, 2018, <https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/>.

¹⁷² UDHR art. 3.

¹⁷³ Hanna Kozłowska, "Facebook Is Revealing Data on How Good It Is at Moderating Content, but the Numbers Have Holes," *Quartz*, May 18, 2018, <https://qz.com/1277729/facebook-is-revealing-data-on-how-good-it-is-at-moderating-content-but-the-numbers-dont-say-much/>.

¹⁷⁴ UDHR art. 2.

¹⁷⁵ Reuben Binns et al., "Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation," *ArXiv:1707.01477 [Cs]* 10540 (2017): 405–15, <https://doi.org/10.1007/978-3-319-67256-4>.

positions to just and favorable conditions of work.¹⁷⁶ The psychological toll that the frontline work of content review and moderation takes is considerable, as these individuals are exposed to the very worst of humanity day in and day out—from child pornography to gruesome acts of violence. Content reviewers are disproportionately female, but reviewers of all genders suffer from depression, burnout, anxiety, sleep difficulties, and even from post-traumatic stress disorder at extraordinary rates.¹⁷⁷ Using AI to lessen the psychological burden associated with this work could well have positive human rights impacts on a group of individuals who are often forgotten in conversations about how best to respond to problematic content online.

¹⁷⁶ UDHR art. 23.

¹⁷⁷ Andrew Arsht and Daniel Etcovitch, “The Human Cost of Online Content Moderation,” *Harvard Journal of Law & Technology Digest*, March 2, 2018, <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation>.

5.5. Human Resources: Recruitment and Hiring

Organizations cannot operate without employees, but the process of identifying, recruiting, and hiring new employees is hard work. Increasingly, public- and private-sector employers are turning to AI to help with the hiring process for at least two reasons.¹⁷⁸ The first is capacity: the number of applicants per position has multiplied in the last several years, while staffing levels at human resources (“HR”) departments remain flat. The second is fairness: there is a growing awareness that hiring processes are rife with implicit bias and discrimination, and that hiring decisions often boil down to “is this person like me?” Many organizations believe that AI may offer at least a partial solution to this challenge.

The responsibility of business to respect human rights applies not just to the services they provide and the products they sell, but also to their internal operations. Flawed hiring processes may have significant implications for the right to freedom from discrimination,¹⁷⁹ the right to equal pay for equal work,¹⁸⁰ and the rights to freedom of expression and association.¹⁸¹ Governments have recognized the need for mechanisms to provide remedy for individuals subjected to discriminatory hiring practices and have created institutions such as the U.S. Equal Employment Opportunity Commission (“EEOC”) and the Canadian Human Rights Commission. As AI-based hiring systems become commonplace, it will be important to evaluate whether these existing mechanisms are up to the task of ensuring that these new technologies are free from bias.

5.5.1. *Traditional Approach to Recruitment and Hiring*

Recruiters have long relied on technology to streamline the hiring process. Today, HR departments commonly use applicant tracking systems (“ATS”) to aggregate applicant information and filter them based on certain criteria, such as years of experience, education, or other keywords. Shortlisted candidates are then interviewed and a decision is made on who to hire. Few quantitative data points are used in this process; instead, it relies on an often-flawed combination of pedigree and gut instinct.

¹⁷⁸ “A New Age of Opportunities: What Does Artificial Intelligence Mean for HR Professionals?” (Ontario: Human Resources Professionals Association, 2017).

¹⁷⁹ UDHR art. 7.

¹⁸⁰ UDHR art. 23(2).

¹⁸¹ UDHR art. 20.

Problems abound in this human-based system. There has been much debate about why we continue to see men from the majority group hired and promoted over women and minorities. Some claim that it may reflect systemic inequities in society leading to men being more highly educated and better prepared for a particular job.¹⁸² But research shows that much of the problem is discrimination resulting from implicit biases that manifest themselves in individual decision-making. In fact, our very definitions of success can exhibit bias. One experimental study shows that individuals are prone to shifting their definition of merit when evaluating applicants to advantage certain groups, which plays a role in gender and racial discrimination.¹⁸³ This distinction is important because, while systemic inequity is a serious problem, decisions marred by individual bias more directly implicate the right to be free from discrimination.¹⁸⁴

Research in the United States has shown, repeatedly, that “white-sounding” names receive 50% more callbacks than “African-American-sounding” names—despite otherwise identical resumes.¹⁸⁵ Moreover, males often receive more callbacks for traditionally “male” jobs. In one experiment designed to explore the dearth of women in STEM professions, researchers found that women were half as likely as men to be hired for a job, based on a flawed assumption that the female candidates performed worse on an arithmetic task. This was the case even when the women actually performed better than their male counterparts. The reason for this was that men were more likely to inflate their performance on the task in an interview and women were more likely to underestimate their performance.¹⁸⁶ Furthermore, the right to equal pay for equal work¹⁸⁷ is implicated by the fact that men receive higher starting salary offers than women for the same job.¹⁸⁸

¹⁸² Matthew Scherer, “AI in HR: Civil Rights Implications of Employers’ Use of Artificial Intelligence and Big Data,” *SciTech Lawyer* 13 (2017 2016): 12-16.

¹⁸³ Eric Luis Uhlmann and Geoffrey L. Cohen, “Constructed Criteria: Redefining Merit to Justify Discrimination,” *Psychological Science* 16, no. 6 (June 1, 2005): 474-80, <https://doi.org/10.1111/j.0956-7976.2005.01559.x>.

¹⁸⁴ UDHR art. 2.

¹⁸⁵ Marianne Bertrand and Sendhil Mullainathan, “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination” (Cambridge, MA: National Bureau of Economic Research, July 2003).

¹⁸⁶ Ernesto Reuben, Paola Sapienza, and Luigi Zingales, “How Stereotypes Impair Women’s Careers in Science,” *Proceedings of the National Academy of Sciences*, March 5, 2014, <https://doi.org/10.1073/pnas.1314788111>.

¹⁸⁷ UDHR art. 23(2)

¹⁸⁸ “2018 State of Wage Inequality in the Workplace Report,” *Hired*, accessed June 21, 2018, <https://hired.com/wage-inequality-report>.

Finally, the right to freedom of association¹⁸⁹ is implicated by human bias in the hiring process. Involvement in certain organizations, such as ethnic affinity groups or LGBTQ networks, can negatively impact job prospects. A résumé audit study found that women with a leadership role in a student LGBTQ organization received 30% fewer callbacks for a job posting than applicants with an identical resume but without the LGBTQ association.¹⁹⁰ Individuals may not know that this information is limiting their job prospects, but if they did, they might feel pressure to disassociate themselves from controversial groups.

Ultimately, the impact of AI hiring systems on human rights will depend, in part, on whether the controls meant to mitigate or to remedy rights-related harms in the existing human-based system can be applied to the new technology. It is to the technology now that we must turn in order to evaluate this question.

5.5.2. *AI Assisted Hiring and Recruitment*

Artificial intelligence is now being used to augment much of the hiring process. Job descriptions can be run through text analysis software to flag gendered language that might discourage highly qualified women from applying.¹⁹¹ Companies can enlist algorithms to advertise openings to eligible candidates through LinkedIn or Google’s ad network.¹⁹² AI is also being used to screen applicants using natural language processing to parse resumes.¹⁹³ Some technologies even draw on social media and other public data to supplement their analyses. After AI is used to narrow down the applicant pool, companies might invite candidates to conduct recorded interviews, where algorithms evaluate word choice, vocal inflection, and even emotions (using facial recognition).¹⁹⁴ These

¹⁸⁹ UDHR art. 20.

¹⁹⁰ Emma Mishel, “Discrimination against Queer Women in the U.S. Workforce: A Résumé Audit Study,” *Socius* 2 (January 1, 2016). <https://doi.org/10.1177/2378023115621316>.

¹⁹¹ Software employed by Textio and Gender Decoder use NLP paired with research on language pattern and implicit word associations to make job descriptions more gender neutral. See Textio, <https://textio.com/>, and <http://gender-decoder.katmatfield.com/>.

¹⁹² John Jersin, “How LinkedIn Uses Automation and AI to Power Recruiting Tools,” *LinkedIn Talent Blog* (blog), October 10, 2017, <https://business.linkedin.com/talent-solutions/blog/product-updates/2017/how-linkedin-uses-automation-and-ai-to-power-recruiting-tools>.

¹⁹³ See examples “Sovren,” accessed June 20, 2018, <https://www.sovren.com/>, and “Textkernel Launches the First Fully Deep Learning Powered CV Parser,” Textkernel (blog), February 8, 2018, <https://www.textkernel.com/extract-4-o-textkernel-launches-the-first-fully-deep-learning-powered-cv-parsing-solution/>.

¹⁹⁴ HireVue.com, “HireVue: Video Interview Software for Recruiting & Hiring,” accessed June 20, 2018, <https://www.hirevue.com>.

technologies purport to identify candidate “personalities” and help establish “fit” within the company. AI is clearly streamlining the hiring process, but the verdict on AI’s ability to mitigate negative human rights impacts is unclear.

Previous sections have noted how the veneer of objectivity that technology provides can be dangerous, because it obscures how AI often replicates human biases at scale. This is particularly worrying when AI is used to devise predictors of success that will determine hiring and advancement opportunities for future applicants and employees. There is already evidence that gender stereotypes have seeped into the “word embedding frameworks”¹⁹⁵ used in many machine learning and natural language processing technologies. One of the more egregious cases revealed in a 2016 study found that an algorithm trained on Google News articles to understand word meanings would respond to the query “man is to computer programmer as woman is to x” with x = homemaker.¹⁹⁶ In view of this example, there is a very real danger that an AI-based hiring algorithm trained on performance reviews,¹⁹⁷ employee surveys, and other data points meant to uncover the attributes of successful employees will reproduce existing patterns of bias in future hiring decisions. The system may produce more consistent results across candidates than human hiring managers, but the outputs of such a system can hardly be described as fair.¹⁹⁸

¹⁹⁵ Tolga Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” ArXiv:1607.06520 [Cs, Stat], July 21, 2016, <http://arxiv.org/abs/1607.06520>.

¹⁹⁶ Word embedding is a set of language and feature modeling techniques used in NLP to map words or phrases onto vectors of real numbers. This allows computer programs to understand and use word meaning, see Tolga Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” ArXiv:1607.06520 [Cs, Stat], July 21, 2016, <http://arxiv.org/abs/1607.06520>.

¹⁹⁷ Performance reviews and promotion structures often exhibit gender and racial bias, which contributes to the paucity of women and minorities as you move up the corporate ladder. See Kieran Snyder, “The Abrasiveness Trap: High-Achieving Men and Women Are Described Differently in Reviews,” *Fortune*, August 26, 2014, <http://fortune.com/2014/08/26/performance-review-gender-bias/>, and Buck Gee and Janet Wong, “Lost in Aggregation: The Asian Reflection in the Glass Ceiling” (The Ascend Foundation, September 2016), and Hannah Riley Bowles, Linda Babcock, and Lei Lai, “Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask,” *Organizational Behavior and Human Decision Processes* 103, no. 1 (May 2007): 84–103, <https://doi.org/10.1016/j.obhdp.2006.09.001>.

¹⁹⁸ Companies like Koru are using this method to establish their client organizations’ “predictive hiring fingerprint” and are claiming that it reduces bias and the proclivity to hire based on pedigree. <https://www.joinkoru.com/>.

5.5.3. Summary of Human Rights Impacts

With the foregoing in mind, the question of how machine learning technologies used in hiring will impact the right to be free from discrimination becomes more complicated.¹⁹⁹ Without intentional intervention in the programming, it seems likely that AI will reproduce the existing systemic patterns of bias and prejudice exhibited in the training data. This may lead AI-based hiring systems to identify metrics for assessing candidates that reflect structural biases rather than the objective determinants of real-world employment performance.

Some technologists and researchers have identified this as a concern and are devising technical solutions. One solution proposes a decoupling technique²⁰⁰ that, in the resume screening context, would allow an algorithm to identify top candidates using variables optimized based on other applicants of a certain category (e.g. race or gender) rather than against the entire applicant pool. In practice, this means the traits to select for a female or minority applicant would be identified based on the trends of other female or minority applicants, and these could differ from the identified successful traits of a male or majority applicant. The feasibility—and legality—of implementing a technical solution that optimizes fairness by distinguishing between individuals on sensitive personal characteristics attributes is highly dependent on jurisdiction.

More hope lies in companies intentionally designing algorithms to control for human biases and implementing auditing systems that can regularly test for bias and errors. Applied and Pymetrics, for example, have worked with academics to devise AI-based hiring systems that use anonymization, skills testing, work product analysis, neurological brain games, and other methods to remove bias based on gender, race, and pedigree.²⁰¹ While these efforts are promising, their outcomes still depend on the reliability of their algorithms and the level of bias in their data. Where AI is being used in the hiring process, it will be important to implement robust

¹⁹⁹ UDHR art. 2.

²⁰⁰ Cynthia Dwork et al., “Decoupled Classifiers for Fair and Efficient Machine Learning,” *ArXiv:1707.06613 [Cs]*, July 20, 2017, <http://arxiv.org/abs/1707.06613>.

²⁰¹ Applied, “Can We Predict Applicant Performance without Requiring CVs? Putting Applied to the Test — Part 1,” *Medium* (blog), September 21, 2016, <https://medium.com/finding-needles-in-haystacks/putting-applied-to-the-test-part-1-9fiad6379e9e>; Josh Constine, “Pymetrics Attacks Discrimination in Hiring with AI and Recruiting Games,” *TechCrunch* (blog), accessed February 22, 2018, <http://social.techcrunch.com/2017/09/20/unbiased-hiring/>.

auditing structures to regularly examine biases in the data and how these are influencing the outputs.

AI-based hiring systems may have a greater negative impact on the freedoms of association²⁰² and expression²⁰³ than the current human-based system. Similar to the concern that people will carefully curate their associations to maximize their credit scores, there is a risk that applicants will feel compelled to disassociate themselves from organizations that might hurt their chances of securing employment, because AI-based systems are more likely to detect such associations than human recruiters. Likewise, AI-based hiring systems may chill individuals from engaging in certain forms of expressive activity out of fear that their words will be used against them in the employment context.

²⁰² UDHR art. 20.

²⁰³ UDHR art. 19.

5.6. Education: Essay Scoring

Access to education is a right in and of itself and a key enabler of a panoply of other human rights. Educational attainment is the primary engine of social mobility;²⁰⁴ those who are more educated are better able to participate in the economy, engage in civic and public life, and improve their personal and local circumstances.

One of the key skills education systems around the world seek to develop in their students is the ability to write. Not only is the quality of one's writing an important factor in the job market, but it is also an important ability for achieving the "full development of the human personality" in the words of the Universal Declaration of Human Rights,²⁰⁵ as well as one of the primary means through which the right to free expression is exercised.

The importance of writing is perhaps why this skill is so frequently tested at every level of the education system. Most countries periodically administer standardized tests of their students' written abilities in the local *lingua franca* as they progress through the educational system. In many countries, students must pass writing tests to graduate from a particular level of schooling, or for admission to the next level. In North America, for example, both the Graduate Management Administrative Test ("GMAT") and the Graduate Record Examinations ("GRE") evaluate the ability of prospective management and law students to produce an analytical writing sample under time constraints.

How writing is marked matters. In the day-to-day school context, the quality of the feedback students receive on their writing will impact the prospects of student improvement over time. And in the context of gatekeeping exams such as the GMAT and the GRE, how test-takers' writing is evaluated may have life-long impacts on their future opportunities. Consequently, as automated techniques are increasingly being used to evaluate student writing, the question arises of what their impacts will be.

5.6.1. Traditional Approach to Essay Grading

The traditional approach to grading writing, whether in the ordinary schooling or the standardized testing context, is for trained individuals to perform this task. When a large volume of writing

²⁰⁴ Michael Greenstone et al., "Thirteen Economic Facts about Social Mobility and the Role of Education" (The Hamilton Project, June 26, 2013).

²⁰⁵ UDHR art. 26(2).

needs to be evaluated against an established performance standard, a common approach is for the individuals responsible for the grading to each evaluate a representative sample of what has been submitted, and to compare results in order to calibrate their approach for the rest of the grading.

In most contexts, evaluating the quality of writing requires not just considering the mechanics of grammar and syntax, but also evaluating the writing for the accuracy of the substance and on considerations such as style and rhetorical impact. Even so, in the context of one common standardized test, a study found that the score assigned to the writing component could be predicted accurately 90% of the time by considering just one variable: length.²⁰⁶

Although educators may have more time to engage with the style of the substance of a student's writing than the evaluator of a standardized test, the resource constraints under which most school systems operate may lead teachers to turn, intentionally or not, to more mechanical assessments of their students' writing. Such assessments may fail to provide the feedback students need for growth and improvement.²⁰⁷ Indeed, studies have found that educators often emphasize form over substance in teaching writing, which may well result in students prioritizing form over substance in their own writing.²⁰⁸

5.6.2. AI-Graded Essays

Following decades of research and many fruitless attempts, automated essay scoring has recently become a reality. Indeed, these technologies are among the more mature current applications of artificial intelligence. Using machine learning, these systems are trained to grade written materials by ingesting a set of exemplars

²⁰⁶ Arthur C. Graesser and Danielle S. McNamara, "Automated Analysis of Essays and Open-Ended Verbal Responses," in *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics.*, ed. Harris Cooper et al. (Washington: American Psychological Association, 2012), 307–25, <https://doi.org/10.1037/13619-017>; Michael Winerip, "SAT Essay Test Rewards Length and Ignores Errors," *The New York Times*, May 4, 2005, sec. Education, <https://www.nytimes.com/2005/05/04/education/sat-essay-test-rewards-length-and-ignores-errors.html>.

²⁰⁷ Deborah Reck and Deb Sabin, "A High-Tech Solution to the Writing Crisis," *The Atlantic*, October 16, 2012, <https://www.theatlantic.com/national/archive/2012/10/a-high-tech-solution-to-the-writing-crisis/263675/>.

²⁰⁸ Gary A. Troia and Steve Graham, "Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know," *Journal of Education and Psychological Consultation* 14, no. 1 (2003): 75–89.

that have been graded by human experts. These systems identify features in the set of training materials that correlate with success, and they then assess any written materials that are fed into the system according to what they have learned.²⁰⁹

Such automated systems are already in use in high-stakes standardized testing environments. For example, the written component of the GMAT exam is currently scored by an automated system and an expert human grader working separately. Should the AI and the human differ in the grade they assign by more than one point, a second human grader acts as a tiebreaker.²¹⁰

The use of automated grading systems has the potential to positively impact the right to education in a number of different ways. Automated systems permit students to engage in more deliberate practice of their writing, receiving more feedback on at least some elements of their writing than they would otherwise. In the context of communities without reliable access to quality writing instruction, whether due to poverty or other forms of marginalization, automated systems may be one of the only reasonably available means to obtain the feedback required to develop one's writing skill. Some studies have demonstrated that the instantaneous feedback of rudimentary grammar checkers have powerful effects on the quality of written expression, so it is reasonable to assume that the same is true of more nuanced AI-based approaches to the same basic endeavor.²¹¹

Relatedly, automating certain aspects of the grading of writing might free educators to spend more time focusing on higher-order teaching tasks, such as engaging with students' ideas and arguments. To many, writing is more than sentence structure and word-choice: it is the expression of ideas and emotions, drawing on cognitive skills distinct from those underpinning grammar and syntax.²¹² Moreover, automated systems have the potential to eliminate bias in grading by removing the opinions of the author

²⁰⁹ Corey Palmero, "A Gentle Introduction to Automated Scoring.Pdf" (Measurement Incorporated, October 2017),

<http://www.measurementinc.com/sites/default/files/2017-10/A%20Gentle%20Introduction%20to%20Automated%20Scoring.pdf>.

²¹⁰ "How the Analytical Writing Assessment Is Scored," Economist GMAT Tutor, April 28, 2017, <https://gmat.economist.com/gmat-advice/gmat-overview/gmat-scoring/how-analytical-writing-assessment-scored/>.

²¹¹ Paul Morphy and Steve Graham, "Word Processing Programs and Weaker Writers/Readers: A Meta-Analysis of Research Findings," *Reading and Writing* 25, no. 3 (March 2012): 641–78, <https://doi.org/10.1007/s11455-010-9292-5>.

²¹² Troia and Graham, "Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know."

and the grader and any relationship between them from the equation.²¹³

On the flipside, there are serious concerns relating to the fact that these systems cannot understand what is written in the same way as human readers.²¹⁴ Some systems might well be able to detect offensive content, but at least for the foreseeable future, artificial intelligence systems will not realistically possess the general intelligence that humans do which enables them to evaluate the validity of written material.²¹⁵ Especially in our era where the truth and accuracy of written materials has become an issue of the utmost public importance, the likely inability of automated grading systems to assess factual validity is of concern.

Furthermore, there are also significant concerns about what incentives these systems create for those who are subject to being evaluated by them.²¹⁶ Consider, for example, that a famous essay by the renowned MIT linguist Noam Chomsky received a grade of “fair” when it was fed into an automated grading system.²¹⁷ If students

²¹³ For examples of bias in writing grading, see John Malouff, “Bias in Grading,” *College Teaching* 56, no. 3 (July 2008): 191–92, <https://doi.org/10.3200/CTCH.56.3.191-192>. Bias may also come in the form of pre-existing stereotypes influencing how critically one reviews the essay. For an example in the legal profession, see Arin N. Reeves, “Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills,” *Yellow Paper* (Nextions, 2014).

²¹⁴ Stephen P. Balfour, “Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review,” *Research & Practice in Assessment* 8, no. 1 (2013): 40–48. For an example of the concerns, see Kai Riemer, “On Rewarding ‘Bullshit’: Algorithms Should Not Be Grading Essays,” *Undark*, accessed April 21, 2018, <https://undark.org/article/rewarding-bullshit-algorithms-classroom/>.

²¹⁵ For example, Dr. Les Perelman of MIT and his team developed a piece of software they call the “Basic Automatic B.S. Essay Language Generator” (BABEL) which, upon the input by the user of three words, auto-generates essays that routinely receive near-perfect scores from various automated grading systems that are currently in use, although their content is nothing but machine-generated nonsense. See Steve Kolowich, “Writing Instructor, Skeptical of Automated Grading, Pits Machine vs. Machine,” *Chronicle of Higher Education*, April 28, 2014, <https://www.chronicle.com/article/Writing-Instructor-Skeptical/146211>. Dr. Perelman’s BABEL system is available to try out at <http://lesperelman.com/>. To be sure, humans are not immune to being fooled by machine-generated gibberish. See Danielle Wiener-Bronner, “More Computer-Generated Nonsense Papers Pulled From Science Journals,” *The Atlantic*, Mar. 3, 2014, <https://www.theatlantic.com/technology/archive/2014/03/more-computer-generated-nonsense-papers-pulled-science-journals/358735/>.

²¹⁶ Jinhao Wang and Michelle Stallone Brown, “Automated Essay Scoring Versus Human Scoring: A Correlational Study,” *Contemporary Issues in Technology and Teacher Education* 8, no. 4 (2008): 16.

²¹⁷ “Parts of Noam Chomsky’s Essay ‘The Responsibility of Intellectuals’ Grammar Checked by ETS’s Criterion and WhiteSmoke | Les Perelman, Ph.D.,” accessed June 21, 2018, <http://lesperelman.com/writing-assessment-robo-grading/parts-noam-chomskys-essay-grammar-checked/>.

respond to the growing prevalence of automated grading systems by focusing on form and length to the detriment of style and substance, these technologies may be doing them a disservice.

Finally, automated grading systems depend on the collection, storage, and analysis of vast quantities of written material. This raises not only the standard privacy-related concerns that accompany most AI systems,²¹⁸ but there is an additional risk that these systems might chill the full enjoyment of the right to free expression. Even in the educational context, a student might be more willing to share writing about something that is deeply personal or controversial with a trusted teacher as opposed to an AI system, if it is going to end up being catalogued in a vast database and subject to being used as training data for some other purpose.

5.6.3. Summary of Human Rights Impacts

On balance, the rise of automated grading systems is likely to have a positive impact on the right to education, as these systems can potentially increase global access to at least some feedback on people's writing. In much of the world today, educational systems are simply too overburdened to provide the kind of individualized evaluation and feedback on writing that is desirable, so the potential of these technologies to improve the situation over the status quo is considerable. The feedback these systems provide might fall short of the Platonic ideal of individualized attention from expert human instructors, but they are a step in the right direction for the vast majority of students worldwide who lack affordable access to high-quality instruction.

Considering that the ability to write is a key enabler of a panoply of civil and political rights, from that of free expression to the right to take part in cultural and scientific life, the improvements that automated grading systems will make in individuals' ability to write when judged on a global basis is likely to improve their ability to enjoy these rights. This is especially true of the crucial right that everyone possesses to an adequate standard of living, as the ability to write is so central to one's employment prospects and ability to participate effectively in various aspects of society.

²¹⁸ Related privacy concerns have been raised about automated plagiarism-detection software, which also require the storage and retention of vast quantities of student-written material to be effective. See Bo Brinkman, "An Analysis of Student Privacy Rights in the Use of Plagiarism Detection Systems," *Science and Engineering Ethics* 19, no. 3 (2013): 1255–1266.

The impact of these automated systems on the right to free expression, however, is more complex. On one hand, anything that improves the ability of people to express themselves in an effective manner would seem to positively impact the enjoyment of this right. On the other hand, as noted above, large-scale collection of written materials that automated systems necessarily entail may chill some individuals from setting down in writing things that they might have said otherwise.

6. Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach

The six use cases explored in the previous section demonstrate how artificial intelligence has the potential to positively impact the full range of human rights. Automating complex tasks that currently require the labor of highly trained professionals could well usher in greater access to specialized healthcare, education, and financial services. Such technologies also have the considerable potential to reduce and correct for various biases that plague human decision-making, from outright discrimination to our reliance on heuristics that sometimes lead us astray. Yet this promise comes at an almost inevitable cost to our privacy due to the data-intensive nature of these technologies which, in turn, may chill the exercise by many of their civil and political rights. Likewise, the possibility that these technologies will reproduce and ossify existing patterns of discrimination and bias, while also producing troubling distributive consequences, must be contended with.

This conundrum gives rise to the question of how we can enjoy the benefits of artificial intelligence—especially the vast potential for positive impacts on human rights— while minimizing its real negative risks.

This is not a new question, but rather one that has arisen with every major technological innovation throughout history. From the development of industrial machinery to the invention of the automobile, transformative technological changes have posed a profound challenge to the existing social order. These technologies utterly transformed society in their time—oftentimes for the better, yet they were frequently accompanied by bad consequences, too. Industrialization, for example, democratized the availability of goods that were once luxuries, though at the cost of widespread economic displacement, Dickensian working conditions, suffocating air pollution, and colonial patterns of natural resource exploitation. Likewise, the automobile revolutionized human mobility and fundamentally transformed the economy, though with the negative consequences of air pollution, urban sprawl, and millions of traffic casualties every year.

The negative impacts of these and other transformative technologies were felt most acutely as they were first coming into widespread use. Over time, however, society responded by developing control mechanisms to attempt to enjoy the good while minimizing the bad. Some controls are regulatory in nature, such as laws and norms that

specify how and when a technology might be used, while others are technological, such as design features that channel a technology towards certain uses and away from others.²¹⁹ Oftentimes, controls are a mix of the two, such as regulatory standards that mandate specific design characteristics.

History shows that it can take quite some time to develop effective mechanisms to control new technologies. The Industrial Revolution began to transform Britain in the 18th century, but it was only in the mid-19th century that Parliament started enacting legislation to address its consequences.²²⁰ Likewise, although automobiles became commonplace in the first decades of the 20th century, the first systematic studies of vehicular safety took place in the 1940s,²²¹ and the first comprehensive automobile safety laws weren't enacted until the 1960s.²²²

Of course, not all of these controls are effective or ideal. Most, if not all of them, are at least partially flawed. Some are too permissive to adequately address the negative consequences of the regulated technology, while others are too restrictive to permit the realization of its benefits. All the same, we must think about which controls are necessary, sufficient, and appropriate to reduce and redress the human rights impacts of artificial intelligence.

We are fortunate to be in a position to design the regulatory and technological controls required to maximize the human rights and other benefits of AI concurrently with the technology itself. AI is the first truly transformative technology to come of age following the articulation of the United Nations Guiding Principles on Business and Human Rights. It is emerging at a time that it is widely understood that businesses have a responsibility to respect human rights, and that due diligence is the key to doing so. Whereas in the past private sector innovators could be ignorant or willfully blind to

²¹⁹ In the U.S. context, Lawrence Lessig famously noted that the “East Coast Code” of laws and regulations promulgated in Washington D.C., and the “West Coast Code” produced by software engineers in Silicon Valley, are both fundamentally forms of regulation. Lawrence Lessig, *Code*, Version 2.0 (New York: Basic Books, 2006), <http://codev2.cc/download+remix/Lessig-Codev2.pdf>.

²²⁰ B.L. Hutchins and A. Harrison, *A History of Factory Legislation*, 2nd ed. (London: P. S. King & Son, 1911), <https://archive.org/details/historyoffactory014402mbp>.

²²¹ For a popular and accessible history of automobile safety, listen to 99% Invisible, *The Nut Behind the Wheel*, accessed June 21, 2018, <https://99percentinvisible.org/episode/nut-behind-wheel/>.

²²² “National Traffic and Motor Vehicle Safety Act of 1966,” Pub. L. No. 89–563 (1966).

the human rights consequences of the technologies they are developing, that is no longer the case.

Due diligence, as the term is used in the Guiding Principles, is the essential first step toward identifying, mitigating, and redressing the adverse human rights impacts of AI. Therefore, as a minimum, public policy efforts should be directed toward ensuring that all who are involved in building these systems engage in the kinds of due diligence that will ensure that they respect human rights by design. Such efforts may be enhanced by mandating or incentivizing the developers and operators of AI systems to make available the training data and the outputs of their systems to external reviewers.²²³

It is heartening that many of the biggest players in developing AI have risk management systems in place that trigger human rights due diligence processes at all appropriate stages in the lifecycle of a technology.²²⁴ That being said, there are at least three challenges endemic to the AI space that may prevent human rights due diligence from being as effective as it might otherwise be.

The first arises from the relatively low awareness among small, early-stage companies of the corporate responsibility to carry out human rights due diligence. This is by no means a challenge that is unique to AI, but given the potential of these technologies to scale up rapidly, it might be more problematic in this space than in other industry verticals.²²⁵ Moreover, given that certain AI systems are often empty vessels into which the end-user can feed whatever training data it wants to automate a formerly manual process, technology developers can be too remote from on-the-ground

²²³ In this vein, New York University's AI Now Institute has developed a framework for public-sector entities in the United States to use in carrying out "algorithmic impact assessments" prior to purchasing or deploying automated decision systems. Dillon Reisman et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (New York University AI Now Institute, April 2018), <https://ainowinstitute.org/aiareport2018.pdf>.

²²⁴ For example, the eleven member-companies of the Global Network Initiative ("GNI"), which include some of the biggest players in the AI space, commit to "carry out human rights due diligence to identify, prevent, evaluate, mitigate and account for risks to the freedom of expression and privacy rights that are implicated by the company's products, services, activities and operations." GNI member-companies are independently assessed every two years to evaluate their compliance with this and other commitments. "Implementation Guidelines" (Global Network Initiative), accessed June 21, 2018, <https://globalnetworkinitiative.org/implementation-guidelines/>.

²²⁵ Dalia Ritvo, Vivek Krishnamurthy, and Sarah Altschuller, "Managing Users' Rights Responsibly—A Guide for Early-Stage Companies," 2016, <http://www.csrandthelaw.com/wp-content/uploads/sites/2/2016/03/Managing-Users-Rights-Responsibly-A-Guide-For-Early-Stage-Companies-no-logos.pdf>.

realities to assess the human rights impacts of the uses of their products.²²⁶ Correspondingly, there is an opportunity to significantly advance human rights by adopting measures to incentivize much wider due diligence efforts throughout the entire AI ecosystem.

The second arises from the difficulty in ascertaining the real-world impacts of any given AI application prospectively. That difficulty arises from the inscrutability of so many AI systems and from the complex interactions that these systems have once they begin to operate in the real world. It is hard enough to predict what human rights impacts a relatively anodyne product will have when it is released into the marketplace, hence the challenge of assessing the human rights impacts of AI systems before they are deployed is all the more considerable. The problem is particularly acute in AI systems which utilize machine or deep learning, such that the AI developer herself may not be able to predict or understand the system's output.²²⁷

To be sure, the Guiding Principles make clear that human rights due diligence is an ongoing responsibility for precisely this reason: not all impacts can be predicted, even with reasonable diligence. Correspondingly, all entities that are involved in the development or use of these technologies must have measures in place to ensure that human rights due diligence is not a matter of “once and done.” Especially given the complexity of AI systems and the fact that the results are often not explainable by conventional means, new analytical techniques and performance metrics may need to be developed to determine whether AI systems are helping or harming human rights. Developing these techniques and metrics is a challenge that the computer science community is working to tackle with alacrity. In Europe, this challenge has been framed in part by

²²⁶ Consider, for example, the controversy that emerged around the time that this report was being finalized regarding the U.S. government's use of facial recognition technology supplied by Microsoft in implementing its (now-rescinded) policy of separating the children of unlawful migrants from their parents. One can speculate that Microsoft could not have foreseen how the U.S. government would use this technology at the time it contracted to provide this system, which highlights the need for companies in this space to conduct *ongoing* due diligence. Catherine Shu, “Microsoft Says It Is ‘Dismayed’ by the Forced Separation of Migrant Families at the Border,” *TechCrunch*, June 19, 2018, <http://social.techcrunch.com/2018/06/18/microsoft-says-it-is-dismayed-by-the-forced-separation-of-migrant-families-at-the-border/>.

²²⁷ To be clear, many AI applications reason in ways beyond human comprehension. This is particularly true for applications based on machine learning and deep learning. Yet, that difference may be insufficient to justify holding AI back. In fact, it may even be a reason to delegate decisions to AI. David Weinberger, “Our Machines Now Have Knowledge We’ll Never Understand,” *WIRED*, April 18, 2017, <https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/>.

the provisions of the General Data Protection Regulation, which requires some human involvement in automated decision-making²²⁸ and encourages the development of “a right to an explanation.”²²⁹

A third complication arises from the uncertainty as to what constitutes an effective remedy to an adverse human rights impact generated by an AI system. Since a right without a remedy is no right at all, the right to remedy is the “third pillar” of the “protect, respect, remedy” framework on which the Guiding Principles rest. Specifically, Guiding Principle 25 recognizes the duty of the state to provide access to effective judicial and other remedies to all those who have been affected by business-related human rights abuses. Judicial remedies in particular may be better suited to addressing the adverse consequences of AI on some human rights over others. For example, judicial remedies may well be more effective in detecting and redressing adverse human rights impacts caused by the use of AI in the criminal justice system, as compared to some other fields of endeavor. Especially since criminal procedural rights are articulated in much more detail than most other human rights in domestic and international law, there is simply more material to work with in terms of identifying when an AI system is adversely impacting rights in this category.²³⁰

Another major challenge facing both judicial and non-judicial remedies is the nature of the harm that AI makes possible. That is, all remedial systems, whether public or private, are much better at remediating substantial harms suffered by the few, as opposed to less significant harms suffered by the many. Consider, for example,

²²⁸ GDPR, art. 22.

²²⁹ Ibid., art 22 and recital 71. For more information, see Bryce Goodman and Seth Flaxman, “European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation,’” *AI Magazine* 38, no. 3 (October 2, 2017): 50, <https://doi.org/10.1609/aimag.v38i3.2741> and Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation,” *International Data Privacy Law* 7, no. 2 (May 2017): 76–99, <https://doi.org/10.1093/idpl/ixp005>.

²³⁰ The fact that the judiciary might be better able to grapple with the adverse impact of AI in its own backyard does not mean that it will reach the right answer in every case, or that the remedies it provides when a violation is found will be sufficient. For example, the U.S. Supreme Court has been roundly criticized for denying review of *Wisconsin v. Loomis*, 881 N.W.2d 749 (2016). See Ellora Israni, “Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis,” *Harvard Journal of Law & Technology Digest*, 2017, <https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1> and Taylor R Moore, “Trade Secrets and Algorithms as Barriers to Social Justice” (Center for Democracy & Technology, August 2017). The point is simply that some alleged rights violations attributed to AI systems are easier to redress judicially in some fields as opposed to others.

some of the most widely-known operational level grievance mechanisms that have been established in the last decade to remedy the adverse human rights impacts of businesses. These include mechanisms established by mining companies to compensate the victims of sexual violence,²³¹ or by global technology companies to vindicate the right that some courts have recognized of individuals to be “forgotten” online.²³² These systems provide remedies to individuals or to small groups of people that have suffered a particularized human rights harm, but they are simply not designed to cope with much more diffuse and oftentimes covert harms that might be every bit as pernicious.²³³

These difficulties are magnified in the AI space by the challenge of detecting the harm and determining and proving causation. Consider, for example, the difficulties that a loan applicant would face in proving that a lending algorithm has discriminated against them, in a situation where seven prospective lenders turned them down, but three others offered them credit. Assuming that the objective truth of the matter is that some of the seven decliners engaged in discrimination, would this person even suspect that they have been the victim of discrimination? What might be the required elements of proof to establish a discrimination claim? How costly would it be to bring such a claim, as against the anticipated value of the remedies available? What expert evidence and analysis would be required to open the “black box” of the algorithm, especially when it is protected by trade secrets and intellectual property law? Now assume that the stakes of what the algorithm is deciding are much lower than a loan, and yet there are adverse consequences distributed over a large population. Who would go through the trouble of seeking a remedy for that harm, and how and where might they do so?

Then there are the additional complications around remedying AI’s impacts on economic, social, and cultural rights. International and domestic legal systems alike have much more developed doctrines and procedures with regard to civil and political rights than they do

²³¹ Yousuf Aftab, “Pillar III on the Ground: An Independent Assessment of the Porgera Remedy Framework” (Enodo Rights, January 2016), <http://www.enodorigths.com/assets/pdf/pillar-III-on-the-ground-assessment.pdf>.

²³² Jacques de Werra, “ADR in Cyberspace: The Need to Adopt Global Alternative Dispute Resolution Mechanisms for Addressing the Challenges of Massive Online Micro-Justice,” *Swiss Review of International & European Law* 26, no. 2 (2016): 289, <https://doi.org/10.2139/ssrn.2783213>.

²³³ To be sure, the same criticism can be levelled against traditional courts which, despite such innovations as class-action lawsuits and contingency fee arrangements, remain an unaffordable and inaccessible option for many victims of rights violations.

for economic, social, and cultural rights. This is due in part to the fact that the international human rights community has prioritized civil and political rights over economic, social, and cultural rights for the last 70 years.²³⁴ But it is also because the state duty in relation to economic, social, and cultural rights is to progressively realize them over time, in view of the available resources, which makes it much harder to identify when these rights are adversely impacted—especially by businesses.

The recent General Comment on “State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities” makes clear the difficulties.²³⁵ The General Comment notes that the state obligation to protect human rights requires them to “prevent effectively infringements of economic, social and cultural rights in the context of business activities” by adopting “legislative, administrative, educational and other appropriate measures, to ensure effective protection against Covenant rights violations linked to business activities.”²³⁶ Yet all of the examples the General Comment provides of “rights violations linked to business activities” are attributable to state failures to regulate the marketplace.²³⁷

The underdevelopment of the regime of economic, social, and cultural rights makes it difficult for businesses engaged in human rights due diligence to know what they should do when their systems adversely impact one of these rights. Consider, for example, the impact that at least some AI systems are sure to have on employment. While it is the duty of the state to protect the right to work, large-scale workforce displacement caused by the deployment of AI obviously burdens this right. As things stand right now, however, it is difficult for a business to determine what if anything it should do to mitigate this impact on the right to work.

The example of the right to work points to the profound challenges related to addressing the distributive consequences of AI—especially with regard to economic rights, in view of the fears that AI might

²³⁴ Samuel Moyn, *Not Enough: Human Rights in an Unequal World* (Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 2018); Samuel Moyn, “How the Human Rights Movement Failed,” *The New York Times*, April 24, 2018, sec. Opinion, <https://www.nytimes.com/2018/04/23/opinion/human-rights-movement-failed.html>.

²³⁵ United Nations Economic and Social Council, Committee on Economic, Social and Cultural Rights, General Comment 24 on “State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities” (2007), U.N. Doc. E/C.12/GC/24.

²³⁶ *Ibid.*, para. 14.

²³⁷ *Ibid.*, paras. 18–22.

trigger widespread unemployment. As several of the case studies find, a particular AI system being used in a particular field of endeavor can positively impact the enjoyment of a particular human right by some individuals, at the same time as it adversely affects that right for others.

It may well be that over the course of time, the international human rights system can develop guidance that is more responsive to such distributive issues in the context of AI. Until that happens, however, there is a strong case to be made that national governments should weigh in on these questions, whether by soft suasion or hard regulation, in a manner that is rights-respecting, yet also reflects the country's particular values and public policy priorities. In other words, when due diligence reveals human rights impacts that have complex distributive consequences, the need for government public policy leadership is at its greatest.²³⁸

There is also an important role for governments to play in creating conditions that encourage businesses to take their human rights responsibilities seriously. In his influential study of private initiatives to improve labor rights in global supply chains, Richard Locke found that “the effectiveness of private regulatory programs is very much tied to the strength of public authoritative rule-making institutions.”²³⁹ In the AI space, governments could consider creating incentives to ensure that effective due diligence is undertaken, and to build capacity among earlier-stage companies to develop their technologies in a rights-respecting manner.

Finally, there is a crucial role for governments to play in creating accountability and redress systems for their own use of algorithmic tools, and for those adverse impacts that cannot easily be addressed by private grievance mechanisms. The General Data Protection Regulation (“GDPR”), which recently came into force in the European Union, is noteworthy in this regard for its provisions requiring “data subjects” to be provided with “meaningful information about the logic involved, as well as the significance and the envisaged consequences” of the automated processing of their personal data.²⁴⁰

²³⁸ As Richard Locke notes in a related context, “[t]he inherent problem with private voluntary initiatives... is their inability to reconcile diverse and conflicting interests and thus promote solutions that require collective action among [] myriad actors...” Richard M. Locke, *The Promise and Limits of Private Power: Promoting Labor Standards in a Global Economy*, Cambridge Studies in Comparative Politics (Cambridge; New York: Cambridge University Press, 2013): 178.

²³⁹ *Ibid.*, 68.

²⁴⁰ GDPR art. 14(2)(g).

It is too soon to tell whether the GDPR's embrace of algorithmic "explainability" will prove to be successful in creating greater accountability for such systems, or whether this approach will instead chill promising developments in AI that produce useful results even if their logic defies human comprehension. What is certain, however, is that collaboration between technology companies, governments, and representatives of the diverse community of stakeholders that AI will impact is required to develop new ways of ensuring that this technology delivers on its promise in a rights-respecting manner.

7. Conclusion

As should now be clear, the relationship between artificial intelligence and human rights is complex. A single AI application can impact a panoply of civil, political, economic, social, and cultural rights, with simultaneous positive and negative impacts on the same right for different people. Multiply these impacts across the full range of cases where AI is already in use or will soon become commonplace, and the magnitude of this technology's impact on society begin to become clear.

Society has dealt with revolutionary technological change in the past, and we have always arrived at a new equilibrium. But we today are better placed than our forebears for the change that is upon us because of the adoption, 70 years ago this December, of the Universal Declaration of Human Rights.

The UDHR gives us a powerful and universally-accepted framework not just for identifying and overcoming past and present wrongs, but also for building a future that respects and honors the rights of every person. This depends, however, on our remaining vigilant to the impacts of our actions on the rights of others. Hence the importance the Guiding Principles place on due diligence both before we deploy these powerful new technologies, and throughout their lifecycle, too.

We are heartened by the growing attention that human rights-based approaches to assessing and addressing the social impacts of AI have begun to receive. We view it as a promising sign that so many of the private enterprises at the forefront of the AI revolution are recognizing their responsibility to act in a rights-respecting manner. But the private sector cannot do it alone, nor should it: governments have a crucial role to play, both in their capacities as developers and deployers of this technology, but also as the guarantors of human rights under international law.

The fundamental role of government in defining and making available remedies for human rights violations cannot be overstated. Of equal or greater importance, however, is the governmental responsibility to evaluate and address the distributive consequences of AI. The institutions and processes of democratic government are the only ones with the legitimacy to determine what distribution of benefits and burdens across society is fair, so now is the time for them to embrace their role in shepherding society through the changes that lie ahead.

8. Further Reading

8.1. Understanding AI

David Weinberger, “Our Machines Now Have Knowledge We’ll Never Understand,” *Wired*, April 18, 2017, <https://www.wired.com/story/our-machines-now-have-knowledge-we-well-never-understand/>.

This article explores how artificial intelligence is changing the way we think about knowledge by delving into the “explainability” debate. It explains machine learning, artificial neural networks and their implications in an accessible manner. Many machine learning models, like Google’s AlphaGo algorithm, are “ineffably complex and conditional”: they make decisions based on opaque and unexplainable patterns, not transparent principles. These systems are becoming more accurate as data becomes more abundant, yet there is a tradeoff between being able to understand why an algorithm makes a decision (a function of its complexity) and its accuracy. Because reality is complex, artificial intelligence may be able to account for an abundance of factors that are beyond human comprehension. Yet the lack of explainability can make it difficult to identify bias in algorithms. (*Category: Concise overview*)

Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society* 3, no. 1 (January 5, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674.

This article describes three different kinds of “opacity” in algorithms: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) opacity arising from the characteristics of machine learning algorithms that make them useful. Burrell argues that recognizing these distinct forms of opacity is important to determining what technical and non-technical solutions can prevent algorithms from causing harm. On (1), secrecy may be essential to the proper function of an algorithm (such as to prevent it from being gamed), but such algorithms are easily reviewable by trusted and independent auditors. Regarding (2), the solution to technical illiteracy is simply greater public education. Finally, (3) is difficult because there may be a trade-off between fairness, accuracy, and interpretability. Certain AI techniques could be avoided in fields where transparency is crucial, or new benchmarks could be developed to assess such

algorithms for discrimination and other issues. (Category: In-depth)

8.2. Defining the Problem: What's at Stake?

Navneet Alang, "Turns Out Algorithms are Racist," *New Republic*, August 31, 2017. <https://newrepublic.com/article/144644/turns-algorithms-racist/>.

This article explains that AI is only as good as the data it is fed. Computers and software, even at their most sophisticated, are essentially input-output systems that are "taught" by feeding them enormous amounts of data. Hence if the input data reflect gender, racial, or other biases, so too will the output. (Category: Concise overview)

Robert Hart, "If you're not a white male, artificial intelligence's use in healthcare could be dangerous," *Quartz*. July 10, 2017, <https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/>..

While acknowledging the potential of AI to revolutionize healthcare, this article points to the danger that this technology will perpetuate healthcare inequalities due to its reliance on existing stores of medical data. Groups including women (especially pregnant women), the young, and the elderly are excluded from many medical research studies, which may result in errors when individuals from these groups are treated by AI systems trained on this data. (Category: Case study – healthcare)

J. Kleinberg et al., "Human Decisions and Machine Predictions," National Bureau of Economic Research, February 2017, <https://www.cs.cornell.edu/home/kleinber/w23180.pdf>.

This article highlights how machine learning can help humans make better decisions, using the example of judges making bail decisions. While acknowledging the difficulties associated in fully automating bail determinations—both due to biases in the training data that would be fed into such a system and the complex mix of factors that judges weigh—AI-based analyses show that the number of people jailed before trial can be substantially reduced without impacting the crime rate. Such insights can then be used by judges to improve their own decision-making. (Category: Case study – criminal justice)

danah boyd, Karen Levy, and Alice Marwick, *The Networked Nature of Algorithmic Discrimination*, Washington, DC: New America Foundation, 2014,
<https://www.danah.org/papers/2014/DataDiscrimination.pdf>.

This study points to the dangers surrounding algorithms making predictions about you based on those you associate with—such as your friends and neighbors. While existing laws prohibit racial and gender discrimination, among other things, an individual’s position in a social network is deeply affected by these and other variables—which algorithms might then use to make predictions about us, leading to unfair results. Consequently, the authors argue that we need to rethink our models of discrimination to consider not just an individual’s immutable characteristics, but also the impacts of how algorithms position us within a network and society, too. (*Category: Novel issues – networked discrimination*)

8.3. Approaches to Regulating AI

S. Wachter et al., “Transparent, explainable, and accountable AI for robotics,” *Science Robotics* 2, no. 6 (May 31, 2017),
<http://robotics.sciencemag.org/content/2/6/eaan6o8o.full>.

This article provides a brief overview of the challenges facing governments seeking to regulate AI. After briefly grounding the regulation of automated systems in its historical context, it raises the critical questions facing regulators seeking to enact optimal laws in the space. (*Category: Concise overview*)

IEEE Global Initiative on Ethics of and Autonomous and Intelligent Systems, “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems” (IEEE, 2017),
<http://standards.ieee.org/develop/indconn/ec/autonomoussystems.html>

This document, prepared by the world’s largest professional organization in the technology space, calls for an ethics- and values-based approach to dealing with the impacts of intelligent and autonomous systems that prioritizes human well-being within a given cultural context. (*Category: In-depth analysis*)

Corrine Cath et al., “AI and the ‘Good Society’: the US, EU, and UK approach,” *SSRN Electronic Journal* (2016).
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906249

In October 2016, the White House, the European Parliament, and the UK House of Commons each issued reports outlining their vision on how to prepare society for the widespread use of AI. This article provides a comparative assessment of these three reports to facilitate the design of policies favorable to the development of a ‘good AI society’. (*Category: Comparative overview*)

National Science and Technology Council: Committee on Technology, *Preparing for the Future of Artificial Intelligence*. Washington, D.C.: Executive Office of the President, October 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

This report surveys the current state of AI research, its existing and potential applications, and the questions that AI raises for society as a whole and for public policy in particular. The report recommends certain specific governmental and non-governmental actions within the U.S. context and sets forth a strategic plan for U.S government funding of AI. (*Category: Regulatory template*)

Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. New York University AI Now Institute, April 2018. <https://ainowinstitute.org/aiareport2018.pdf>.

This report proposes an “Algorithmic Impact Assessment” framework to be used by public entities in the U.S. in procuring and deploying AI systems. The framework is designed to support affected communities and stakeholders as they seek to assess the claims made about automated decision systems, and to determine where their use is acceptable. It offers key elements for the construction of a public agency algorithmic impact assessment, and a practical accountability framework combining agency review and public input. (*Category: Regulatory template*)

8.4. Business and AI

Sherif Elsayed-Ali, “Why Embracing Human Rights Will Ensure AI Works for All,” April 2018, <https://www.weforum.org/agenda/2018/04/why-embracing-human-rights-will-ensure-AI-works-for-all/>.

This article recommends four actions to prevent discriminatory outcomes in machine learning based on the UN Guiding Principles on Business and Human Rights. These are: (1) active inclusion of underrepresented populations in datasets and AI

development, (2) fairness in the interpretation of biased data, (3) a right to understand how algorithmic decisions are made, and (4) access to redress when wrong decisions are made. (*Category: Concise overview*)

R. Jorgenson, “What Platforms Mean When They Talk About Human Rights,” *Policy & Internet* 9, no. 3 (May 29, 2017) <https://doi.org/10.1002/poi3.152>.

This article examines how two major Internet platforms—Google and Facebook—make sense of human rights. Based on primary research, the authors find that the companies frame themselves as strongly committed to human rights. Yet this framing focuses primarily on government rights violations, rather than the companies’ own adverse impacts on its users’ rights. (*Category: Case study—internet platforms*)

Shift, Oxfam and Global Compact Network Netherlands, *Doing Business with Respect for Human Rights: A Guidance Tool for Companies*, 2016, https://www.businessrespecthumanrights.org/image/2016/10/24/business_respect_human_rights_full.pdf.

This report provides comprehensive guidance to businesses on what they should do to operationalize their responsibility to respect human rights, as recognized in the UN Guiding Principles on Business and Human Rights. (*Category: In-depth*)

The Future Computed

Artificial Intelligence and its role in society

By Microsoft

Foreword by Brad Smith and Harry Shum



The Future Computed

Artificial Intelligence
and its role in society

By Microsoft

With a foreword by
Brad Smith and Harry Shum

Published by Microsoft Corporation
Redmond, Washington. U.S.A.
2018

First published 2018 by Microsoft Corporation
One Microsoft Way
Redmond, Washington 98052

© 2018 Microsoft. All rights reserved

ISBN 978-0-9997508-1-0

Table of contents

Foreword	
The Future Computed	1
Chapter 1	
The Future of Artificial Intelligence	23
Microsoft's Approach to AI	34
The Potential of Modern AI -	44
Addressing Societal Challenges	
The Challenges AI Presents	49
Chapter 2	
Principles, Policies and Laws for the	51
Responsible Use of AI	
Ethical and Societal Implications	57
Developing Policy and Law for	74
Artificial Intelligence	
Fostering Dialogue and the Sharing of	83
Best Practices	

Chapter 3	
AI and the Future of Jobs and Work	85
The Impact of Technology on Jobs and Work	92
The Changing Nature of Work, the Workplace and Jobs	102
Preparing Everyone for the Future of Work	108
Changing Norms of Changing Worker Needs	123
Working Together	134
 Conclusion	
AI Amplifying Human Ingenuity	135
 Endnotes	 139

Foreword

The Future Computed

By Brad Smith and Harry Shum



Twenty years ago, we both worked at Microsoft, but on opposite sides of the globe. In 1998, one of us was living and working in China as a founding member of the Microsoft Research Asia lab in Beijing. Five thousand miles away, the other was based at the company's headquarters, just outside of Seattle, leading the international legal and corporate affairs team. While we lived on separate continents and in quite different cultures, we shared a common workplace experience within Microsoft, albeit with differing routines before we arrived at the office.

At that time in the United States, waking to the scent of brewing coffee was a small victory in technology automation. It meant that you had remembered to set the timer on the programmable coffee maker the night before. As you drank that first cup of coffee, you typically watched the morning news on a standard television or turned the pages of the local newspaper to learn what had happened while you slept. For many people a daily diary was your lifeline, reminding you of the coming day's activities: a morning meeting at the office, dial-in numbers and passcodes for conference calls, the address for your afternoon doctor's appointment, and a list of to-dos including programming the VCR to record your favorite show. Before you left for the day, you might have placed a few phone calls (and often left messages on answering machines), including to remind sitters when to pick up children or confirm dinner plans.

Twenty years ago, for most people in China, an LED alarm clock was probably the sole digital device in your bedroom. A bound personal calendar helped you track the day's appointments, addresses, and phone numbers. After sending your kids off to school, you likely caught up on the world's happenings from a radio broadcast while you ate a quick breakfast of soya milk with Youtiao at your neighborhood restaurant. In 1998, commuters in Beijing buried their noses in newspapers and books – not smartphones and laptops – on the crowded trains and buses traveling to and from the city's centers.

But today, while many of our fundamental morning routines remain the same, a lot has also changed as technology has altered how we go about them. Today a morning in Beijing is still different from a morning in Seattle, but not as different as it used to be. Consider for a moment that in both places the smartphone charging on your bedside table is the device that not only wakes you, but serves up headlines and updates you on your friends' social lives. You read all the email that arrived overnight, text your sister to confirm dinner plans, update the calendar invite to your sitter with details for soccer practice, and then check traffic conditions. Today, in 2018, you can order and pay for a double skinny latte or tea from Starbucks and request a ride-share to drive you to work from that same smartphone.

Compared with the world just 20 years ago, we take a lot of things for granted that used to be the stuff of science fiction. Clearly much can change in just two decades.

Twenty years from now, what will your morning look like? At Microsoft, we imagine a world where your personal digital assistant Cortana talks with your calendar while you sleep. She works with your other smart devices at home to rouse you at the end of a sleep cycle when it's easiest to wake and ensures that you have plenty of time to shower, dress, commute and prepare for your first meeting. As you get ready, Cortana reads the latest news, research reports and social media activity based on your current work, interests and tasks, all of which she gleaned from your calendar, meetings, communications, projects and writings. She updates you on the weather, upcoming meetings, the people you will see, and when you should leave home based on traffic projections.

Acting on the request you made a year before, Cortana also knows that it's your sister's birthday and she's ordered flowers (lilies, your sister's favorite) to be delivered later that day. (Cortana also reminds you about this so that you'll know to say, "you're welcome" when your sister thanks you.) Cortana has also booked a reservation for a restaurant that you both like at a time that's convenient for both of your schedules.

In 2038, digital devices will help us do more with one of our most precious commodities: time.

In 20 years, you might take your first meeting from home by slipping on a HoloLens or other device where you'll meet and interact with your colleagues and clients around a virtual boardroom powered by mixed reality. Your presentation and remarks will be translated automatically into each participant's native language, which they will hear through an earpiece or phone. A digital assistant like Cortana will then automatically prepare a summary of the meeting with tasks assigned to the participants and reminders placed on their schedules based on the conversation that took place and the decisions the participants made.

In 2038, a driverless vehicle will take you to your first meeting while you finalize a presentation on the car's digital hub. Cortana will summarize research and data pulled from newly published articles and reports, creating infographics with the new information for you to review and accept. Based on your instructions, she'll automatically reply to routine emails and reroute those that can be handled by others, which she will request with a due date based on the project timeline. In fact, some of this is already happening today, but two decades from now everyone will take these kinds of capabilities for granted.

Increasingly, we imagine that a smart device will monitor your health vitals. When something is amiss, Cortana will schedule an appointment, and she will also track and schedule routine checkups, vaccines and tests. Your digital assistant will book appointments and reserve time on your

calendar on days that are most convenient. After work a self-driving car will take you home, where you'll join your doctor for a virtual checkup. Your mobile device will take your blood pressure, analyze your blood and oxygen level, and send the results to your doctor, who will analyze the data during your call. Artificial intelligence will help your doctor analyze your results using more than a terabyte of health data, helping her accurately diagnose and prescribe a customized treatment based on your unique physiological traits. Within a few hours, your medication will arrive at your door by drone, which Cortana will remind you to take. Cortana will also monitor your progress and, if you don't improve, she'll ask your permission to book a follow-up appointment with the doctor.

When it's time to take a break from the automated world of the future, you won't call a travel agent or even book online your own flight or hotel as you do today. You'll simply say, "Hey, Cortana, please plan a two-week holiday." She'll propose a custom itinerary based on the season, your budget, availability and interests. You'll then decide where you want to go and stay.

Looking back, it's fascinating to see how technology has transformed the way we live and work over the span of twenty years. Digital technology powered by the cloud has made us smarter and helped us optimize our time, be more productive and communicate with one another more effectively. And this is just the beginning.

Before long, many mundane and repetitive tasks will be handled automatically by AI, freeing us to devote our time and energy to more productive and creative endeavors. More broadly, AI will enable humans to harness vast amounts of data and make breakthrough advances in areas like healthcare, agriculture, education and transportation. We're already seeing how AI-bolstered computing can help doctors reduce medical mistakes, farmers improve yields, teachers customize instruction and researchers unlock solutions to protect our planet.

But as we've seen over the past 20 years, as digital advances bring us daily benefits they also raise a host of complex questions and broad concerns about how technology will affect society. We have seen this as the internet has come of age and become an essential part of our work and private lives. The impact ranges from debates around the dinner table about how distracting our smartphones have become to public deliberations about cybersecurity, privacy, and even the role social media plays in terrorism. This has given birth not just to new public policies and regulations, but to new fields of law and to new ethical considerations in the field of computer science. And this seems certain to continue as AI evolves and the world focuses on the role it will play in society. As we look to the future, it's important that we maintain an open and questioning mind while we seek to take advantage of the opportunities and address the challenges that this new technology creates.

The development of privacy rules over the past two decades provides a good preview of what we might expect to see more broadly in the coming years for issues relating to AI. In 1998, one would have been hard-pressed to find a full-time “privacy lawyer.” This legal discipline was just emerging with the advent of the initial digital privacy laws, perhaps most notably the European Community’s Data Protection Directive, adopted in 1995. But the founding of the International Association of Privacy Professionals, or IAPP, the leading professional organization in the field, was still two years away.

Today, the IAPP has over 20,000 members in 83 countries. Its meetings take place in large convention centers filled with thousands of people. There’s no shortage of topics for IAPP members to discuss, including questions of corporate responsibility and even ethics when it comes to the collection, use, and protection of consumer information. There’s also no lack of work for privacy lawyers now that data protection agencies — the privacy regulators of our age — are operating in over 100 countries. Privacy regulation, a branch of law that barely existed two decades ago, has become one of the defining legal fields of our time.

What will the future bring when it comes to the issues, policies and regulations for artificial intelligence? In computer science, will concerns about the impact of AI mean that the study of ethics will become a requirement for computer programmers and researchers? We believe that’s

a safe bet. Could we see a Hippocratic Oath for coders like we have for doctors? That could make sense. We'll all need to learn together and with a strong commitment to broad societal responsibility. Ultimately the question is not only what computers can do. It's what computers should do.

Similarly, will the future give birth to a new legal field called "AI law"? Today AI law feels a lot like privacy law did in 1998. Some existing laws already apply to AI, especially tort and privacy law, and we're starting to see a few specific new regulations emerge, such as for driverless cars. But AI law doesn't exist as a distinct field. And we're not yet walking into conferences and meeting people who introduce themselves as "AI lawyers." By 2038, it's safe to assume that the situation will be different. Not only will there be AI lawyers practicing AI law, but these lawyers, and virtually all others, will rely on AI itself to assist them with their practice.

The real question is not whether AI law will emerge, but how it can best come together — and over what timeframe. We don't have all the answers, but we're fortunate to work every day with people who are asking the right questions. As they point out, AI technology needs to continue to develop and mature before rules can be crafted to govern it. A consensus then needs to be reached about societal principles and values to govern AI development and use, followed by best practices to live up to them. Then we're likely to be in a better position for governments to create legal and regulatory rules for everyone to follow.

This will take time — more than a couple of years in all likelihood, but almost certainly less than two decades. Already it's possible to start defining six ethical principles that should guide the development and use of artificial intelligence. These principles should ensure that AI systems are fair, reliable and safe, private and secure, inclusive, transparent, and accountable. The more we build a detailed understanding of these or similar principles — and the more technology developers and users can share best practices to implement them — the better served the world will be as we begin to contemplate societal rules to govern AI.

Today, there are some people who might say that ethical principles and best practices are all that is needed as we move forward. They suggest that technology innovation doesn't really need the help of regulators, legislators and lawyers.

While they make some important points, we believe this view is unrealistic and even misguided. AI will be like every technology that has preceded it. It will confer enormous benefits on society. But inevitably, some people will use it to cause harm. Just as the advent of the postal service led criminals to invent mail fraud and the telegraph was followed by wire fraud, the years since 1998 have seen both the adoption of the internet as a tool for progress and the rise of the internet as a new arena for fraud, practiced in increasingly creative and disturbing ways on a global basis.

We must assume that by 2038, we'll grapple with the issues that arise when criminal enterprises and others use AI in ways that are objectionable and even harmful. And undoubtedly other important questions will need to be addressed regarding societally acceptable uses for AI. It will be impossible to address these issues effectively without a new generation of laws. So, while we can't afford to stifle AI technology by adopting laws before we understand the issues that lie ahead of us, neither can we make the mistake of doing nothing now and waiting for two decades before getting started. We need to strike a balance.

As we consider principles, policies and laws to govern AI, we must also pay attention to AI's impact on workers around the globe. What jobs will AI eliminate? What jobs will it create? If there has been one constant over 250 years of technological change, it has been the ongoing impact of technology on jobs — the creation of new jobs, the elimination of existing jobs, and the evolution of job tasks and content. This too is certain to continue with the adoption of AI.

Will AI create more jobs than it will eliminate? Or will it be the other way around? Economic historians have pointed out that each prior industrial revolution created jobs on a net basis. There are many reasons to think this will also be the case with AI, but the truth is that no one has a crystal ball.

It's difficult to predict detailed employment trends with certainty because the impact of new technology on jobs is often indirect and subject to a wide range of interconnected innovations and events. Consider the automobile. One didn't need to be a soothsayer to predict that the adoption of cars

would mean fewer jobs producing horse-drawn carriages and new jobs manufacturing automobile tires. But that was just part of the story.¹

The transition to cars initially contributed to an agricultural depression that affected the entire American economy in the 1920s and 1930s. Why? Because as the horse population declined rapidly, so did the fortunes of American farmers. In the preceding decade roughly a quarter of agricultural output had been used to feed horses. But fewer horses meant less demand for hay, so farmers shifted to other crops, flooding the market and depressing agricultural prices more broadly. This agricultural depression impacted local banks in rural areas, and then this rippled across the entire financial system.

Other indirect effects had a positive economic impact as the sale of automobiles led to the expansion of industry sectors that, at first glance, appear disconnected from cars. One example was a new industry to provide consumer credit. Henry Ford's invention of the assembly line made cars affordable to a great many families, but cars were still expensive and people needed to borrow money to pay for them. As one historian noted, "installment credit and the automobile were both cause and consequence of each other's success."² In short, a new financial services market took flight.

Something similar happened with advertising. As passengers traveled in cars driving 30 miles per hour or more, "a sign had to be grasped instantly or it wouldn't be grasped at all."³ Among other things, this led to the creation of corporate

logos that could be recognized immediately wherever they appeared.

Consider the indirect impact of the automobile on the island of Manhattan alone. The cars driving down Broadway contributed to the creation of new financial jobs on Wall Street and new advertising positions on Madison Avenue. Yet there's little indication that anyone predicted either of these new job categories when cars first appeared on city streets.

One of the lessons for AI and the future is that we'll all need to be alert and agile to the impact of this new technology on jobs. While we can predict generally that new jobs will be created and some existing jobs will disappear, none of us should develop such a strong sense of certainty that we lose the ability to adapt to the surprises that probably await us.

But as we brace ourselves for uncertainty, one thing remains clear. New jobs will require new skills. Indeed, many existing jobs will also require new skills. That is what always happens in the face of technological change.

Consider what we've seen over the past three decades. Today every organization of more than modest size has one or more employees who support its IT, or information technology. Very few of these jobs existed 30 years ago. But it's not just IT staff that needed to acquire IT skills. In the early 1980s, people in offices wrote with a pen on paper, and then secretaries used typewriters to turn that prose into something that was actually legible. By the end of the decade, secretaries learned to use word processing terminals. And then in the 1990s, everyone learned to do their own writing on a PC and

the number of secretaries declined. IT training wasn't just reserved for IT professionals.

In a similar way, we're already seeing increasing demand for new digital and other technical skills, with critical shortages appearing in some disciplines. This is expanding beyond coding and computer science to data science and other fields that are growing in importance as we enter the world's Fourth Industrial Revolution. More and more, this isn't just a question of encouraging people to learn new skills, but of finding new ways to help them acquire the skills they will need. Surveys of parents show that they overwhelmingly want their children to have the opportunity to learn to code. And at Microsoft, when we offer our employees new courses on the latest AI advances, demand is always extremely high.

The biggest challenges involve the creation of ways to help people learn new skills, and then rethinking how the labor market operates to enable employers and employees to move in more agile ways to fill new positions. The good news is that many communities and countries have developed new innovations to address this issue, and there are opportunities to learn from these emerging practices. Some are new approaches to longstanding programs, like Switzerland's successful youth apprenticeships. Others are more recent innovations spurred by entities such as LinkedIn and its online tools and services and nonprofit ventures like the Markle Foundation's Skillful initiative in Colorado.

The impact of AI, the cloud and other new technologies won't stop there. A few decades ago, workers in many countries mostly enjoyed traditional employer-employee

relationships and worked in offices or manufacturing facilities. Technology has helped upend this model as more workers engage in alternative work arrangements through remote and part-time work, as contractors or through project-based engagements. And most studies suggest that these trends will continue.

For AI and other technologies to benefit people as broadly as possible, we'll need to adapt employment laws and labor policies to address these new realities. Many of our current labor laws were adopted in response to the innovations of the early 20th century. Now, a century later, they're no longer suited to the needs of either workers or employers. For example, employment laws in most countries assume that everyone is either a full-time employee or an independent contractor, making no room for people who work in the new economy for Uber, Lyft or other similar services that are emerging in every field from tech support to caregiving.

Similarly, health insurance and other benefits were designed for full-time employees who remain with a single employer for many years. But they aren't as effective for individuals who work for multiple companies simultaneously or change jobs more frequently. Our social safety net — including the United States' Social Security system — is a product of the first half of the last century. There is an increasingly pressing need to adapt these vital public policies to the world that is changing today.

As we all think about the future, the pace of change can feel more than a little daunting. By looking back to technology in 1998, we can readily appreciate how much change we've lived through already. Looking ahead to 2038, we can begin to anticipate the rapid changes that lie ahead — changes that will create opportunities and challenges for communities and countries around the world.

For us, some key conclusions emerge.

First, the companies and countries that will fare best in the AI era will be those that embrace these changes rapidly and effectively. The reason is straightforward: AI will be useful wherever intelligence is useful, helping us to be more productive in nearly every field of human endeavor and leading to economic growth. Put simply, new jobs and economic growth will accrue to those that embrace the technology, not those that resist it.

Second, while we believe that AI will help improve daily life in many ways and help solve big societal problems, we can't afford to look to this future with uncritical eyes. There will be challenges as well as opportunities. This is why we need to think beyond the technology itself to address the need for strong ethical principles, the evolution of laws, the importance of training for new skills, and even labor market reforms. This must all come together if we're going to make the most of this new technology.

Third, we need to address these issues together with a sense of shared responsibility. In part this is because AI technology won't be created by the tech sector alone. At Microsoft we're working to "democratize AI" in a manner that's similar to the way we "democratized the PC." Just as our work that started in the 1970s enabled organizations across society to create their own custom applications for the PC, the same thing will happen with AI. Our approach to AI is making the fundamental AI building blocks like computer vision, speech, and knowledge recognition available to every individual and organization to build their own AI-based solutions. We believe this is far preferable to having only a few companies control the future of AI. But just as this will spread broadly the opportunity for others to create AI-based systems, it will spread broadly the shared responsibility needed to address AI issues and their implications.

As technology evolves so quickly, those of us who create AI, cloud and other innovations will know more than anyone else how these technologies work. But that doesn't necessarily mean that we will know how best to address the role they should play in society. This requires that people in government, academia, business, civil society, and other interested stakeholders come together to help shape this future. And increasingly we need to do this not just in a single community or country, but on a global basis. Each of us has a responsibility to participate — and an important role to play.

All of this leads us to what may be one of the most important conclusions of all. We're reminded of something that Steve Jobs famously talked about repeatedly: he always sought to work at the intersection of engineering and the liberal arts.

One of us grew up learning computer science and the other started in the liberal arts. Having worked together for many years at Microsoft, it's clear to both of us that it will be even more important to connect these fields in the future.

At one level, AI will require that even more people specialize in digital skills and data science. But skilling-up for an AI-powered world involves more than science, technology, engineering and math. As computers behave more like humans, the social sciences and humanities will become even more important. Languages, art, history, economics, ethics, philosophy, psychology and human development courses can teach critical, philosophical and ethics-based skills that will be instrumental in the development and management of AI solutions. If AI is to reach its potential in serving humans, then every engineer will need to learn more about the liberal arts and every liberal arts major will need to learn more about engineering.

We're all going to need to spend more time talking with, listening to, and learning from each other. As two people from different disciplines who've benefited from doing just that, we appreciate firsthand the valuable and even enjoyable opportunities this can create.

We hope that the pages that follow can help as we all get started.

Brad Smith President and Chief Legal Officer

Harry Shum Executive Vice President, Artificial Intelligence and Research
Microsoft Corporation

1. See Brad Smith and Carol Ann Browne, "Today in Technology: The Day the Horse Lost its Job," at <https://www.linkedin.com/pulse/today-technology-day-horse-lost-its-job-brad-smith/>

2. Lendol Calder, *Financing the American Dream: A Cultural History of Consumer Credit* (Princeton: Princeton University Press, 1999), p. 184.

3. John Steele Gordon, *An Empire of Wealth: The Epic History of American Economic Power* (New York: HarperCollins Publishers, 2004), p. 299-300.



Brad Smith



Harry Shum

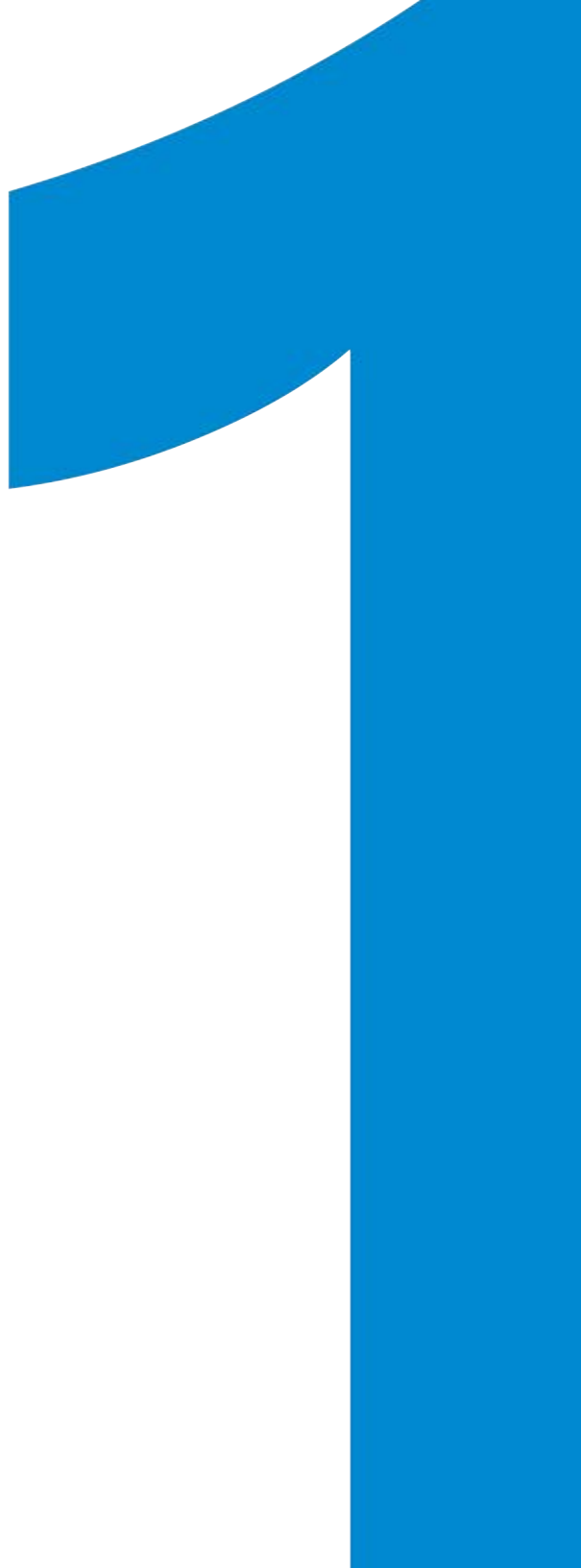
ACKNOWLEDGEMENTS

We would like to thank the following contributors for providing their insights and perspectives in the development of this book.

Benedikt Abendroth, Geff Brown, Carol Ann Browne, Dominic Carr, Pablo Chavez, Steve Clayton, Amy Colando, Jane Broom Davidson, Mariko Davidson, Paul Estes, John Galligan, Sue Glueck, Cristin Goodwin, Mary Gray, David Heiner, Merisa Heu-Weller, Eric Horvitz, Teresa Hutson, Nicole Isaac, Lucas Joppa, Aaron Kleiner, Allyson Knox, Cornelia Kutterer, Jenny Lay-Flurrie, Andrew Marshall, Anne Nergaard, Carolyn Nguyen, Barbara Olagaray, Michael Philips, Brent Sanders, Mary Snapp, Dev Stahlkopf, Steve Sweetman, Lisa Tanzi, Ana White, Joe Whittinghill, Joshua Winter, Portia Wu

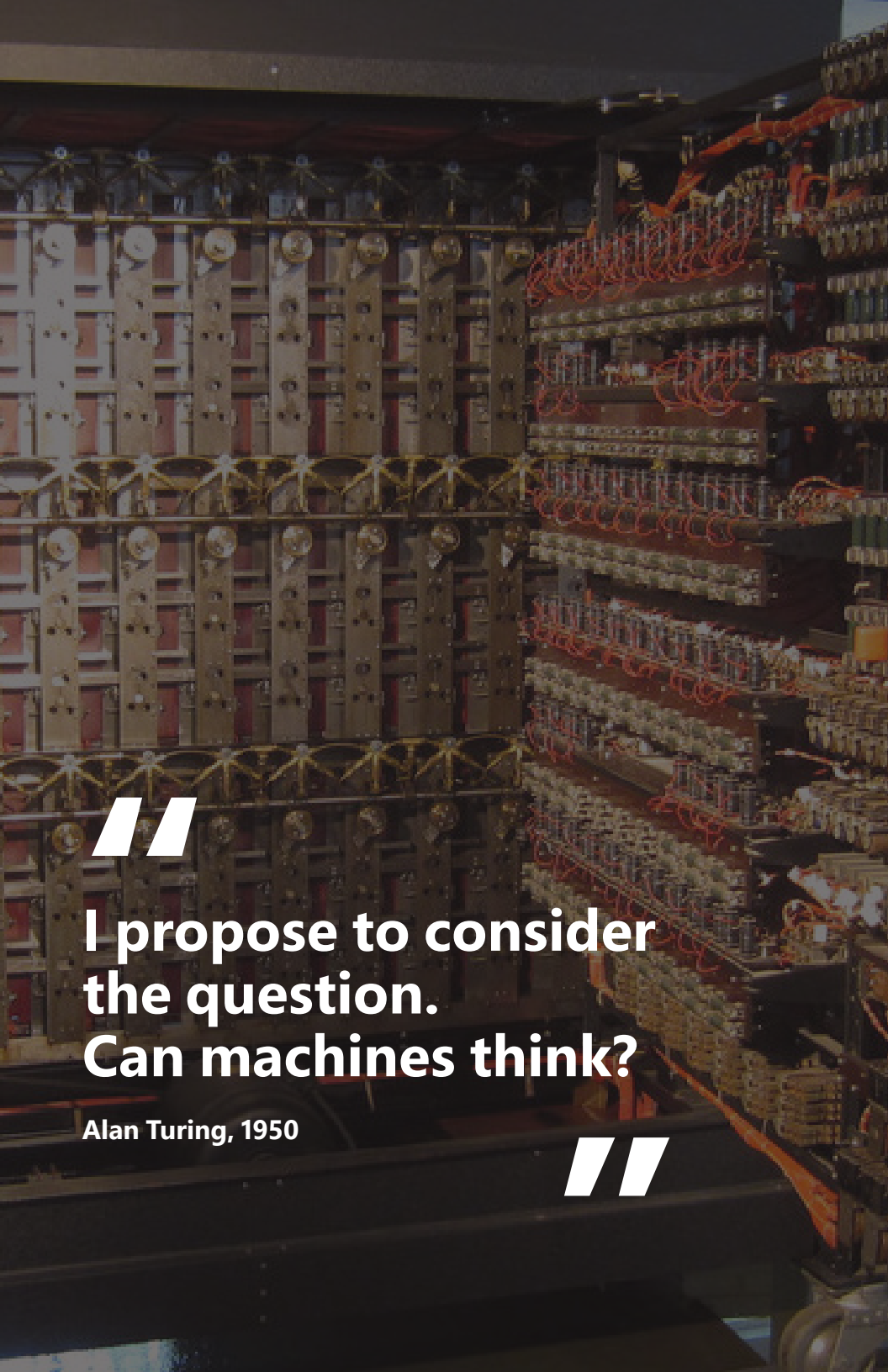
Chapter 1

The Future of Artificial Intelligence









//

**I propose to consider
the question.
Can machines think?**

Alan Turing, 1950

//

In the summer of 1956, a team of researchers at Dartmouth College met to explore the development of computer systems capable of learning from experience, much as people do. But, even this seminal moment in the development of AI was preceded by more than a decade of exploration of the notion of machine intelligence, exemplified by Alan Turing's quintessential test: a machine could be considered "intelligent" if a person interacting with it (by text in those days) could not tell whether it was a human or a computer.

Researchers have been advancing the state of the art in AI in the decades since the Dartmouth conference. Developments in subdisciplines such as machine vision, natural language understanding, reasoning, planning and robotics have produced an ongoing stream of innovations, many of which have already become part of our daily lives. Route-planning features in navigation systems, search engines that retrieve and rank content from the vast amounts of information on the internet, and machine vision capabilities that enable postal services to automatically recognize and route handwritten addresses are all enabled by AI.

At Microsoft, we think of AI as a set of technologies that enable computers to perceive, learn, reason and assist in decision-making to solve problems in ways that are similar to what people do. With these capabilities, how computers understand and interact with the world is beginning to feel far more natural and responsive than in the past, when computers could only follow pre-programmed routines.

Not so long ago we interacted with computers via a command line interface. And while the graphical user interface was an important step forward, we will soon be routinely interacting with computers just by talking to them, just as we would to a person. To enable these new capabilities, we are, in effect, teaching computers to see, hear, understand and reason.¹ Key technologies include:

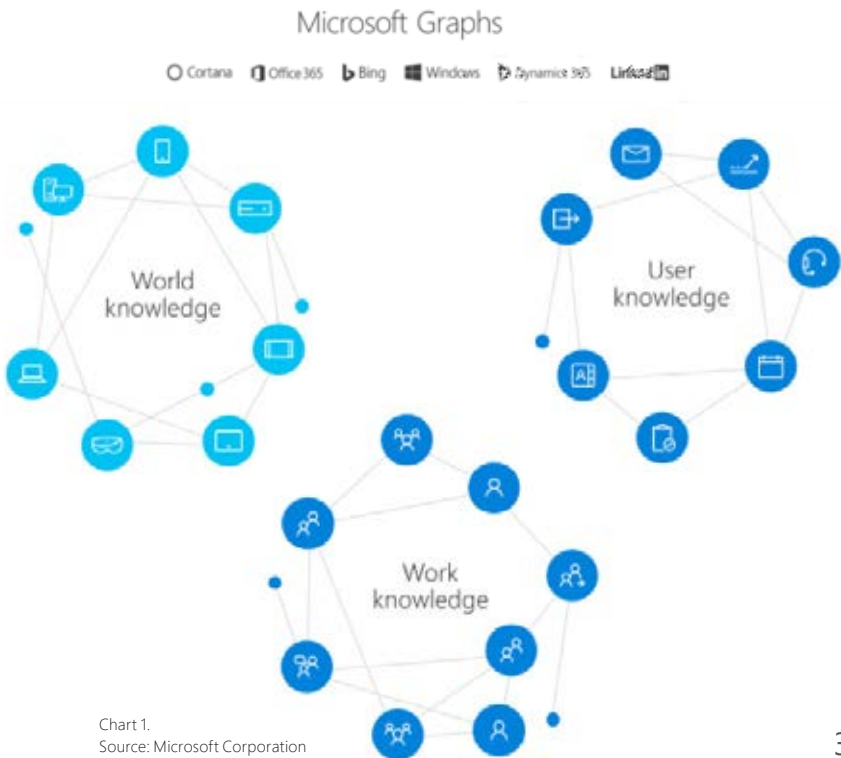
Vision: the ability of computers to “see” by recognizing what is in a picture or video.

Speech: the ability of computers to “listen” by understanding the words that people say and to transcribe them into text.

Language: the ability of computers to “comprehend” the meaning of the words, taking into account the many nuances and complexities of language (such as slang and idiomatic expressions).

Knowledge: the ability of a computer to “reason” by understanding the relationship between people, things, places, events and the like. For instance, when a search result for a movie provides information about the cast and other movies those actors were in, or at work when you participate in a meeting and the last several documents that you shared with the person you’re meeting with are automatically delivered to you. These are examples of a computer reasoning by drawing conclusions about which information is related to other information.

Computers are learning the way people do; namely, through experience. For computers, experience is captured in the form of data. In predicting how bad traffic will be, for example, computers draw upon data regarding historical traffic flows based on the time of day, seasonal variations, the weather, and major events in the area such as concerts or sporting events. More broadly, rich “graphs” of information are foundational to enabling computers to develop an understanding of relevant relationships and interactions between people, entities and events. In developing AI systems, Microsoft is drawing upon graphs of information that include knowledge about the world, about work and about people.



Thanks in part to the availability of much more data, researchers have made important strides in these technologies in the past few years. In 2015, researchers at Microsoft announced that they had taught computers to identify objects in a photograph or video as accurately as people do in a test using the standard ImageNet 1K database of images.² In 2017, Microsoft's researchers announced they had developed a speech recognition system that understood spoken words as accurately as a team of professional transcribers, with an error rate of just 5.1 percent using the standard Switchboard dataset.³ In essence, AI-enhanced computers can, in most cases, see and hear as accurately as humans.

Much work remains to be done to make these innovations applicable to everyday use. Computers still may have a hard time understanding speech in a noisy environment where people speak over one another or when presented with unfamiliar accents or languages. It is especially challenging to teach computers to truly understand not just what words were spoken, but what the words mean and to reason by drawing conclusions and making decisions based on them. To enable computers to comprehend meaning and answer more complex questions, we need to take a big-picture view, understand and evaluate context, and bring in background knowledge.

Why Now?

Researchers have been working on AI for decades. Progress has accelerated over the past few years thanks in large part to three developments: the increased availability of data; growing cloud computing power; and more powerful algorithms developed by AI researchers.

As our lives have become increasingly digitized and sensors have become cheap and ubiquitous, more data than ever before is available for computers to learn from.

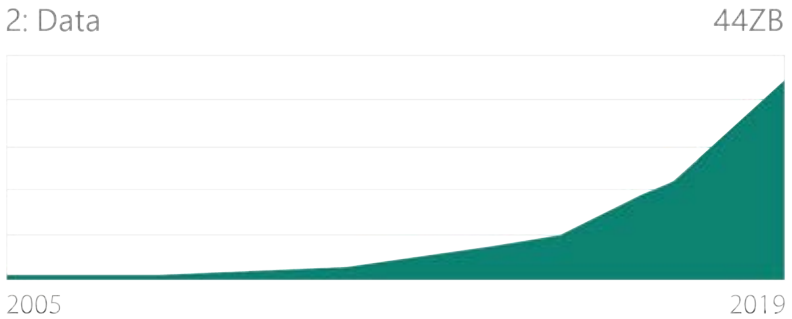


Chart 2.
Source: IDC Digital Universe Forecast, 2014

Only with data can computers discern the patterns, often subtle, that enable them to “see,” “hear” and “understand.”

Analyzing all this data requires massive computing power, which is available thanks to the efficiencies of cloud computing. Today, organizations of any type can tap into the power of the cloud to develop and run their AI systems.

Researchers at Microsoft, other technology firms, universities and governments have drawn upon this combination of the availability of this data, and with it ready access to powerful computing and breakthroughs in AI techniques — such as “deep learning” using so-called “deep neural nets”— to enable computers to mimic how people learn.

In many ways, AI is still maturing as a technology. Most of the progress to date has been in teaching computers to perform narrow tasks — play a game, recognize an image, predict traffic. We have a long way to go to imbue computers with “general” intelligence. Today’s AI cannot yet begin to compete with a child’s ability to understand and interact with the world using senses such as touch, sight and smell. And AI systems have only the most rudimentary ability to understand human expression, tone, emotion and the subtleties of human interaction. In other words, AI today is strong on “IQ” but weak on “EQ.”

At Microsoft, we’re working toward endowing computers with more nuanced capabilities. We believe an integrated approach that combines various AI disciplines will lead to the development of more sophisticated tools that can help people perform more complex, multifaceted tasks. Then, as we learn how to combine multiple IQ functions with abilities that come naturally to people — like applying knowledge of one task to another, having a commonsense understanding of the world, interacting naturally, or knowing when someone

is trying to be funny or sarcastic, and the difference between those — AI will become even more helpful. While this is clearly a formidable challenge, when machines can integrate the smarts of IQ and the empathy of EQ in their interactions, we will have achieved what we call “conversational AI.” This will be an important step forward in the evolution of computer-human interaction.

Microsoft’s Approach to AI

When Bill Gates and Paul Allen founded Microsoft over 40 years ago, their aim was to bring the benefits of computing — then largely locked up in mainframes — to everyone. They set out to build a “personal” computer that would help people be more productive at home, at school and at work. Today, Microsoft is aiming to do much the same with AI. We’re building AI systems that are designed to amplify natural human ingenuity. We’re deploying AI systems with the goal of making them available to everyone and aspiring to build AI systems that reflect timeless societal values so that AI earns the trust of all.⁴

Amplifying Human Ingenuity

We believe that AI offers incredible opportunities to drive widespread economic and social progress. The key to attaining these benefits is to develop AI in such a way that it is human-centered. Put simply, we aim to develop AI in order to augment human abilities, especially humankind's innate ingenuity. We want to combine the capabilities of computers with human capabilities to enable people to achieve more.

Computers are very good at remembering things. Absent a system failure, computers never forget. Computers are very good at probabilistic reasoning, something many people are not so good at. Computers are very good at discerning patterns in data that are too subtle for people to notice. With these capabilities, computers can help us make better decisions. And this is a real benefit, because, as researchers in cognitive psychology have established, human decision-making is often imperfect. Broadly speaking, the kind of “computational intelligence” that computers can provide will have a significant impact in almost any field where intelligence itself has a role to play.



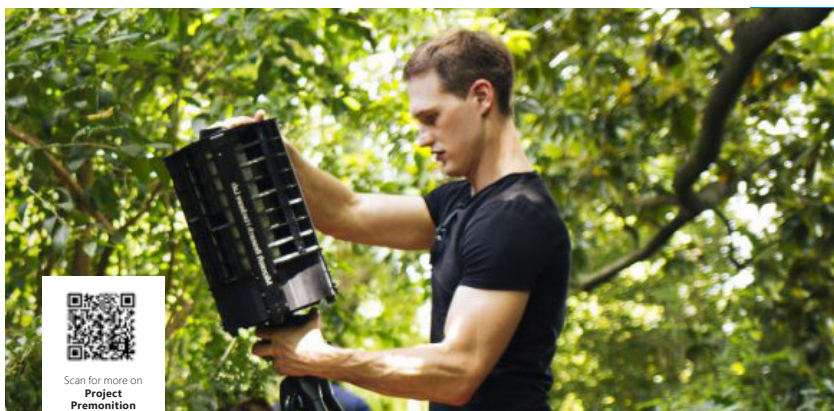
AI improving medical image analysis for clinicians

AI systems are already helping people tackle big problems. A good example of this is “InnerEye,” a project in which U.K.-based researchers at Microsoft have teamed up with oncologists to develop an AI system to help treat cancer more effectively.⁵

InnerEye uses AI technology originally developed for video gameplay to analyze computed tomography (CT) and magnetic resonance imaging (MRI), and helps oncologists target cancer treatment more quickly. CT and MRI scans allow doctors to look inside a patient’s body in three dimensions and study anomalies, such as tumors. For cancer patients who are undergoing radiation therapy, oncologists use such scans to delineate tumors from the surrounding

healthy tissue, bone and organs. In turn, this helps focus the cell-damaging radiation treatment on the tumor while avoiding healthy anatomy as much as possible. Today, this 3-D delineation task is manual, slow and error-prone. It requires a radiation oncologist to draw contours on hundreds of cross-sectional images by hand, one at a time — a process that can take hours. InnerEye is being designed to accomplish the same task in a fraction of that time, while giving oncologists full control over the accuracy of the final delineation.

To create InnerEye's automatic segmentation, researchers used hundreds of raw CT and MRI scans (with all identifying patient information removed). The scans were fed into an AI system that learned to recognize tumors and healthy anatomical structures with a clinical level of accuracy. As part of the process, once the InnerEye automatic segmentation is complete, the oncologist goes in to fine-tune the contours. The doctor is in control at all times. With further advances, InnerEye may be helpful for measuring and tracking tumor changes over time, and even assessing whether a treatment is working.



AI helping researchers prevent disease outbreaks

Another interesting example is “Project Premonition.” We’ve all seen the heartbreaking stories of lives lost in recent years to dangerous diseases like Zika, Ebola and dengue that are transmitted from animals and insects to people. Today, epidemiologists often don’t learn about the emergence of these pathogens until an outbreak is underway. But this project — developed by scientists and engineers at Microsoft Research, the University of Pittsburgh, the University of California Riverside and Vanderbilt University — is exploring ways to detect pathogens in the environment so public health officials can protect people from transmission before an outbreak begins.⁶

What epidemiologists need are sensors that can detect when pathogens are present. The researchers on this project hit

upon an ingenious idea: why not use mosquitoes as sensors? There are plenty of them and they feed on a wide range of animals, extracting a small amount of blood that contains genetic information about the animal bitten and pathogens circulating in the environment.

The researchers use advanced autonomous drones capable of navigating through complex environments to identify areas where mosquitoes breed. They then deploy robotic traps that can distinguish between the types of mosquitoes researchers want to collect and other insects, based on wing movement patterns. Once specimens are collected, cloud-scale genomics and advanced AI systems identify the animals that the mosquitoes have fed on and the pathogens that the animals carry. In the past, this kind of genetic analysis could take a month; now the AI capabilities of Project Premonition have shortened that to about 12 hours.

During a Zika outbreak in 2016, Project Premonition drones and traps were tested in Houston. More than 20,000 mosquitoes were collected from nine different species, including those known to carry Zika, dengue, West Nile virus and malaria. Because the traps also gather data on environmental conditions when an insect is collected, the test provided useful data not only about pathogens in the environment but also about mosquito behavior. This helped Project Premonition researchers improve their ability to target hotspots where mosquitoes breed. Researchers are also working to improve how to identify known diseases and detect the presence of previously unknown pathogens.

While the project is still in its early stages, it may well point the way toward an effective early warning system that will detect some of the world's most dangerous diseases in the environment and help prevent deadly outbreaks.

Making Human-Centered AI Available to All

We cannot deliver on the promise of AI unless we make it broadly available to all. People around the world can benefit from AI — but only if AI technologies are available for them. For Microsoft, this begins with basic R&D. Microsoft Research, with its 26-year history, has established itself as one of the premier research organizations in the world contributing both to the advancement of computer science and to Microsoft products and services. Our researchers have published more than 22,000 papers in all areas of study — from the environment to health, and from privacy to security. Recently, we announced the creation of Microsoft Artificial Intelligence and Research, a new group that brings together approximately 7,500 computer scientists, researchers and engineers. This group is chartered with pursuing a deeper understanding of the computational foundations of intelligence, and is focused on integrating research from all fields of AI research in order to solve some of AI's most difficult challenges.

We continue to encourage researchers to publish their results broadly so that AI researchers around the world — at universities, at other companies and in government settings — can build on these advances.

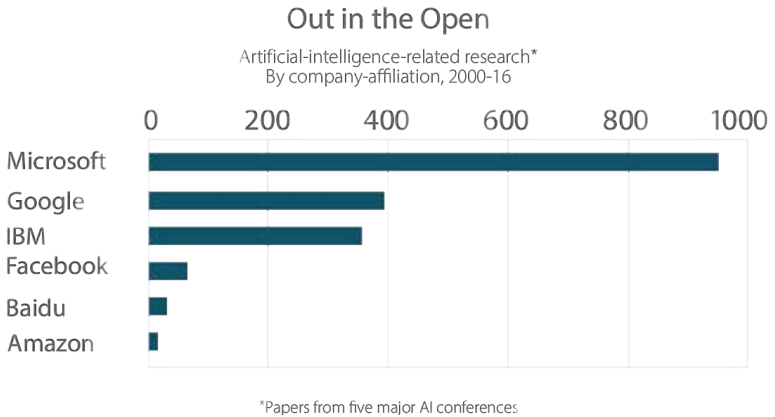


Chart 3.
Source: The Economist

For our customers, we’re building AI capabilities into our most popular products, such as Windows and Office. Windows is more secure thanks to AI systems that detect malware and automatically protect computers against it. In Office, Researcher for Word helps you write more compelling documents. Without leaving a document, you can find and incorporate relevant information from across the web using Bing “Knowledge Graph.” If you are creating a PowerPoint presentation, PowerPoint Designer assesses the images and text you’ve used, and provides design tips to create more professional-looking slides, along with suggestions for text captions for images to improve accessibility. And PowerPoint Presentation Translator lets you engage diverse audiences more effectively by breaking down language barriers through auto-captioning in over 60 languages. This feature will also aid people with hearing loss.

AI is the enabling technology behind Cortana, Microsoft’s personal digital assistant. Cortana is young, but she’s learning fast. Already Cortana can help you schedule a meeting, make a restaurant reservation and find answers to questions on

a broad range of topics. Over time, Cortana will be able to interact with other personal digital assistants to automatically handle tasks that take up time and follow familiar patterns. One of the key technologies that Cortana builds upon is Bing, our search service. But instead of just providing links to relevant information, Cortana uses Bing to discover answers to your questions and provide them in a variety of more context-rich ways.⁷

Microsoft is not only using AI technologies to create and enhance our own products, we are also making them available to developers so that they can build their own AI-powered products. The Microsoft AI Platform offers services, tools and infrastructure making AI development easier for developers and organizations of any size. Our service offerings include Microsoft Cognitive Services, a set of pre-built AI capabilities including vision, speech, language and search. All of these are hosted in the cloud and can be easily integrated into applications. Some of these are also customizable so that they can be better optimized to help transform and improve business processes specific to an organization's industry and business needs. You can see the breadth of these offerings below.

Microsoft Cognitive Services



Chart 4.
Source: Microsoft Corporation

We also have technologies available to simplify the creation of “bots” that can engage with people more naturally and conversationally. We offer a growing collection of coding and management tools to make the AI development process easier. And our infrastructure offerings help others develop and deploy algorithms, and store their data and derive insights from it.

Finally, with Microsoft’s AI Business Solutions, we are building systems of intelligence so organizations can better understand and act on the information they collect in order to be more productive.

One example of an AI Business Solution is Customer Care Intelligence, currently being used by the Department of Human Services (DHS) in Australia to transform how it delivers services to citizens. At the heart of the program is

an expert system that uses a virtual assistant named “Roxy” who helps claims processing officers answer questions and solve problems. Roxy was trained using the DHS operational blueprint that includes all of the agency’s policies and procedures, and fed all of the questions that passed between claims officers and DHS managers over a three-month period. In early use, the system was able to answer nearly 80 percent of the questions it was asked. This is expected to translate to about a 20 percent reduction in workload for claims officers.

The internal project with Roxy was so successful that DHS is now developing virtual assistants that will interact directly with citizens. One of these projects will target high school seniors to help them decide whether to apply for a university or enroll in a vocational program through Australia’s Technical and Further Education program by helping them navigate the qualification process.

The Potential of Modern AI – Addressing Societal Challenges

At Microsoft, we aim to develop AI systems that will enable people worldwide to more effectively address local and global challenges, and to help drive progress and economic opportunity.

Today's AI enables faster and more profound progress in nearly every field of human endeavor, and it is essential to enabling the digital transformation that is at the heart of worldwide economic development. Every aspect of a business or organization — from engaging with customers to transforming products, optimizing operations and empowering employees — can benefit from this digital transformation.

But even more importantly, AI has the potential to help society overcome some of its most daunting challenges. Think of the most complex and pressing issues that humanity faces: from reducing poverty and improving education, to delivering healthcare and eradicating diseases, addressing sustainability challenges such as growing enough food to feed our fast-growing global population through to advancing inclusion in our society. Then imagine what it would mean in lives saved, suffering alleviated and human potential unleashed if we could harness AI to help us find solutions to these challenges.

Providing effective healthcare at a reasonable cost to the approximately 7.5 billion people on the planet is one of society's most pressing challenges. Whether it's analyzing massive amounts of patient data to uncover hidden patterns that can point the way toward better treatments, identifying compounds that show promise as new drugs or vaccines, or unlocking the potential of personal medicine based on

in-depth genetic analysis, AI offers vast opportunities to transform how we understand disease and improve health. Machine reading can help doctors quickly find important information amid thousands of documents that they otherwise wouldn't have time to read. By doing so, it can help medical professionals spend more of their time on higher value and potentially lifesaving work.

Providing safe and efficient transportation is another critical challenge where AI can play an important role. AI-controlled driverless vehicles could reduce traffic accidents and expand the capacity of existing road infrastructure, saving hundreds of thousands of lives every year while improving traffic flow and reducing carbon emissions. These vehicles will also facilitate greater inclusiveness in society by enhancing the independence of those who otherwise are not able to drive themselves.

In education, the ability to analyze how people acquire knowledge and then use that information to develop predictive models for engagement and comprehension points the way toward new approaches to education that combine online and teacher-led instruction and may revolutionize how people learn.

As demonstrated by Australia's Department of Human Services' use of the natural language capabilities of Customer Care Intelligence to answer questions, AI also has the potential to improve how governments interact with their citizens and deliver services.



AI enabling people with low vision to hear information about the world around them

Another area where AI has the potential to have a significant positive impact is in serving the more than 1 billion people in the world with disabilities. One example of how AI can make a difference is a recent Microsoft offering called “Seeing AI,” available in the iOS app store, that can assist people with blindness and low vision as they navigate daily life.

Seeing AI was developed by a team that included a Microsoft engineer who lost his sight at 7 years of age. This powerful app, while still in its early stages, demonstrates the potential for AI to empower people with disabilities by capturing images from the user’s surroundings and instantly describing what is happening. For example, it can read signs and menus, recognize products through barcodes, interpret handwriting, count currency, describe scenes and objects in the vicinity, or, during a meeting, tell the user that there is a man and a woman sitting across the table who are smiling and paying close attention.⁸



AI empowering farmers to be more productive and increase their yield

And with the world's population expected to grow by nearly 2.5 billion people over the next quarter century, AI offers significant opportunities to increase food production by improving agricultural yield and reducing waste. For example, our “FarmBeats” project uses advanced technology, existing connectivity infrastructure, and the power of the cloud and machine learning to enable data-driven farming at low cost. This initiative provides farmers with easily interpretable insights to help them improve agricultural yield, lower overall costs and reduce the environmental impact of farming.⁹

Given the significant benefits that stem from using AI — empowering us all to accomplish more by being more productive and efficient, driving better business outcomes, delivering more effective government services and helping

to solve difficult societal issues — it's vital that everyone has the opportunity to use it. Making AI available to all people and organizations is foundational to enabling everyone to capitalize on the opportunities AI presents and share in the benefits it delivers.

The Challenges AI Presents

As with the great advances of the past on which it builds — including electricity, the telephone and transistors — AI will bring about vast changes, some of which are hard to imagine today. And, as was the case with these previous significant technological advances, we'll need to be thoughtful about how we address the societal issues that these changes bring about. Most importantly, we all need to work together to ensure that AI is developed in a responsible manner so that people will trust it and deploy it broadly, both to increase business and personal productivity and to help solve societal problems.

This will require a shared understanding of the ethical and societal implication of these new technologies. This, in turn, will help pave the way toward a common framework of principles to guide researchers and developers as they deliver a new generation of AI-enabled systems and capabilities, and governments as they consider a new generation of rules and regulations to protect the safety and privacy of citizens and ensure that the benefits of AI are broadly accessible.

In Chapter 2, we offer our initial thinking on how to move forward in a way that respects universal values and addresses the full range of societal issues that AI will raise, while ensuring that we achieve the full potential of AI to create opportunities and improve lives.

Investigations in an Era of AI

I



Investigations in the Era of A.I.

Serge Jorgensen, Sylint Group

David K. A. Mordecai, PhD, Risk Economics Inc

Paul Starrett, Starrett Consulting Inc

Panel Discussion Topics



Investigative Process in the age
of AI and Robotics



Using AI to Detect & Investigate



Admissibility of AI-related
evidence at hearings or in trial

What is the
Investigation
Process in the
age of AI &
Robotics?

Metadata

Evidence identification &
collection

Generative AI systems and
artifacts



Blood Alcohol Content Testing

The ensuing dispute about the disclosure of the software used to operate the device, called firmware, and the source codes needed for an analysis of that software, caused significant disruption in the orderly completion of the proceedings and eventually led to our further remand for additional proceedings.

State v. Chun, 943 A.2d 114



Robot Data Identification, Collection & Analysis

...framework that uses autonomous mobile indoor robots for gathering actionable building information in real-time, and discusses how this information can be further utilized for various analyses and critical decision-making.

Robotic Data Collection
& Simulation for
Evaluation of Building
Retrofit Performance

Using A.I. to Detect & Investigate Incidents

01


Discovering Corrupt Practices

02

Tracking Insider Trading

03

Detecting Corporate Espionage and Trade Secret Misappropriation



Prior Offenses

2 armed robberies, 1
attempted armed
robbery

Subsequent Offenses

1 grand theft

Prior Offenses

4 juvenile
misdemeanors

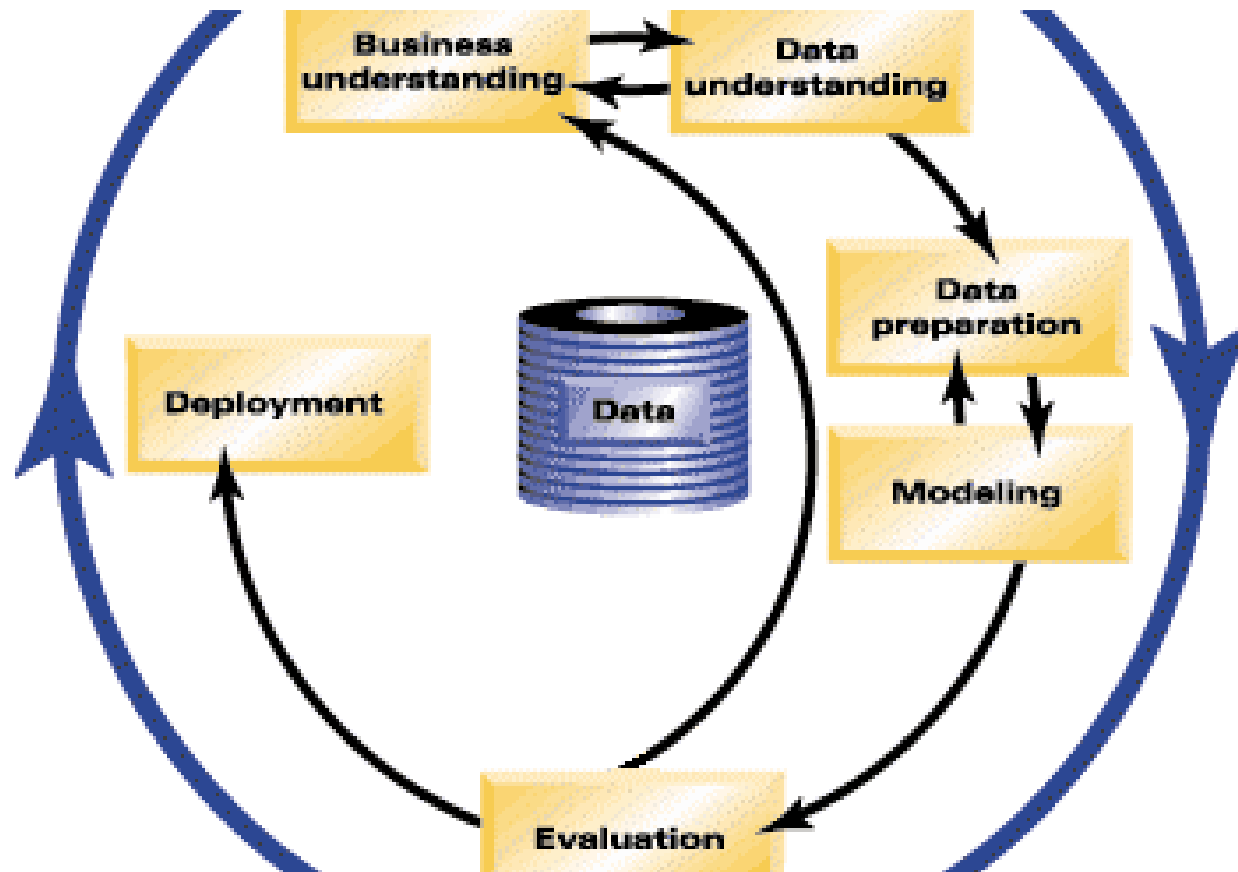
Subsequent Offenses

None

The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Laws, Decisions & Challenges



CRISP DM Cross-Industry Standard Process for Data Mining

When an algorithm prescreens resumes and, not by intentional design, discounts the resumes of women or minorities, is the employer liable for discrimination?

Machine Learning
Evidence: Admissibility &
Weight (Nutter)

Admissibility of A.I.-Related Evidence

Existing Laws
& Expectations

Challenges &
Considerations

Training Sets,
Algorithms and
Statistics



Admissibility, Statistics & Methodologies

some machine evidence, like human testimony, depends on the credibility of a source

<https://www.law.berkeley.edu/spotlight/machine-testimony/>



Whew!

Serge Jorgensen, Sylint Group

David K. A. Mordecai, PhD, Risk Economics Inc

Paul Starrett, Starrett Consulting Inc

Investigations in an Era of AI:

Online Resources

Compas Recidivism Algo Bias

<https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say>

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.propublica.org/series/machine-bias/p3>

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

<https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/>

<https://towardsdatascience.com/mitigating-algorithmic-bias-in-predictive-justice-ux-design-principles-for-ai-fairness-machine-learning-d2227ce28099>

Prescriptive Actions

<https://ieeexplore.ieee.org/ielx7/44/7947239/07947257.pdf?tp=&arnumber=7947257&isnumber=7947239&ref=aHR0cDovL3NjaG9sYXluZ29vZ2xLMmNhLw==>

Algo Bias in Autonomous Systems

https://www.researchgate.net/profile/Alex_London/publication/318830422_Algorithmic_Bias_in_Autonomous_Systems/links/5a4bb017aca2729b7c893d1b/Algorithmic-Bias-in-Autonomous-Systems.pdf

Legal Scholarship for Machine Testimony

<https://www.law.berkeley.edu/spotlight/machine-testimony/>

<https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=5810&context=yjl>

<https://georgetownlawtechreview.org/wp-content/uploads/2019/01/3.1-Sites-pp-1-27.pdf>

<https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1691&context=jcl>

Breathalyzer

<https://www.nytimes.com/2019/11/03/business/breathalyzer-investigation-takeaways.html?smid=nytcore-ios-share>

<https://www.nytimes.com/2019/11/03/business/drunk-driving-breathalyzer.html?smid=nytcore-ios-share>

<https://www.nytimes.com/2019/11/01/the-weekly/breathalyzer-drunk-driving.html?smid=nytcore-ios-share>

Established Scientific Principles of Evidence

<https://www.sciencemag.org/news/2016/03/reversing-legacy-junk-science-courtroom>

<https://www.sciencemag.org/topic/forensics>

<https://www.pnas.org/content/pnas/115/18/4541.full.pdf>

<https://www.jstor.org/stable/41307115>

<https://www.innocenceproject.org/lasting-impact-of-2009-nas-report/>

<https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf> (e.g., See chapters 3, 4 & 10)

<https://www.nationalacademies.org/includes/SciMat.pdf> (e.g., p. 34-35, 36, 72-119 with particular emphasis on the fingerprint image discussion; sections beginning 211, 303, ~811, 897)

<https://www.nature.com/articles/s41586-019-1138-y>

<https://towardsdatascience.com/what-is-machine-behavior-35a97e8ed4c9>

<https://www.ncbi.nlm.nih.gov/pubmed/31019318>

https://spiral.imperial.ac.uk:8443/bitstream/10044/1/70375/4/20190225_Machine%20behaviour%20revised.pdf

AI and Robotics Intellectual Property and Licensing Issues

J

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

Promoting the Progress through AI

January 10, 2020



THE PREMIER SOURCE FOR CLE

“Congress shall have the power... to promote the **progress** of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries.”

US Const, Article I, Section 8, Clause 8



Patents: protects new, useful and nonobvious inventions that are “patent-eligible.” Must be applied for and granted by the Patent Office. ex: computer chip.

Copyrights: protect original works of authorship, e.g. writings or expressions, but not facts or ideas. Arises automatically. example: software on the chip.

Trade Secret: protects information, e.g. a formula, pattern, compilation, that is economically valuable because it has been kept secret. ex: chip customer list.

Protections for Data: generally secured through the above, contract

Intellectual Property Rights are:

- protections over nonrival intangible property
- theorized to incentivize innovation, make transacting in intangible assets, and solve the Arrow information paradox
- negative, not positive rights, they give the right to exclude others from using the protected work

Intellectual Property Rights are:

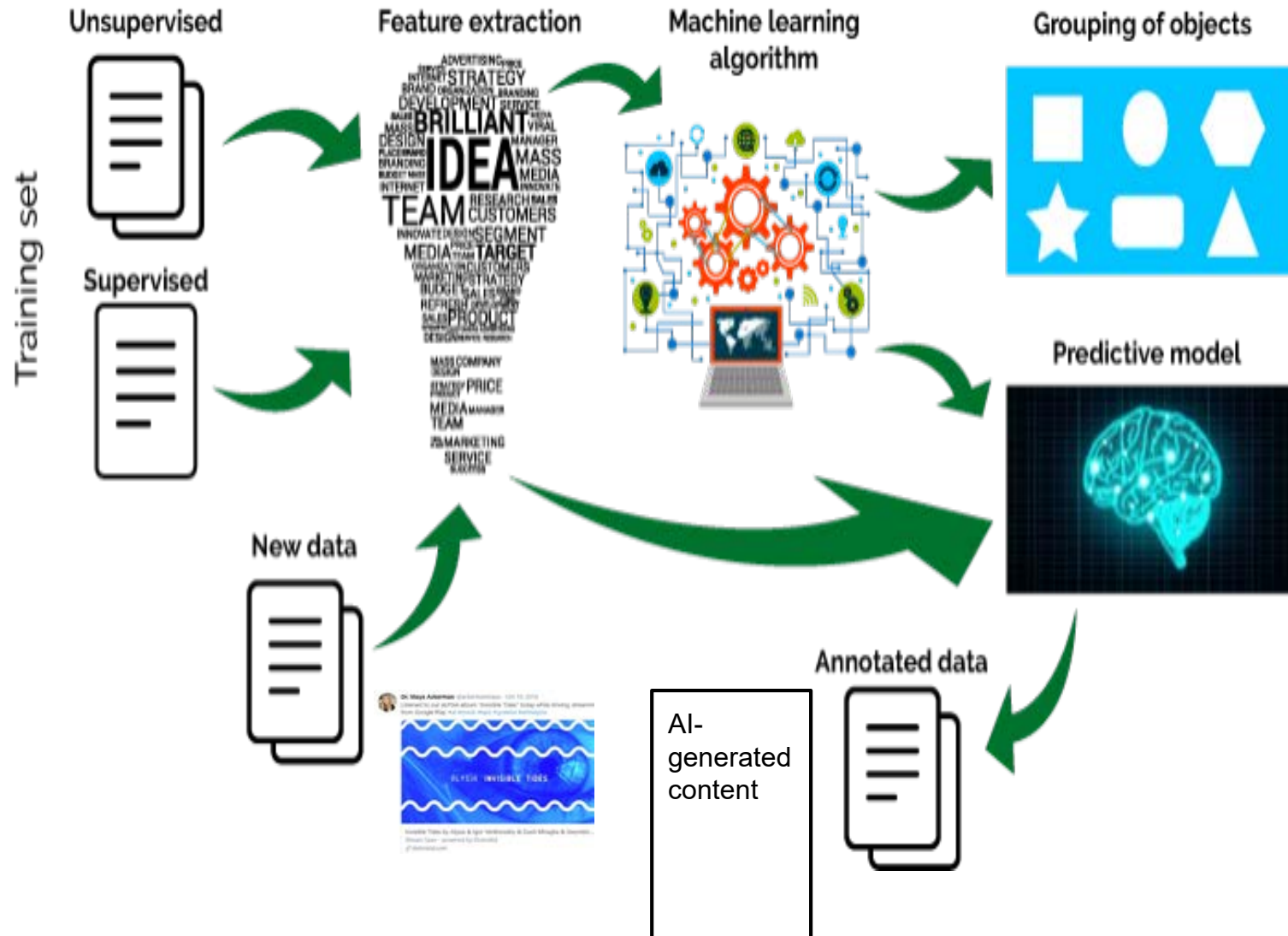
- are not exclusive of each other - a single product can and usually is covered
- “one size fits all” - not generally tailored to technology types (although see copyright)
- reflect a balance between innovators of today and tomorrow.

Speakers

- Professor Tyler Ochoa, Santa Clara University School of Law
- Dr. Christian Mammen, Partner, Womble Bond Dickinson LLP
- Brian Adams, Associate Patent Counsel, Qualcomm
- Hogene Choi, Partner, Baker Botts
- **Professor Colleen Chien, Santa Clara University School of Law**

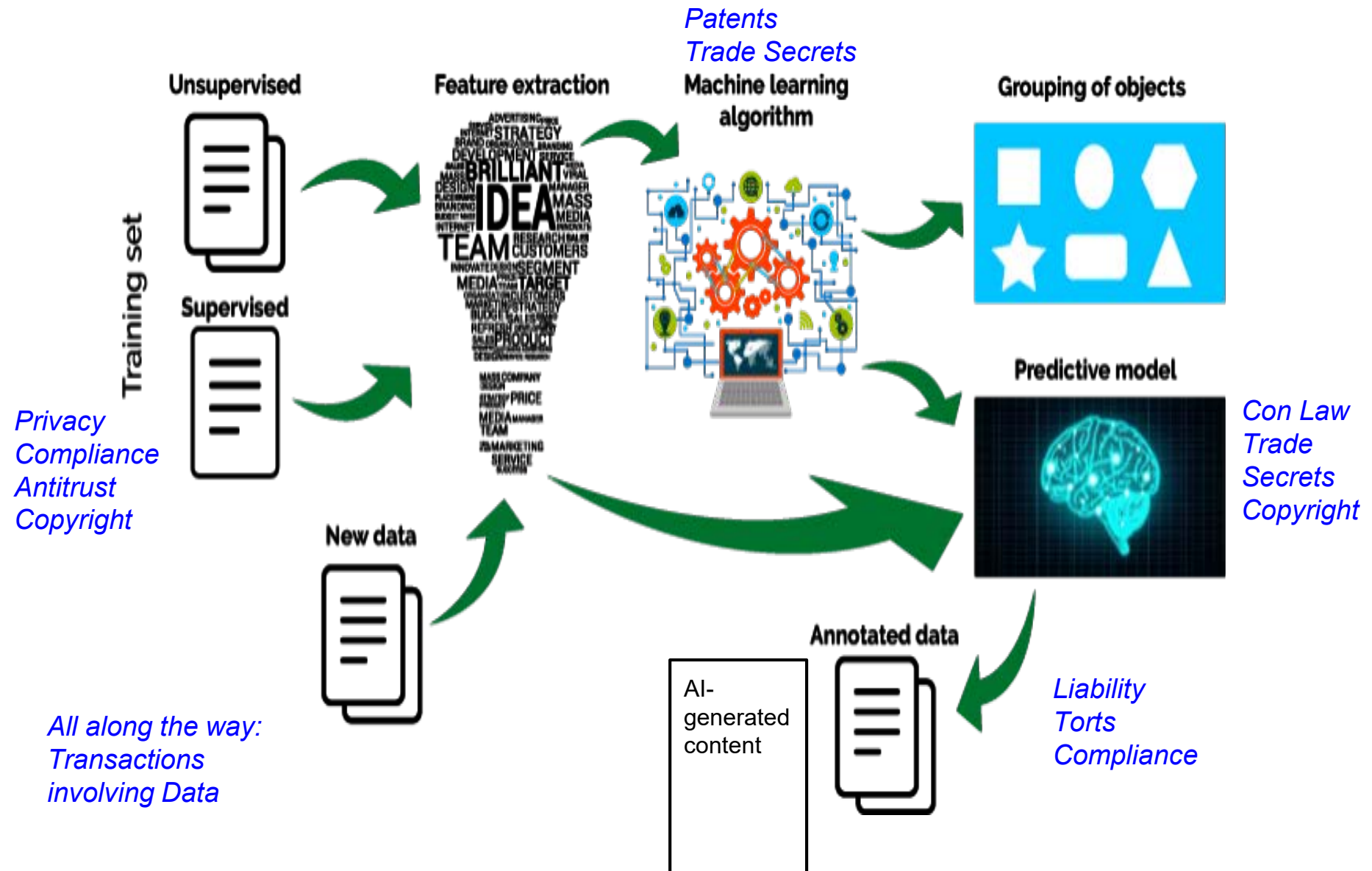
NATIONAL INSTITUTES

Artificial Intelligence and Robotics



NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Progress in AI...



EP3564144A1 Pending

Quilify: C Value: C

(Origin) FOOD CONTAINER
(FR) RÉCIPIENT ALIMENTAIRE
(DE) LEBENSMITTELBEHÄLTNER

Full Text Simple Family Extended Family Citations History Original Document

Abstract

A container (10) for use, for example, for beverages, has a wall (12) with an external surface (14) and an internal wall (16) of substantially uniform thickness. The wall (12) has a fractal profile which provides a series of fractal elements (18-28) on the interior and exterior surfaces (14-16), forming pits (40) and bulges (42) in the profile of the wall and in which a pit (40) as seen from one of the exterior or interior surfaces (12, 14) forms a bulge (42) on the other of the exterior or interior surfaces (12, 14). The profile enables multiple containers to be coupled together by inter-engagement of pits and bulges on corresponding ones of the containers. The profile also improves grip, as well as heat transfer into and out of the container.

Figure (7)

Bitography

Earliest Priority: 2015-10-17
Legal Status: Application yet to be examined
Designated In: AT, BE, CH, DE, DK, ES, FR, GB, GR, IT, U, LU, NL, SE, MC, PT, IE, SI, LT, LV, FI, RO, MK, CY, AL, TR, BG, CZ, EE, HU, PL, SK, BA, HR, IS, MT, NO, RS, ME, SM, MA, KH, TN, MD
Cur. Assignee: THALER STEPHEN L 2019-11-07
Assignee (Std): THALER STEPHEN L [+Cur.Assignee]
Patent Family: 1 Member(Search ID: 83913228)
EP(1)
Patent Type: Utility Patent
Show All

Framing Questions

- 1. Provide two concrete examples of how AI issues are being raised in your area of law.**
- 2. What makes the issues raised by AI different?**
- 3. Areas of legal or policy uncertainty?**
- 4. Best practices or promising models?**
- 5. What issues are of most urgent of policy attention?**



Ownership of Copyright Created with AI Tools

Prof. Tyler Ochoa

High Tech Law Institute

Santa Clara University School of Law

Ownership of Copyright in Works Created with Artificial Intelligence Tools

Professor Tyler T. Ochoa
High Tech Law Institute
Santa Clara University School of Law

January 10, 2020

Five Types of Ownership

- = **§201(a): Sole authorship**
 - Single author (any other contributions are *de minimis*).
- = **§201(a): Joint authorship**
 - If both intended to merge expression into a unitary whole.
- = **§201(b): Work made for hire (§101)**
 - If work created by employee, or commissioned work falls within one of nine categories and there is signed writing.
- = **§103(b): Derivative work (§101)**
 - If a second author adds significant original expression.
- = **§201(c): Collective work**
 - If elements are independently copyrightable, and user selects and arranges them into a collective whole.

Who is an “Author”?

- = The term “author” is not defined in the statute (except for works made for hire); but there are several indications that Congress contemplated a human author.**
 - §101 defines “widow” or “widower” as “the author’s surviving spouse under the law of the author’s domicile at the time of his or her death.”**
 - An author’s children, “whether legitimate or not” can inherit rights under the statute.**
 - An author’s grandchildren can inherit termination rights.**

Who is an “Author”?

- = **Naruto v. Slater, 888 F.3d 418 (9th Cir. 2018):**
- “The terms ‘children,’ ‘grandchildren,’ ‘legitimate,’ ‘widow,’ and ‘widower’ all imply humanity and necessarily exclude animals that do not marry and do not have heirs entitled to property by law.”



Who is an “Author”?

- = The Copyright Office states: “To qualify as a work of ‘authorship’ a work must be created by a human being. Works that do not satisfy this requirement are not copyrightable.”
[Compendium III, § 313.2]**

Who is an “Author”?

- = “the Office will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author.” [Compendium III, § 313.2]**
- = “music generated entirely by a mechanical or an automated process is not copyrightable. ... Nor could a musical composition created solely by a computer algorithm be registered.” [Compendium III, § 802.5(C)]**

Computer Assisted Works

- = Authors use tools. That doesn't give the creator of the tool a claim to authorship of the works that result from use of those tools.**
 - Traditional tools: pencil and paper, brush and paint, cameras and film, magnetic tape**
 - Digital tools: word processor, paint program, 3D printer, digital cameras, computer editing**

Computer Assisted Works

- = Rearden, LLC v. Walt Disney Co., 293 F. Supp. 3d 963 (N.D. Cal. 2018):**
 - Owner of copyright in performance-capture software sued three movie studios (Disney, Fox, Paramount) for copyright infringement based on third-party's use of the software to create sequences for major motion pictures.
 - e.g., *Beauty and the Beast*, *Deadpool*, *Terminator: Genisys*
- = HELD:** Assuming without deciding that the copyright in a software program may extend to the program's output, it does so only if the program does the "lion's share" of the work, and the user's input is so marginal that the output reflects the program's contents.

Computer Generated Works

- = The “tool” is programmed to “create” new works in a non-deterministic (“random”) manner when run by a user.
- = Programmer is the author of the tool (code), but process/algorithm is non-protectable under § 102(b).
- = User does not appear to add original expression.
- = Possible copyright claim to selection of inputs; but doubtful copyright would extend to the output.

AI and Trade Secret

Dr. Christian Mammen

10/30/19: USPTO Request for Comments (84 FR 58141)

- 10. How, if at all, does AI impact trade secret law? Is the Defend Trade Secrets Act (DTSA), 18 U.S.C. 1836 *et seq.*, adequate to address the use of AI in the marketplace?
- 11. Do any laws, policies, or practices need to change in order to ensure an appropriate balance between maintaining trade secrets on the one hand and obtaining patents, copyrights, or other forms of intellectual property protection related to AI on the other?

How is AI implicated in trade secret protection?

1. The AI algorithm is the trade secret
1. AI involved in generating the trade secret
1. AI involved in misappropriating trade secrets

Quick Primer on US Trade Secret Law

- Sources
 - Uniform Trade Secrets Act – e.g., Cal. Civ. Code § 3426
 - Defend Trade Secrets Act – 18 U.S.C. § 1936
- Trade Secret Definition
 - Non-public (secret) information
 - Independent economic value from being non-public
 - Subject to reasonable efforts to maintain secrecy
- Misappropriation Definition
 - Acquired through improper means
 - Disclosure or use by someone who acquired through improper means or violated duty of secrecy

Trade Secrets Compared with Other IP Types

- Advantages
 - Potentially indefinite duration
 - No registration requirement
 - State and federal protections
 - Non-monetary remedies relatively easier?
- Disadvantages
 - Must show actual taking (theft) from right holder
 - Easy to lose via
 - public disclosure
 - reverse engineering
 - inadequate lockdown

Fundamental Tensions in Trade Secret Law

- To qualify for trade secret status, the information must be locked down ... but any claim for misappropriation means that someone, somehow, improperly took the information out of the secure environment.
- To assert a claim for trade secret misappropriation, the owner of the trade secret must make some, sufficiently detailed, disclosure to describe the misappropriated trade secrets.

(1) The AI Algorithm is the Trade Secret

- Examples: Search, autonomous vehicles, speech recognition
 - AI algorithm, combined with training data, produces many discrete results that are useful and accurate, but not individually valuable
- *Loop AI Labs v. Gatti*, 195 F.Supp.3d 1107 (N.D. Cal. 2016)
 - Publicly-filed document listing 55 general categories is insufficient disclosure under CCP § 2019.210
- *LivePerson v. 24/7 Customer*, 83 F.Supp.3d 501 (S.D. N.Y. 2015)
 - Misappropriation claim involving “predictive algorithms” adequately pled
- *DigitalGlobe, Inc. v. Paladino*, 269 F.Supp.3d 1112 (D. Colo. 2017) (motion for PI on breach of contract/noncompete claim)
 - “algorithms and methods for applying machine learning to large volumes of geospatial data” *can* qualify as trade secrets *but* allegations of misappropriation too generic

(2) AI Involved in Generating the Trade Secret

- Example: AI algorithm comes up with a unique and valuable solution to a known problem – e.g., new medical treatment
- Impediments:
 - Can the end product be reverse-engineered (not the process by which the end product was created, but the product itself)?
 - AI “black box” problem: how to make a CCP § 2019.210 disclosure of the trade secret?

Training Data as the Trade Secret?

- When, and under what circumstances, might the training data used to train the algorithm be considered a trade secret?
 - Nonpublic, independent economic value, reasonable efforts to maintain secrecy
 - Separate claim from algorithm and/or product of AI?

(3) AI Involved in Misappropriating Trade Secrets

- Example: AI is somehow able to “break” or derive the claimed trade secrets
- Return to Definition of Misappropriation (Civ. Code § 3426.1)
 - Acquisition by improper means
 - Disclosure or use by a person who acquired by improper means or had a duty to maintain secrecy
 - “Improper means” includes theft, bribery, misrepresentation, breach or inducement of a breach of a duty to maintain secrecy, or **espionage through electronic or other means**. Reverse engineering or independent derivation alone shall not be considered improper means.³¹

NATIONAL INSTITUTES

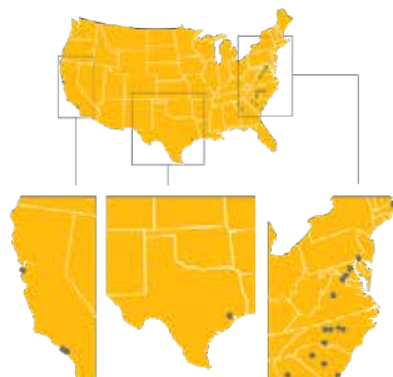
Artificial Intelligence and Robotics



27 Locations
across the
US and UK



More than
400 Partners
1,000 Lawyers



US

Boston
Wilmington
Baltimore
Washington, D.C.
Tysons Corner
Charlottesville
Raleigh
Research Triangle Park
Greensboro
Winston-Salem
Charlotte
Greenville
Columbia
Charleston
Atlanta
Houston
Silicon Valley
Los Angeles
Orange County



UK

Aberdeen
Edinburgh
Newcastle
Leeds
London
Southampton
Bristol
Plymouth

Consolidating
our national
reputations and
regional heritage
under one powerful
transatlantic brand

Our sectors



Energy &
Natural Resources



Healthcare



Manufacturing



Wealth
Management



Transport,
Logistics &
Infrastructure



Financial
Institutions



Insurance



Life Sciences and
Pharmaceuticals



Real
Estate



Retail &
Consumer



Technology

Representing
more than

250

Publicly Traded
Companies
in the US and UK



+150
Chambers
rankings



UK Top
20 Law
firm



US Top
80 Law
firm



Global
Top 100
Law firm by
revenue

IP Protection for Data

Brian Adams, Qualcomm

Data's Importance

“Data is becoming a new natural resource. It promises to be for the 21st century what steam power was for the 18th, electricity for the 19th, and hydrocarbons for the 20th.”

- Ginni Rometty (IBM), 2013

Artificial Intelligence systems are only as good as the data used to train them

...so how do we protect the data?

What is Data

2.5 quintillion bytes of data created every day, and by 2020 1.7MB of data every second for every person on earth



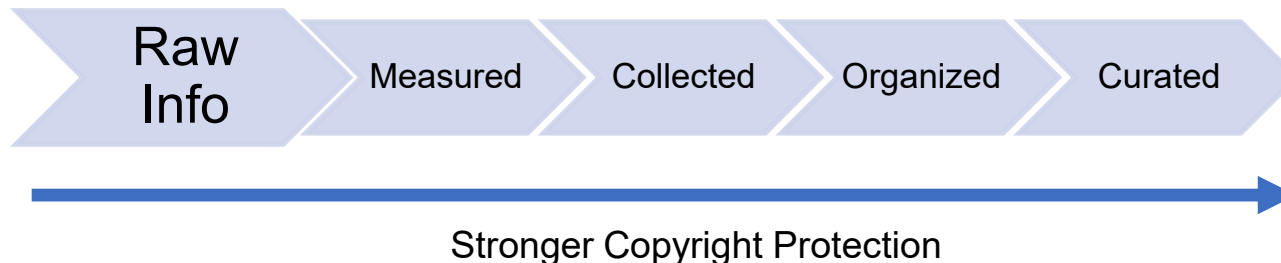
Data and Copyright Law

facts are not protectable by copyright

"In no case does copyright protection . . . extend to an idea . . . concept, principle or discovery . . ." 17 U.S.C. § 102(b)

databases as compilations

a compilation is a "collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship." 17 U.S.C. § 101



Data and Copyright Law

- Feist Publications v. Rural Telephone Service (S. Ct. 1991)
 - “Facts are never original, so the compilation author can claim originality, if at all, only in the way the facts are presented.”
 - Individual *datum* are not protectable; but a mass of data may be
 - rejected “sweat of the brow” because it extended copyright to the facts
 - for compilation to be protectable, consideration is whether the “selection, arrangement and coordination” of the data is original
 - “Originality requires only that the author make the selection or arrangement independently...and that it display some minimal level of creativity.”
 - but phone book listing in alphabetical order is not enough

Data and Copyright Law

- *Digital Drilling Data Sys. v. Petrolink Servs* (S.D. Tex. 2018)
 - The technology at issue is used in directional drilling for oil and gas
 - Digital Drilling's program collects raw data from the drill, does data manipulations and reporting, and stores the raw and derived data in a database
 - Defendant Petrolink's program displays drill data visually
 - Rather than collect its own data, Petrolink's program copied and used the data stored in the Digital Drilling database
 - Held on SJ: The organizational structure of the Digital Drilling database is copyrightable, but the actual data stored in the database is not
 - The copyright claim was dropped prior to trial
 - After trial, the plaintiff was awarded \$414k for unjust enrichment

Data and Copyright Law

- **Other Considerations:**
 - Fair Use – is AI transformative?
 - EU Database Directive
 - Text and Data Mining (TDM)

Key take away:

Copyright protection may be easy to obtain but may provide thin protection

Data as Trade Secrets

- Unlike copyright, trade secret protection extends to facts, and does not require any creativity
- What if your use of the data exposes it to the public?

Protecting Data by Other Measures

- Attorneys should encourage the business to think early on about ways to protect their data
- Contracts
- Unjust enrichment
- Technical Measures

Limitations on the Rights to Collect and Use Data

- Data protection and privacy laws and regulations
 - The EU General Data Protection Regulations (GDPR)
 - California data privacy law
 - HIPA
- If data was obtained from customers or users, they may have rights in the data
 - Think early on about data ownership (contracts)

Open Source Business Models

- “If you love someone, set them free.”
- The open source software business model
 - 15 years ago, the business case for contributing to open source software was not well understood
 - Today, many companies contribute to open source software projects (charge for ancillary services or help improve OSS infrastructure)
- Open Data
 - Open data licenses were created to enable sharing of data as easily as we currently share open source software code

Open Source Business Models

In discussing IP protection for Data, there are two buckets

- 1.) Data with Inherent value (i.e. Customer lists)
- 2.) Data as a means to an end (i.e. Training AI systems)

- Employ a **nuanced** IP strategy
 - Inherently valuable Data might be kept proprietary
 - Other Data might be released as Open Source
 - Or some other combination

Open Source Business Models

- Example – The Waymo Open Dataset
 - Released in August 2019 to aid the research community in making advancements in machine perception and self-driving technology.
 - Contains a curated set of self-driving data that can be used for research on machine learning models.
 - Released under the license agreement “Waymo Dataset License Agreement for Non-Commercial Use”

Open Data Licenses Examples

- CC0 1.0 Universal (CC0 1.0)
 - Public Domain Dedication
- Community Data License Agreement (CDLA)
 - CDLA Sharing
 - Copyleft-style (GPL like) license
 - CDLA Permissive
 - Apache-style license.

Open Data Projects

- Standardized Data Definitions
 - Example of temperature (C or F, integer increments, etc...)
- Project Maintainers also play a key role
 - Quality v. Quantity
 - Goals and direction of the project

AI and Patents

Hogene Choi, Baker Botts

NATIONAL INSTITUTES

Artificial Intelligence and Robotics



Patenting Artificial Intelligence

ABA National Institute: AI & Robotics

January 10, 2020

Hogene L. Choi

01

PATENTABILITY OF AI

AI-Specific Considerations

Subject Matter Eligibility (§ 101): foundational machine-learning techniques

- Example: U.S Patent No. 10,311,342

1. A method for processing an image, comprising:

by a camera of a mobile computing device, capturing the image;

by a processor of the mobile computing device, inputting data representing the image into a convolutional neural network (CNN), the CNN including a plurality of convolutional layers, a set of weights or filters for at least one of the layers, and a set of input data to the at least one of the layers;

by the processor of the mobile computing device, representing a convolution operation between the set of input data and the set of filters or weights by a product of a scaling factor and a binary representation of the set of filters or weights convolved with the set of input data, wherein the binary representation is the sign of the weight values, and the scaling factor is the average of the absolute weight values; and

by the processor of the mobile computing device, applying a classification operation to an output of the last of the plurality of convolutional layers.

AI-Specific Considerations

Subject Matter Eligibility (§ 101): practical applications of AI

- Example: U.S Patent No. 10,426,442

1. A method for processing digital images in assisted reproductive technologies, the method comprising:

- obtaining one or more digital images of a reproductive anatomy of a patient through one or more imaging modalities;
- processing the one or more digital images to detect one or more reproductive anatomical structures;
- processing the one or more digital images to annotate, segment, or classify one or more anatomical features of the one or more reproductive anatomical structures;
- analyzing the one or more anatomical features according to at least one linear or non-linear framework; and,
- predicting at least one time-to-event outcome of an assisted reproductive procedure according to the at least one linear or non-linear framework, the at least one time-to-event outcome comprising an egg retrieval date.

AI-Specific Considerations

Subject Matter Eligibility (§ 101): techniques for creating training data

- Example: U.S Patent No. 9,173,614

1. A method for automating the identification of meaningful features and the formulation of expert rules for classifying magnetocardiography data, comprising:
 applying a direct kernel transform to sensed data acquired from sensors sensing magnetic fields generated by a patient's heart activity, resulting in transformed data; and
 identifying said meaningful features and formulating said expert rules from said transformed data, using machine learning.

AI-Specific Considerations

§ 112: Written Description: Enablement

- For foundational machine-learning techniques
- For practical applications of AI
- For techniques for creating training data

§§ 102 and 103:

- Anticipation
- Obviousness

Patent It? Or Keep It As a Trade Secret?

Is the invention detectable and embodied in the product or service?

- Is your business built on a SAAS model?
- Are you distributing trained ML models, software to train a ML model, or software to create training data?

Do you need to disclose your invention to obtain funding or as part of a joint development effort or other partnership?

Is the invention likely to be independently discovered or invented?

How prevalent is the risk of trade secret theft?

02

AI AND INVENTORSHIP

Artificial Inventor Project

Can an AI be an inventor?

- DABUS “start[s] as a swarm of many disconnected neural nets, each containing interrelated memories, perhaps of a linguistic, visual, or auditory nature. These nets are constantly combining and detaching due to carefully controlled chaos introduced within and between them. Then, through cumulative cycles of learning and unlearning, a fraction of these nets interconnect into structures representing complex concepts. In turn these concept chains tend to connect with other chains representing the anticipated consequences of any given concept. Thereafter, such ephemeral structures fade, as others take their place, in a manner reminiscent of what we humans consider stream of consciousness.” http://imagination-engines.com/iei_dabus.php
- DABUS “only received training in general knowledge in the field, and proceeded to independently conceive of the invention, and to identify it as novel and salient. [It] was not created to solve any particular problem, nor was trained on any special data relevant to the instant invention.” <http://artificialinventor.com/patent-applications/>

Artificial Inventor Project

“Food container” (EP3564144A1; GB1816909.4)

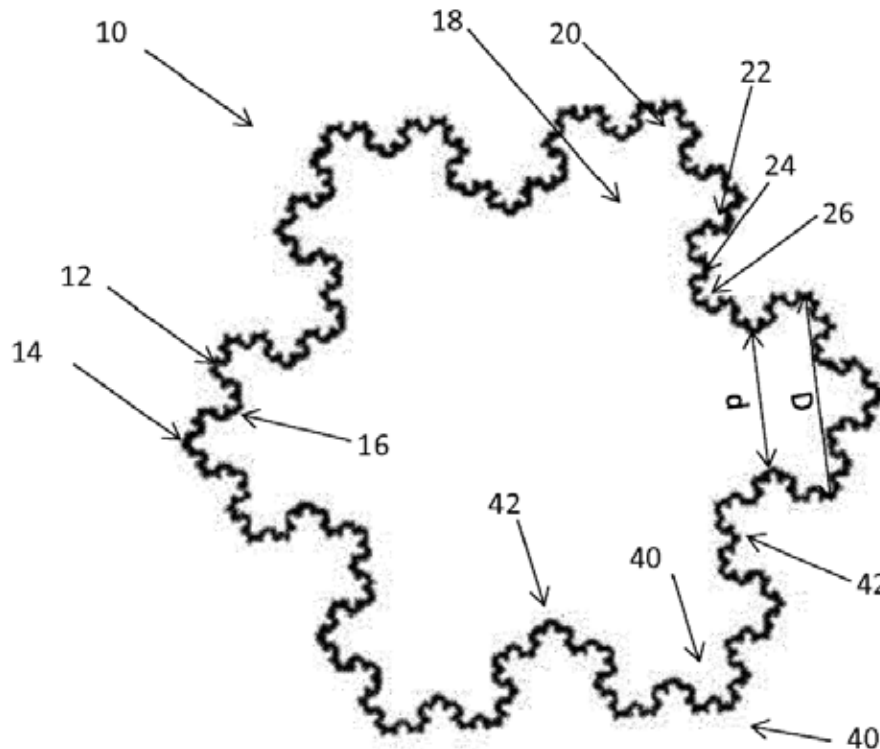


Fig. 1

Artificial Inventor Project

“Devices and methods for attracting enhanced attention” (EP3563896A1;
GB1818161.0)

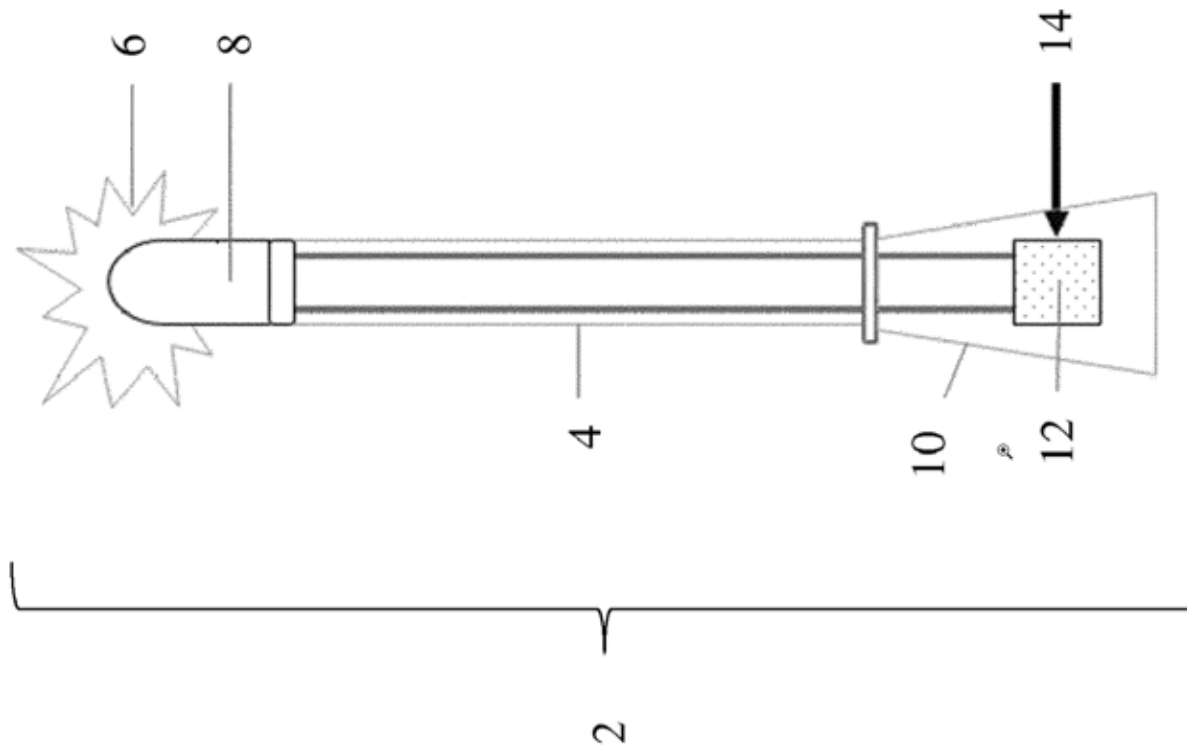


Figure 1

Can an AI be an inventor?

European Patent Office (applications refused): *reasoning not published yet*

United Kingdom Intellectual Property Office (applications withdrawn):

- “The fundamental function of the patent system is to encourage innovation by granting time-limited monopolies in exchange for public disclosure.”
- “[An] AI machine is unlikely to be motivated to innovate by the prospect of obtaining patent protection. Instead, the motivation to innovate will have been implemented as part of the development of the machine; in essence, it will have been instructed to innovate.”
- “DABUS, as a machine, cannot own intellectual property . . . DABUS has no rights to its inventions and cannot enter into any contract to assign its right to apply for a patent to the applicant. It is unclear, therefore, as to how precisely the applicant has derived the right to the inventions from their creator, DABUS.”

Other counterpart patent applications filed and pending in the United States, United Kingdom, Germany, Israel, China, Korea and Taiwan.

AI and Robotics Standards, Certifications, and Auditing

K

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

AI and Robotics Standards, Certifications, and Auditing

January 10, 2019 / 2:15 - 3:15 pm



THE PREMIER SOURCE FOR CLE

AI and Robotics Standards, Certifications, and Auditing

January 10, 2019/ 2:15 - 3:15 pm

Speakers

- Eric Hibbard, CISSP-ISSAP, CIPT, CISA, CCSP
 - ABA SciTech Leadership
 - SDO Leadership (U.S., ISO, & IEEE)
 - Industry, Government, & Academia Experience
- Sumit Kalra (CISA, CISSP, SSCP, ISO Lead Auditor)
 - Former Partner at BPM LLP.
 - Sr. Manager at Grant Thornton LLP
 - Sumit@SecurityCrisis.com

Current State – Standards (Sample)

ISO/IEC 22989:--, *Artificial intelligence — Concepts and terminology*

ISO/IEC 23894:--, *Information Technology — Artificial Intelligence — Risk Management*

ISO/IEC TR 24027:--, *Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making*

ISO/IEC TR 24028:--, *Information technology — Artificial Intelligence (AI) — Overview of trustworthiness in Artificial Intelligence*

ISO/IEC TR 24368:--, *Information Technology — Artificial Intelligence — Overview of ethical and societal concerns*

NOTE: All listed ISO/IEC documents currently in draft states.

AI Definitions

Existing ISO Definition

- capability of a functional unit to perform functions that are generally associated with human intelligence such as reasoning and learning
- interdisciplinary field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning
- branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement

Chartered Professional Accountants of Canada

- AI is the science of teaching programs and machines to complete tasks that normally require human intelligence.

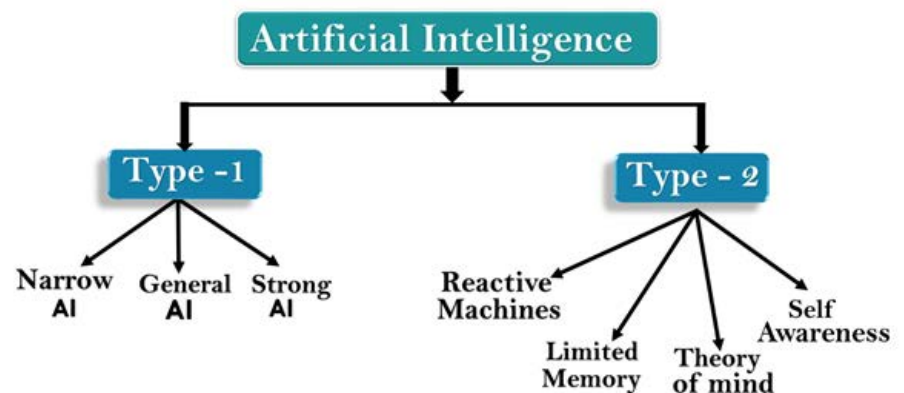
AI Types (Based on Classifications)

Capabilities

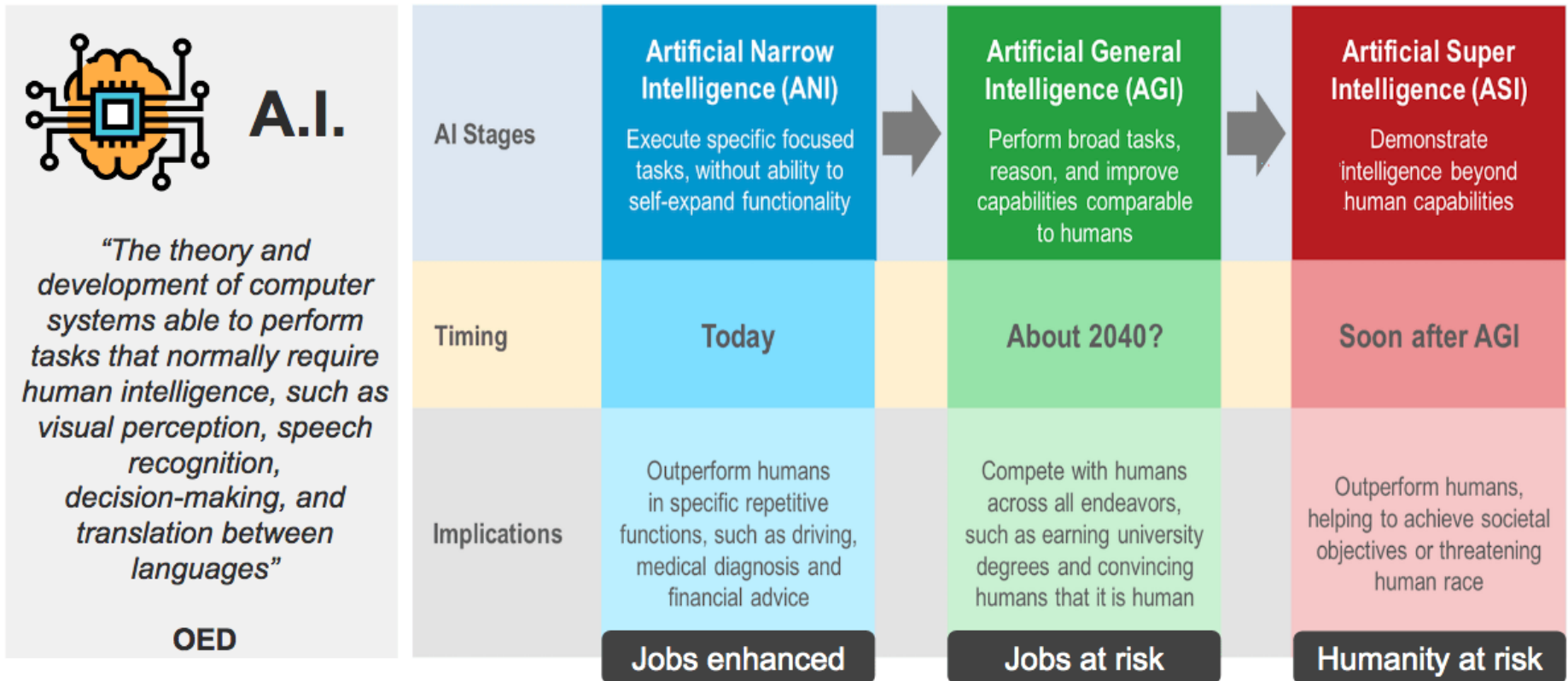
- Artificial Narrow Intelligence (ANI)
- Artificial General Intelligence (AGI)
- Artificial Superintelligence (ASI)

Functionalities

- Reactive Machines
- Limited Memory
- Theory of Mind
- Self-awareness



Type I - Enhancement



Current State - Certifications

- International Standards Organization (ISO)
- Information Security Management System: ISO 27001 (Information Security Management System)
- Cloud: ISO 27017; FedRAMP; CSA STAR
- Privacy: ISO 27701 (draft)

Current State - Audits/Attestations

- Frameworks: AICPA, PCI, FDA, FAA, Etc.
 - Assertion based;
 - Compliance based;
 - Subject matter based.
- Enable Trust Amongst:
 - Business2Business (B2B);
 - Business2Consumer (B2C);
 - Business2Government (B2G);
 - and All of the various combinations of the three...

Current State - Audits/Attestations (Cont.)

- Security and Compliance Audit Practices:
 - Infrastructure
 - Supporting Roles
 - Application Code
 - Policy, Processes and Procedures
- Social Engineering
- Physical Location
- Ethical Hacking

Ethical vs. Unethical?

- Various Global Definitions
- Geographically Determined
- Baseline Algorithms
- Dynamic learning
- Cognitive Decision making

Much of it is unauditable...

What is Missing in Audit Standards?

- Clear definition of scope?
- Can the opinion be extended:
 - to dynamic scope?
 - to projected results?
 - Who is held:
 - responsible?
 - accountable?
 - financially liable?
- Continuous Auditing or Management Operations?
- Litigation: who is going to get sued?

Potential Trustworthiness Properties

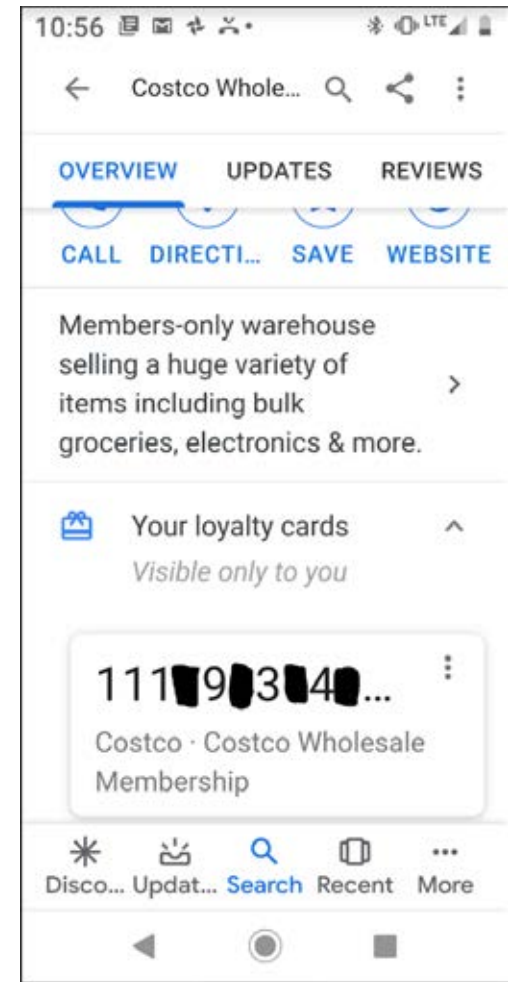
- Security
- Privacy
- Safety (environmental, living entities, operational)
- Reliability (perform consistently)
- Resiliency (recover quickly)
- Not Included:
 - Interoperability
 - Scalability
 - Dependability
 - Robustness

Trustworthiness

- Who is listening? Digital Assistants
- What are they doing? Creating Digital profiles
- What are they doing with information? Training AI Bots
- Data Collected: Record Amount
- Data Anonymization: Where in the workflow?
- Data Sharing: Raw, Processed, Anonymized, Bigdata?
- Data Collaboration: who has a copy of it?
- Data Processing: Multiple applications

Trustworthiness

- I made a call to Costco Tire Center;
- Gave them my Membership #;
- AI Engine recorded and Analyzed my conversation;
- After I hung up; notice that Google wanted to know what i would like to do with my information it captured;



Trustworthiness:

- Historical Perspective:
 - Audits/Certifications - Vehicles to Communicate Compliance Status through Examination of historical records.
 - Records Evaluation - Existence, Completeness, Accuracy, Timely etc.
 - Auditing the past to demonstrate trust
- Current Auditing Techniques focus on the past
- Current Auditing Standards do not allow to project on the future...

Common Sources of Bias in AI/ML

- Unintentionally uploading the implicit human biases that pervade our culture. How should we codify definitions of fairness?
- Poorly-selected training data for machine learning, or poorly reasoned rules; underlying data rather than the algorithm itself are most often the main source of the issue.
- Intentional bias by evil programmers, corporations, or governments

There's no real way to eliminate cultural bias without fixing our culture first, so we need to compensate for it when we design our systems.

Bias

- Similar to Trustworthiness, Bias cannot be audited
- Under current standards, for Bias:
 - Auditable criteria does not exist
 - Measurement Baselines do not exist
 - Cannot provide opinion on a future event

All algorithms that affect people's lives should be subject to audit.

Current & Future State of the Profession

Current

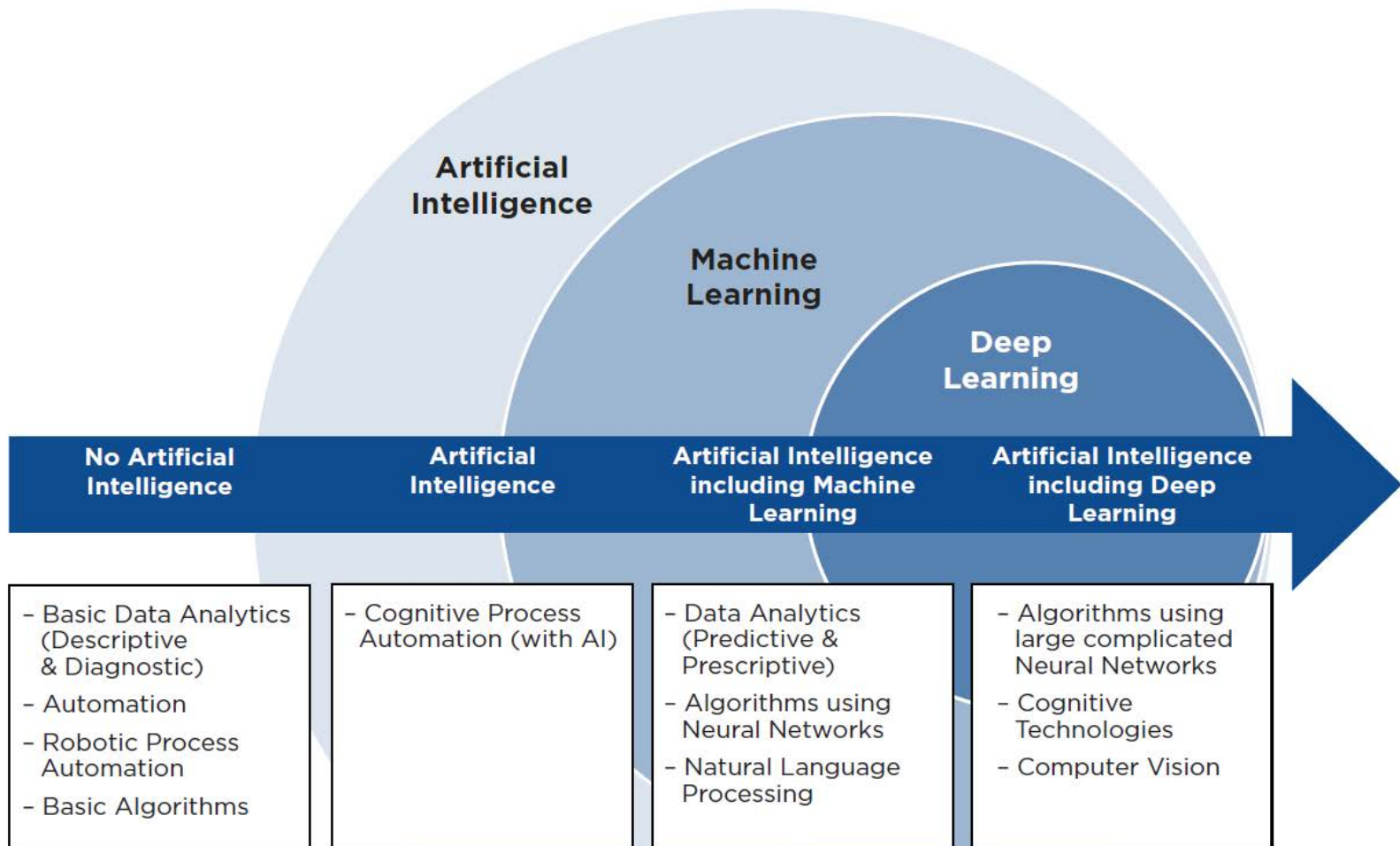
- Audit Data Normalization
 - Audit Documentation
 - Evidence Collection
 - Evidence Evaluation
- Auditing Tools Growth
 - Word & Excel
 - Disjointed Toolset
 - Data Analytics
 - Around the System
- In-summary - Manual

Future

- Use of Captured Normalized Data
 - Workflow Creation
 - Task Automation
 - AI Engine Training
- Auditing Tools
 - Apply Cognition
 - Evaluate Data
 - Audit the System
 - Reach Conclusions
- In-summary - Automated

NATIONAL INSTITUTES

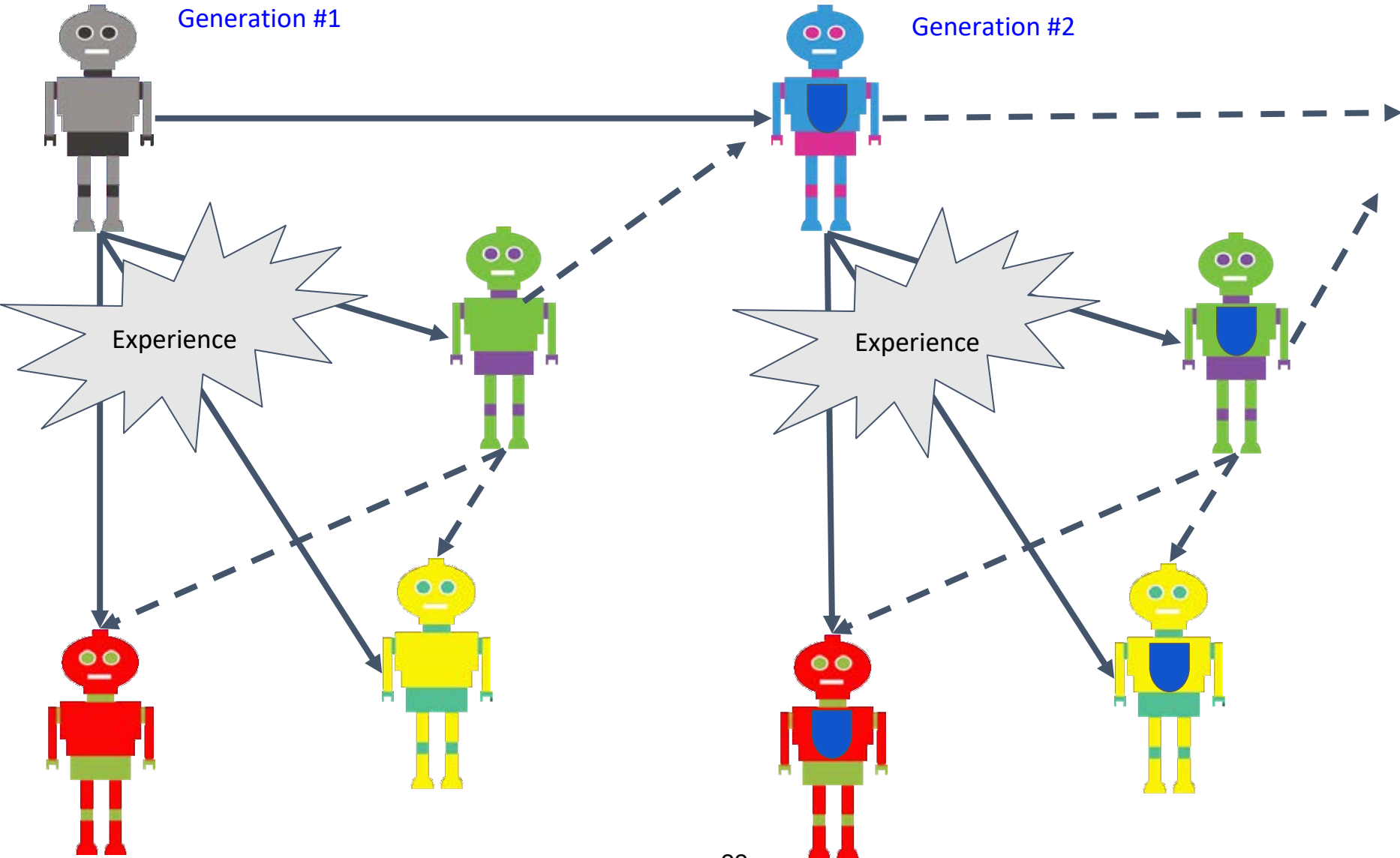
Artificial Intelligence and Robotics



NATIONAL INSTITUTES Artificial Intelligence and Robotics

Generation #1

Generation #2



AI Beyond Controlled Baseline

- Leveraging unique experiences
- Propagation of “good experiences” to other AIs of same model (same generation)
 - Used as initial starting point
 - Reset/override of AIs in the field
 - Complete versus partial
- Capturing “good experiences” for next generation AIs models
- How much human involvement/supervision?

Legal Implications of Experience

- Who is liable?
- What if someone gets a hold of bad models?
- Intentional use for harm? Unethical use?
- Miss appropriation of tools...who is liable?
- Responsibility for the unintended impact for the AI model?

References

- *A CPA's Introduction to AI: From Algorithms to Deep Learning, What You Need to Know*; Chartered Professional Accountants of Canada, 2019,
<https://www.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/cpas-introduction-to-ai-from-algorithms.pdf>
- the rise of artificial intelligence: a critical inflection point for the accounting profession
- <https://www.cpa.com/whitepapers/rise-artificial-intelligence-critical-inflection-point-accounting-profession>
- ISO Online Browsing Platform (OBP),
<https://www.iso.org/obp/>
- *AI for Legal: ANI, AGI and ASI*; Lawtomed, 2019,
<https://lawtomed.com/ai-for-lawyers-ani-agi-and-asi/>

AI, Robotics, Ethics, and the Public Good

L

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

AI, Robotics, Ethics and the Public Good

January 10, 2010 / 3:30 pm



THE PREMIER SOURCE FOR CLE

Speakers

- Ryan Budish (Asst. Research Director, Berkman Klein Center for Internet & Society, Harvard)
- Theresa Harris (Project Director, Scientific Responsibility, Human Rights and Law Program, American Association for the Advancement of Science)
- Shannon Vallor (McKenna Professor of Philosophy, Santa Clara University)
- Cynthia Cwik (Continuing Fellow, Stanford Distinguished Careers Institute) [Moderator]

Cynthia Cwik

- **Areas of Interest:** Issues at intersection of law, science, technology and policy
- **Presently** Fellow and Continuing Fellow, Stanford Distinguished Careers Institute, where focusing on cutting-edge technology issues, including the ethical, economic, legal, policy and societal implications of emerging technologies
- **ABA Experience:** Chair of Section of Science & Technology Law (2015-2016); While Chair, launched the first ABA National Institute on the Internet of Things (2016); Co-Editor: *The Internet of Things: Legal Issues, Policy and Practical Strategies* (2019); Co-Chair, Science & Technology for the Public Good Committee (2019)
- **Partner with Jones Day and Latham & Watkins:** Represented Fortune 500 corporations in high-stakes, high-profile matters involving science and technology issues.

Issues to Be Addressed:

- The most urgently needed applications of AI for the public good
- Meaning of phrase: “AI for the public good”
- Role of ethics in AI for the public good vs. role of law and regulations
- Role of corporate social responsibility practices and ethical codes regarding AI

Shannon Vallor

- **My Research Areas:** Ethics of Data and Artificial Intelligence, Philosophy of Science and Technology
- **Presently** the Regis & Dianne McKenna Professor at Santa Clara University where I've taught since 2003
- **In February 2020:** moving to the University of Edinburgh as the Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence
- **Formerly:** Visiting Researcher/Consulting AI Ethicist at Google (2018-2020) and President of the Society for Philosophy and Technology (2015-2017)
- **Boards:** Foundation for Responsible Robotics (NL), Chair, Data Delivery Group (Scottish Government)

Shannon Vallor

- **Past Research:** *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (2016, Oxford University Press)
- **Current Research:** Follow-up book on AI: *The AI Mirror: Rebuilding our Humanity in an Age of Machine Thinking*
- **Primary AI Ethics Research Themes and Questions:**
 - AI's Impact on Our Moral Habits, Skills & Virtues
 - Designing AI as a Support for Human Moral Intelligence
 - AI/Robotic Mediation of Human Moral Relations
 - Sustainable AI for Long-term Human Flourishing

Shannon Vallor

Outreach with SCU's Markkula Center for Applied Ethics:

- **Practical Ethics Education for STEM Disciplines (requested for use by 200+ higher ed instructors in 21 countries)**
 - *An Introduction to Software Engineering Ethics (2013/2015)*
 - *An Introduction to Data Ethics (2017)*
 - *An Introduction to Cybersecurity Ethics (2018)*
 - Mozilla Responsible CS Challenge Grantee (2019)
- **Practical Ethics Training for Working Technologists**
 - *Ethics in Tech Practice (2018)* – funded by Omidyar Network's Tech Society & Solutions Lab, released under CC license
 - Piloted at X (Alphabet) in 2018-2019 and since adapted for use at Google and other large tech companies
- **Summer Institute in Tech Ethics (SITE) at SCU July-August, 2020**

Theresa Harris

- **Presently** Project Director, AAAS Scientific Responsibility, Human Rights and Law Program
- **Focus Areas:** Intersections of Science, Technology and Human Rights including use of scientific evidence in human rights courts, scientific responsibility and human rights, and applications of emerging technologies to protect and advance human rights

Theresa Harris

- **Examples of AAAS projects:**
 - On-call Scientists pro bono network
 - Science and Human Rights Coalition
 - *Location-Based Data in Crisis Situations: Principles and Guidelines* [2019]
 - *Geospatial Evidence in International Human Rights Litigation: Technical and Legal Considerations* [2018]
 - *Broadly Accepted Practices Regarding the Use of Geospatial Technologies for Human Rights* [2016]
- **AAAS and ABA joint committee** National Conference of Lawyers and Scientists

Ryan Budish

- **Presently:** Assistant Director for Research, Berkman Klein Center for Internet & Society
- **Focus Areas:** Ethics & Governance of AI, multistakeholder governance models, cybersecurity, online privacy & surveillance, online censorship

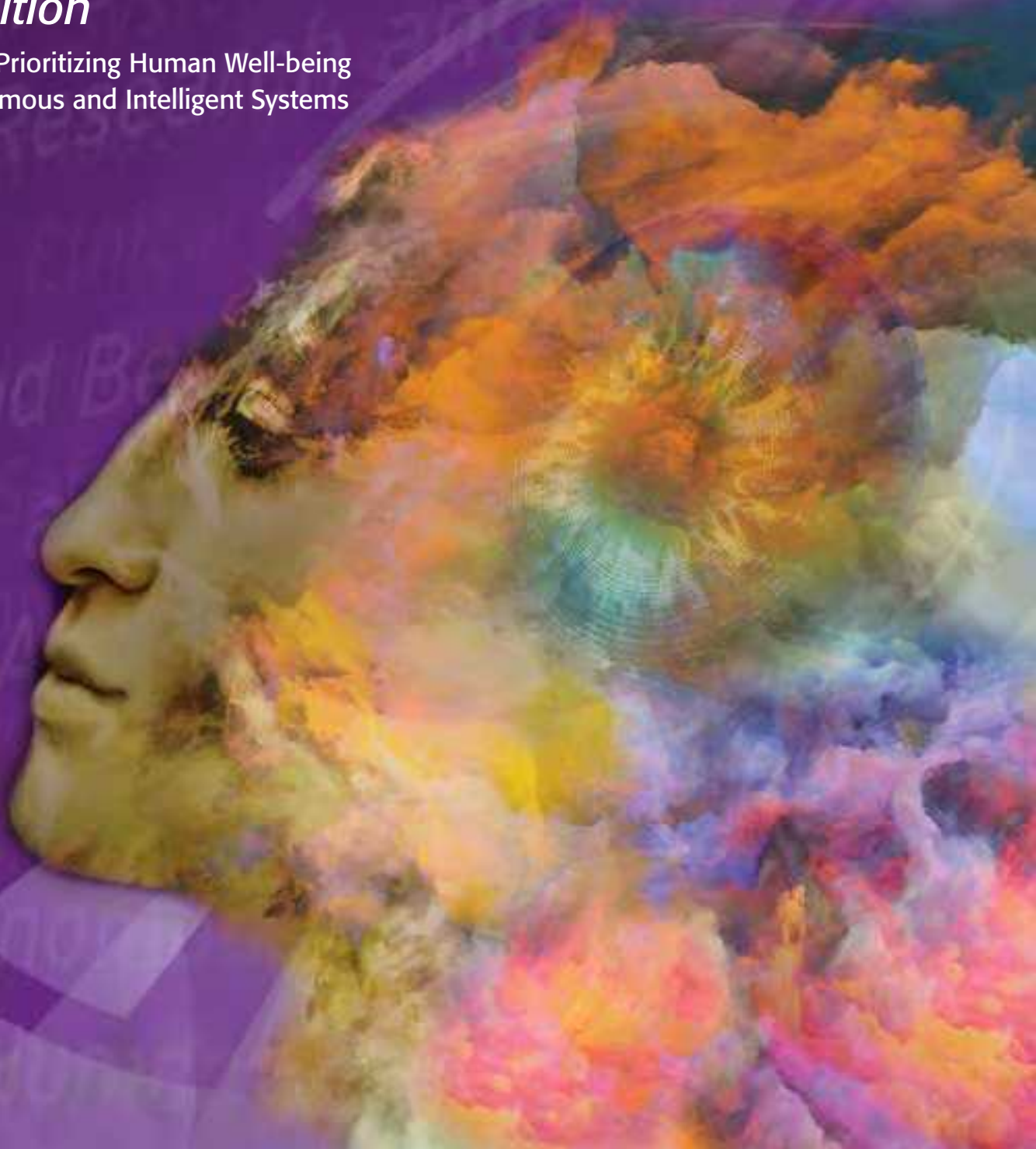
Ryan Budish


- **Examples of BKC activities:**
 - *OECD AI Principles*: Member of the AI Governance Expert Group that drafted AI principles ultimately adopted by 42 countries
 - *United Nations AI Strategy*: Worked with ITU and other UN agencies in developing UN-system strategy for AI
 - *AGTech Forum*: Training state AGs and their senior staff about AI issues
 - *Challenges Forum*: Workshops with companies about their ethical challenges deploying AI technologies
 - *“Accountability of AI Under the Law: The role of explanation”*: review of areas where the law requires explanations from AI systems (2017)

ETHICALLY ALIGNED DESIGN

First Edition

A Vision for Prioritizing Human Well-being
with Autonomous and Intelligent Systems





The views and opinions expressed in this collaborative work are those of the authors and do not necessarily reflect the official policy or position of their respective institutions or of the Institute of Electrical and Electronics Engineers (IEEE). This work is published under the auspices of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the purposes of furthering public understanding of the importance of addressing ethical considerations in the design of autonomous and intelligent systems.

Please see page 290, How the Document Was Prepared, for more details regarding the preparation of this document.

Table of Contents

Introduction	2
Executive Summary	3-6
Acknowledgements	7-8
 <i>Ethically Aligned Design</i>	
From Principles to Practice	9-16
General Principles	17-35
Classical Ethics in A/IS	36-67
Well-being	68-89
Affective Computing	90-109
Personal Data and Individual Agency	110-123
Methods to Guide Ethical Research and Design	124-139
A/IS for Sustainable Development	140-168
Embedding Values into Autonomous and Intelligent Systems	169-197
Policy	198-210
Law	211-281
 <i>About Ethically Aligned Design</i>	
The Mission and Results of The IEEE Global Initiative	282
From Principles to Practice—Results of Our Work to Date	283-284
IEEE P7000™ Approved Standardization Projects	285-286
Who We Are	287
Our Process	288-289
How the Document was Prepared	290
How to Cite <i>Ethically Aligned Design</i>	290
Key References	291

Introduction

As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles. These systems must be developed and should operate in a way that is beneficial to people and the environment, beyond simply reaching functional goals and addressing technical problems. This approach will foster the heightened level of trust between people and technology that is needed for its fruitful use in our daily lives.

To be able to contribute in a positive, non-dogmatic way, we, the techno-scientific communities, need to enhance our self-reflection. We need to have an open and honest debate around our explicit or implicit values, including our imaginary¹ around so-called "Artificial Intelligence" and the institutions, symbols, and representations it generates.

Ultimately, our goal should be *eudaimonia*, a practice elucidated by Aristotle that defines human well-being, both at the individual and collective level, as the highest virtue for a society. Translated roughly as "flourishing", the benefits of *eudaimonia* begin with conscious contemplation, where ethical considerations help us define how we wish to live.

Whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, 墨家/School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability.

The goal of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") is that *Ethically Aligned Design* will provide pragmatic and directional insights and recommendations, serving as a key reference for the work of technologists, educators and policymakers in the coming years.

Ethically Aligned Design sets forth scientific analysis and resources, high-level principles, and actionable recommendations. It offers specific guidance for standards, certification, regulation or legislation for design, manufacture, and use of A/IS that provably aligns with and improves holistic societal well-being.

¹The symbols, values, institutions, and norms of a societal group through which people imagine their lives and constitute their societies.

Introduction

Executive Summary

I. Purpose of *Ethically Aligned Design, First Edition (EAD1e)*

Autonomous and intelligent technical systems are specifically designed to reduce the necessity for human intervention in our day-to-day lives. In so doing, these new systems are also raising concerns about their impact on individuals and societies. Current discussions include advocacy for a positive impact, such as optimization of processes and resource usage, more informed planning and decisions, and recognition of useful patterns in big data. Discussions also include warnings about potential harm to privacy, discrimination, loss of skills, adverse economic impacts, risks to security of critical infrastructure, and possible negative long-term effects on societal well-being.

Because of their nature, the full benefit of these technologies will be attained only if they are aligned with society's defined values and ethical principles. Through this work we intend, therefore, to establish frameworks to guide and inform dialogue and debate around the non-technical implications of these technologies, in particular related to ethical aspects. We understand "ethical" to go beyond moral constructs and include social fairness, environmental sustainability, and our desire for self-determination.

Our analyses and recommendations in *Ethically Aligned Design* address values and intentions as well as implementations, both legal and technical. They are both aspirational, what we hope or wish should happen, and practical, what we—the techno-scientific community and every group involved with and/or affected by these technologies—could do for society to advance in positive directions. The analyses and recommendations in EAD1e are offered as guidance for consideration by governments, businesses, and the public at large in the advancement of technology for the benefit of humanity.

Chapters in *Ethically Aligned Design, First Edition*

- | | |
|--|---|
| 1. From Principles to Practice | 7. Methods to Guide Ethical Research and Design |
| 2. General Principles | 8. A/IS for Sustainable Development |
| 3. Classical Ethics in A/IS | 9. Embedding Values into Autonomous and Intelligent Systems |
| 4. Well-being | 10. Policy |
| 5. Affective Computing | 11. Law |
| 6. Personal Data and Individual Agency | |

Introduction

II. General Principles

The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

1. Human Rights

A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

2. Well-being

A/IS creators shall adopt increased human well-being as a primary success criterion for development.

3. Data Agency

A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

4. Effectiveness

A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

5. Transparency

The basis of a particular A/IS decision should always be discoverable.

6. Accountability

A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

7. Awareness of Misuse

A/IS creators shall guard against all potential misuses and risks of A/IS in operation.

8. Competence

A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

III. Ethical Foundations

Classical Ethics

By drawing from over two thousand five hundred years of classical ethics traditions, the authors of *Ethically Aligned Design* explored established ethics systems, addressing both scientific and religious approaches, including secular philosophical traditions, to address human morality in the digital age. Through reviewing the philosophical foundations that define autonomy and ontology, this work addresses the alleged potential for autonomous capacity of intelligent technical systems, morality in amoral systems, and asks whether decisions made by amoral systems can have moral consequences.

IV. Areas of Impact

A/IS for Sustainable Development

Through affordable and universal access to communications networks and the Internet, autonomous and intelligent systems can be made available to and benefit populations anywhere. They can significantly alter institutions and institutional relationships toward more human-centric structures, and they can address humanitarian and sustainable development issues resulting in increased individual societal and environmental well-being. Such efforts could be facilitated through the recognition of and adherence to established indicators of societal flourishing such as the United Nations Sustainable Development Goals so that human well-being is utilized as a primary success criteria for A/IS development.

Introduction

Personal Data Rights and Agency Over Digital Identity

People have the right to access, share, and benefit from their data and the insights it provides. Individuals require mechanisms to help create and curate the terms and conditions regarding access to their identity and personal data, and to control its safe, specific, and finite exchange. Individuals also require policies and practices that make them explicitly aware of consequences resulting from the aggregation or resale of their personal information.

Legal Frameworks for Accountability

The convergence of autonomous and intelligent systems and robotics technologies has led to the development of systems with attributes that simulate those of human beings in terms of partial autonomy, ability to perform specific intellectual tasks, and even a human physical appearance. The issue of the legal status of complex autonomous and intelligent systems thus intertwines with broader legal questions regarding how to ensure accountability and allocate liability when such systems cause harm. It is clear that:

- Autonomous and intelligent technical systems should be subject to the applicable regimes of property law.
- Government and industry stakeholders should identify the types of decisions and operations that should never be delegated to such systems. These stakeholders should adopt rules and standards that ensure effective human control over those decisions and how to allocate legal responsibility for harm caused by them.

- The manifestations generated by autonomous and intelligent technical systems should, in general, be protected under national and international laws.
- Standards of transparency, competence, accountability, and evidence of effectiveness should govern the development of autonomous and intelligent systems.

Policies for Education and Awareness

Effective policy addresses the protection and promotion of human rights, safety, privacy, and cybersecurity, as well as the public understanding of the potential impact of autonomous and intelligent technical systems on society. To ensure that they best serve the public interest, policies should:

- Support, promote, and enable internationally recognized legal norms.
- Develop government expertise in related technologies.
- Ensure governance and ethics are core components in research, development, acquisition, and use.
- Regulate to ensure public safety and responsible system design.
- Educate the public on societal impacts of related technologies.

Introduction

V. Implementation

Well-being Metrics

For autonomous and intelligent systems to provably advance a specific benefit for humanity, there need to be clear indicators of that benefit. Common metrics of success include profit, gross domestic product, consumption levels, and occupational safety. While important, these metrics fail to encompass the full spectrum of well-being for individuals, the environment, and society. Psychological, social, economic fairness, and environmental factors matter. Well-being metrics include such factors, allowing the benefits arising from technological progress to be more comprehensively evaluated, providing opportunities to test for unintended negative consequences that could diminish human well-being. A/IS can improve capturing of and analyzing the pertinent data, which in turn could help identify where these systems would increase human well-being, providing new routes to societal and technological innovation.

Embedding Values into Autonomous and Intelligent Systems

If machines engage in human communities as quasi-autonomous agents, then those agents must be expected to follow the community's social and moral norms. Embedding norms in such quasi-autonomous systems requires a clear delineation of the community in which they are to be deployed. Further, even within a particular community, different types of technical embodiments will demand different sets of norms. The first step is to identify the norms of the specific community in which the systems

are to be deployed and, in particular, norms relevant to the kinds of tasks that they are designed to perform.

Methods to Guide Ethical Research and Design

To create autonomous and intelligent technical systems that enhance and extend human well-being and freedom, values-based design methods must put human advancement at the core of development of technical systems. This must be done in concert with the recognition that machines should serve humans and not the other way around. Systems developers should employ values-based design methods in order to create sustainable systems that can be evaluated in terms of not only providing increased economic value for organizations but also of broader social costs and benefits.

Affective Computing

Affect is a core aspect of intelligence. Drives and emotions such as anger, fear, and joy are often the foundations of actions throughout our lives. To ensure that intelligent technical systems will be used to help humanity to the greatest extent possible in all contexts, autonomous and intelligent systems that participate in or facilitate human society should not cause harm by either amplifying or dampening human emotional experience.

Introduction

Acknowledgements

Our progress and the ongoing positive influence of this work are due to the volunteer experts serving on all our Committees and IEEE P7000™ Standards Working Groups, along with the IEEE professional staff who support our efforts. Thank you for your dedication toward defining, designing, and inspiring the ethical principles and standards that will ensure that autonomous and intelligent systems and the technologies associated with them will positively benefit humanity.

We wish to thank the Executive Committee and Committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems:

Executive Committee Officers

Raja Chatila, *Chair*

Kay Firth-Butterfield, *Vice Chair*

John C. Havens, *Executive Director*

Executive Committee Members

Dr. Greg Adamson, Karen Bartleson, Virginia Dignum, Danit Gal, Malavika Jayaram, Sven Koenig, Eileen M. Lach, Raj Madhavan, Richard Mallah, AJung Moon, Monique Morrow, Francesca Rossi, Alan Winfield, and Hagit Messer Yaron

Committee Chairs

- **General Principles:** Mark Halverson, and Peet van Biljon
- **Embedding Values into Autonomous Intelligent Systems:** Francesca Rossi and Bertram F. Malle
- **Methodologies to Guide Ethical Research and Design:** Raja Chatila and Corinne Cath
- **Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI):** Malo Bourgon and Richard Mallah
- **Personal Data and Individual Agency:** Katryna Dow and John C. Havens
- **Reframing Autonomous Weapons Systems:** Peter Asaro
- **Sustainable Development:** Elizabeth Gibbons
- **Law:** Nicolas Economou and John Casey
- **Affective Computing:** John Sullins and Joanna J. Bryson
- **Classical Ethics in A/IS:** Jared Bielby
- **Policy:** Peter Brooks and Mina Hannah
- **Extended Reality:** Monique Morrow and Jay Iorio
- **Well-being:** Laura Musikanski and John C. Havens
- **Editing:** Karen Bartleson and Eileen M. Lach
- **Outreach:** Maya Zuckerman and Ali Muzaffar
- **Communications:** Leanne Seeto and Mark Halverson
- **High School:** Tess Posner
- **Global Coordination:** Victoria Wang, Arisa Ema, Pavel Gotovtsev

Introduction

Programs and Projects Inspired by The IEEE Global Initiative:

- ***Ethically Aligned Design University Consortium:*** Hagit Messer, *Chair*
- ***Ethically Aligned Design Community:*** Lisa Morgan, Program Director, Content and Community
- ***Ethics Certification Program for Autonomous and Intelligent Systems:*** Meeri Haataja, *Chair*; Ali Hessami, *Vice-Chair*
- ***Glossary:*** Sara M. Jordan, *Chair*

People

We would like to warmly recognize the leadership and constant support of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems by Dr. Ing. Konstantinos Karachalios, Managing Director of the IEEE Standards Association.

We would also like to thank Stephen Welby, Executive Director and Chief Operating Officer of IEEE for his generous and insightful support of the *Ethically Aligned Design*, First Edition process and The IEEE Global Initiative overall.

We would especially like to thank Eileen M. Lach, the former IEEE General Counsel and Chief Compliance Officer, whose heartfelt conviction that there is a pressing need to focus the global community on highlighting ethical considerations in the development of autonomous and intelligent systems served as a strong catalyst for the development of the Initiative within IEEE.

Finally, we would like to also acknowledge the ongoing work of three Committees of The IEEE Global Initiative regarding their chapters of *Ethically Aligned Design* that, for timing reasons, we were not able to include in *Ethically Aligned Design*, First Edition. These Committees include: Reframing Autonomous Weapons Systems, Extended Reality (formerly Mixed Reality) and Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI). We would like to thank Peter Asaro, Monique Morrow and Jay Iorio, Malo Bourgon and Richard Mallah for their leadership in these groups along with all their Committee Members. Once these chapters have completed their review and been accepted by IEEE they could either be included in *Ethically Aligned Design*, published by The IEEE Global Initiative, or in other publications of IEEE.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

From Principles to Practice

Ethically Aligned Design Conceptual Framework

Ethically Aligned Design, First Edition (EAD1e) represents more than a comprehensive report, distilling the consensus of its vast community of creators into a set of high-level ethical principles, key issues, and practical recommendations. EAD1e is an in-depth seminal work, a one-of-a-kind treatise, intended not only to inform a broader public but also to inspire its audience and readership of academics, engineers, policy makers, and manufacturers of autonomous and intelligent systems¹ (A/IS) to take action.

This Chapter, “From Principles to Practice”, provides a mapping of the conceptual framework of *Ethically Aligned Design*. It outlines the logic behind “Three Pillars” that form the basis of EAD1e, and it connects the Pillars to high-level “General Principles” which guide all manner of ethical A/IS design. Following this, the content of the Chapters of EAD1e is mapped to the Principles. Finally, examples of EAD1e already in practice are described.

Sections in this Chapter:

- The Three Pillars of the *Ethically Aligned Design* Conceptual Framework
- The General Principles of *Ethically Aligned Design*
- Mapping the Pillars to the Principles
- Mapping the Principles to the Content of the Chapters
- From Principles to Practice
- *Ethically Aligned Design* in Implementation

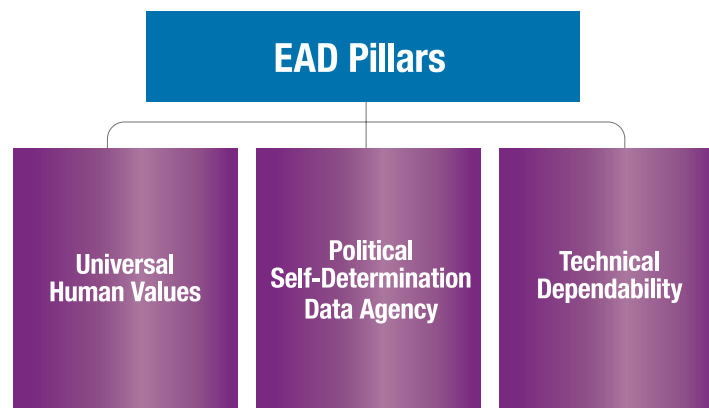
From Principles to Practice

Ethically Aligned Design Conceptual Framework

The Three Pillars of the *Ethically Aligned Design* Conceptual Framework

The Pillars of the *Ethically Aligned Design* Conceptual Framework fall broadly into three areas, reflecting anthropological, political, and technical aspects:

- 1. Universal Human Values:** A/IS can be an enormous force for good in society provided they are designed to respect human rights, align with human values, and holistically increase well-being while empowering as many people as possible. They should also be designed to safeguard our environment and natural resources. These values should guide policy makers as well as engineers, designers, and developers. Advances in A/IS should be in the service of all people, rather than benefiting solely small groups, a single nation, or a corporation.
- 2. Political Self-Determination and Data Agency:** A/IS—if designed and implemented properly—have a great potential to nurture political freedom and democracy, in accordance with the cultural precepts of individual societies, when people have access to and control over the data constituting and representing their identity. These systems can improve government effectiveness and accountability, foster trust, and protect our private sphere, but only when people have agency over their digital identity and their data is provably protected.
- 3. Technical Dependability:** Ultimately, A/IS should deliver services that can be trusted.² This trust means that A/IS will reliably, safely, and actively accomplish the objectives for which they were designed while advancing the human-driven values they were intended to reflect. Technologies should be monitored to ensure that their operation meets predetermined ethical objectives aligning with human values and respecting codified rights. In addition, validation and verification processes, including aspects of explainability, should be developed that could lead to better auditability and to certification³ of A/IS.



The General Principles of *Ethically Aligned Design*

The General Principles of *Ethically Aligned Design* have emerged through the continuous work of dedicated, open communities in a multi-year, creative, consensus-building process. They articulate high-level principles that should apply to all types of autonomous and intelligent systems (A/IS). Created to guide behavior and inform standards and policy making, the General Principles define imperatives for the ethical design, development, deployment, adoption, and decommissioning of autonomous and intelligent systems. The Principles consider the role of A/IS creators, i.e., those who design and manufacture, of operators, i.e., those with expertise specific to use of A/IS, other users, and any other stakeholders or affected parties.

The General Principles⁴ of *Ethically Aligned Design*

- 1. Human Rights**—A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
- 2. Well-being**—A/IS creators shall adopt increased human well-being as a primary success criterion for development.
- 3. Data Agency**—A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

- 4. Effectiveness**—A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
- 5. Transparency**—The basis of a particular A/IS decision should always be discoverable.
- 6. Accountability**—A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
- 7. Awareness of Misuse**—A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
- 8. Competence**—A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.



From Principles to Practice

Ethically Aligned Design Conceptual Framework

Mapping the Pillars to the Principles

Whereas the Pillars of the *Ethically Aligned Design* Conceptual Framework represent broad anthropological, political, and technical aspects relating to autonomous and intelligent systems, the General Principles provide contextual filters for deeper analysis and pragmatic implementation.

It is also important to recognize that the General Principles do not live in isolation of EAD's Pillars and vice versa. While the General Principle of "Transparency" may inform the design of a specific autonomous or intelligent system, the A/IS must also account for universal human values, political self-determination, and data agency. Moreover, Transparency goes beyond technical features. It is an important requirement also for the processes of policy and lawmaking. In this way, EAD1e's Pillars form the holistic ethical grounding upon which the Principles can build, and the latter may apply in various spheres of human activity.

EAD1e Pillars Mapped to General Principles

		EAD Pillars		
		Universal Human Values	Political Self-Determination Data Agency	Technical Dependability
EAD General Principles	Human Rights	■	■	
	Well-being	■	■	
	Data Agency	■	■	■
	Effectiveness			■
	Transparency	■	■	■
	Accountability	■	■	■
	Awareness of Misuse			■
	Competence			■

■ Indicates General Principle mapped to Pillar.

From Principles to Practice

Ethically Aligned Design Conceptual Framework

Mapping the Principles to the Content of the Chapters

The Chapters of *Ethically Aligned Design* provide in-depth subject matter expertise that allows readers to move from the General Principles to more deeply analyze ethical A/IS issues within the context of their specific work.

The mapping or indexing provided in the table below serve as directional starting points since elements of a Principle like “Competence” may resonate in several EAD1e Chapters. In addition, where core subjects are primarily covered by specific Chapters, we have done our best to indicate this via our mapping below.

EAD1e General Principles Mapped to Chapters

		EAD Chapters									
		General Principles	Classical Ethics in A/IS	Well-being	Affective Computing	Data & Individual Agency	Methods A/IS Design	A/IS for Sustainable Dev.	Embedding Values into A/IS	Policy	Law
EAD General Principles	Human Rights	■	■	■	■	■	■	■	■	■	■
	Well-being	■	■	■ ■ ■	■	■		■	■	■	■
	Data Agency	■		■	■	■ ■ ■	■	■	■	■	
	Effectiveness	■			■		■		■	■	■
	Transparency	■			■		■		■	■	■
	Accountability	■			■		■	■	■	■	■
	Awareness of Misuse	■	■		■		■		■	■	■
	Competence	■			■		■		■	■	■

■ Indicates General Principle mapped to Chapter.

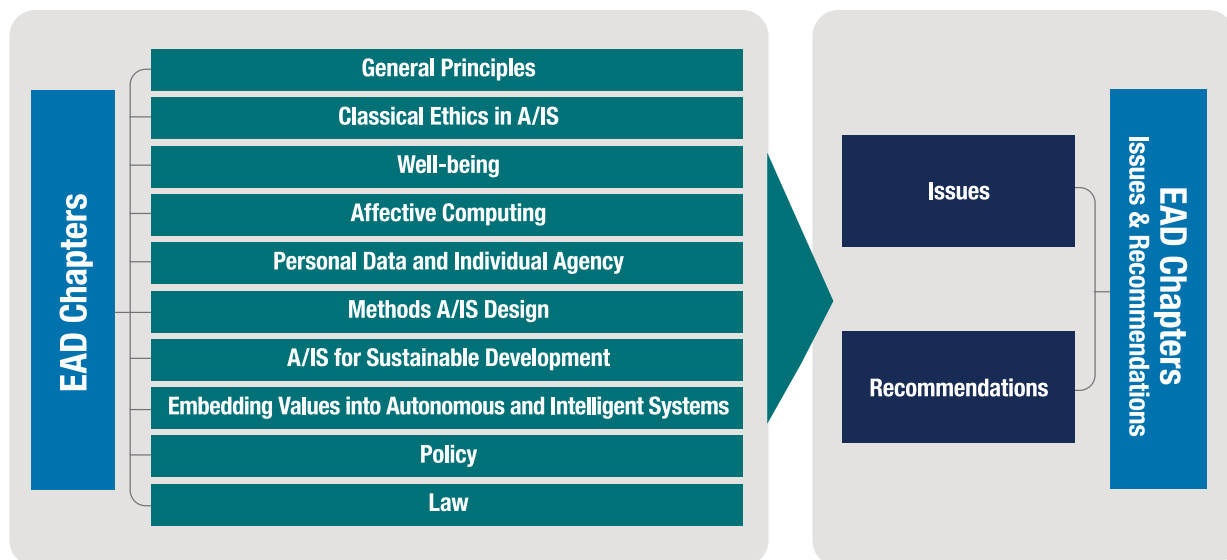
■ Indicates primary EAD Chapter providing elaboration on a General Principle.

From Principles to Practice

Ethically Aligned Design Conceptual Framework

From Principles to Practice

It is at this step of the *Ethically Aligned Design* Conceptual Framework that readers will be able to identify the Principles and Chapters of key relevance to their work. Content provided in EAD1e Chapters is organized by “Issues” identified as the most pressing ethical matters surrounding A/IS design to address today and “Recommendations” on how it should be done. By reviewing these Issues and Recommendations in light of a specific A/IS product, service, or system being designed, readers are provided with a simple form of impact assessment and due diligence process to help put their “Principles into Practice” for themselves. Of course, more fine-tuned customization and adaptation of the content of EAD1e to fit specific sectors or applications are possible and will be pursued in the near future. See below for some implementation examples already happening.



Ethically Aligned Design in Implementation

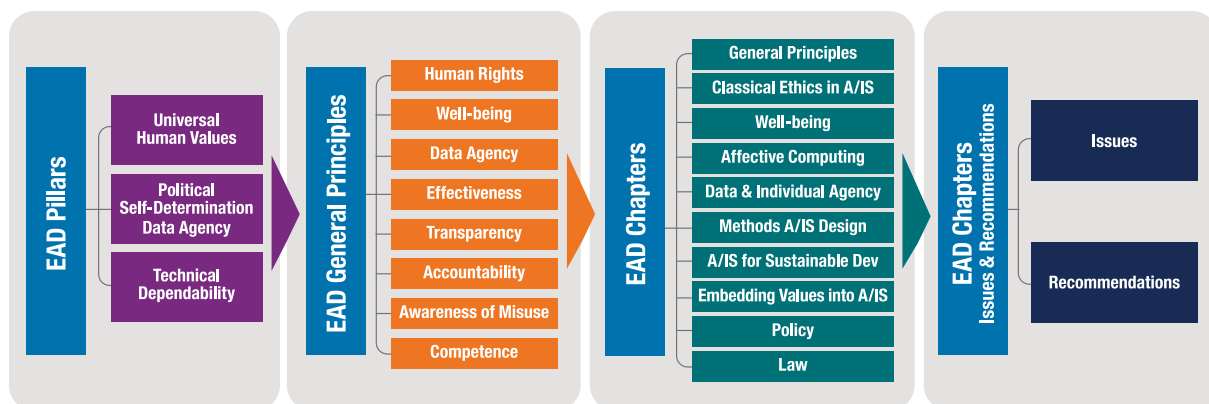
Ethically Aligned Design, First Edition represents the culmination of a three-year process guided bottom-up since 2015 by the rigor and standards of the engineering profession and by a globally open and iterative process involving hundreds of global experts. The analysis of the Principles, Issues, and Recommendations generated as part of an iterative process have already inspired the creation of fourteen IEEE Standardization Projects, a Certification Program, A/IS Ethics Courses, and multiple other action-oriented programs currently in development.

In its earlier manifestations, *Ethically Aligned Design* informed collaborations on A/IS governance with a broad range of governmental and civil society organizations, including the United Nations, the European Commission, the Organization for Economic Cooperation and Development and many national and municipal governments and institutions.⁵ Moreover, the engagement in all of these arenas and with such partners has put the collective knowledge and creativity of The IEEE Global Initiative in the service of global policy-making with tangible and visible results. Beyond inspiring the policy arena, EAD1e and this growing body of work has also been influencing the development of industry-related resources.⁶

It is time to move “From Principles to Practice” in society regarding the governance of emerging autonomous and intelligent systems. The implementation of ethical principles must be validated by dependable applications of A/IS in practice while honoring our desire for political self-determination and data agency. To achieve societal progress, the autonomous and intelligent systems we create must be trustworthy, provable, and accountable and must align to our explicitly formulated human values.

It is our hope that *Ethically Aligned Design* and this conceptual framework will provide action-oriented inspiration for your work as well.

Ethically Aligned Design Conceptual Framework—From Principles to Practice



For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

From Principles to Practice

Ethically Aligned Design Conceptual Framework

Endnotes

¹ We prefer not to use—as far as possible—the vague term “AI” and use instead the term autonomous and intelligent systems (A/IS). This terminology is applied throughout *Ethically Aligned Design, First Edition* to ensure the broadest possible application of ethical considerations in the design of the addressed technologies and systems.

² See also [Draft Ethics Guidelines for Trustworthy AI](#) of The European Commission’s High Level Expert Group on AI.

³ A/IS should be subject to specific certification procedures by competent and qualified agencies with participation or control of public authorities in the same way other technical systems require certification before deployment. The IEEE has launched one of the world’s first programs dedicated to creating A/IS certification processes. [The Ethics Certification Program for Autonomous and Intelligent Systems](#) (ECPAIS) offers processes by which organizations can seek certified A/IS products, systems, and services. It is being developed through an extensive and open public-private collaboration.

⁴ For their overall framing, see the “General Principles” Chapter.

⁵ As an example, the recently published report [Draft Ethics Guidelines for Trustworthy AI](#) of The European Commission’s High Level Expert Group on AI explicitly mentions EAD as a major source of their inspiration. EAD has also been guiding policy creation for efforts of the United Nations and the Organization for Economic Cooperation and Development.

⁶ [Everyday Ethics for Artificial Intelligence: A Practical Guide for Designers and Developers](#)

General Principles

The General Principles of *Ethically Aligned Design* articulate high-level ethical principles that apply to all types of autonomous and intelligent systems (A/IS), regardless of whether they are physical robots, such as care robots or driverless cars, or software systems, such as medical diagnosis systems, intelligent personal assistants, or algorithmic chat bots, in real, virtual, contextual, and mixed-reality environments.

The General Principles define imperatives for the design, development, deployment, adoption, and decommissioning of autonomous and intelligent systems. The Principles consider the role of A/IS creators, i.e., those who design and manufacture, of operators, i.e., those with expertise specific to use of A/IS, other users, and any other stakeholders or affected parties.

We have created these ethical General Principles for A/IS that:

- Embody the highest ideals of human beneficence within human rights.
- Prioritize benefits to humanity and the natural environment from the use of A/IS over commercial and other considerations. Benefits to humanity and the natural environment should not be at odds—the former depends on the latter. Prioritizing human well-being does not mean degrading the environment.
- Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems, in particular by ensuring actions of A/IS are accountable and transparent.

These General Principles are elaborated in subsequent sections of this chapter of *Ethically Aligned Design*, with specific contextual, cultural, and pragmatic explorations which impact their implementation.

General Principles

General Principles as Imperatives

We offer high-level General Principles in *Ethically Aligned Design* that we consider to be imperatives for creating and operating A/IS that further human values and ensure trustworthiness. In summary, our General Principles are:

1. **Human Rights**—A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being**—A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency**—A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people’s capacity to have control over their identity.
4. **Effectiveness**—A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. **Transparency**—The basis of a particular A/IS decision should always be discoverable.
6. **Accountability**—A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse**—A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence**—A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

General Principles

Principle 1—Human Rights

A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

Background

Human benefit is a crucial goal of A/IS, as is respect for human rights set out in works including, but not limited to: [The Universal Declaration of Human Rights](#), the [International Covenant on Civil and Political Rights](#), the [Convention on the Rights of the Child](#), the [Convention on the Elimination of all forms of Discrimination against Women](#), the [Convention on the Rights of Persons with Disabilities](#), and the [Geneva Conventions](#).

Such rights need to be fully taken into consideration by individuals, companies, professional bodies, research institutions, and governments alike to reflect the principle that A/IS should be designed and operated in a way that both respects and fulfills human rights, freedoms, human dignity, and cultural diversity.

While their interpretation may change over time, “human rights”, as defined by international law, provide a unilateral basis for creating any A/IS, as these systems affect humans, their emotions,

data, or agency. While the direct coding of human rights in A/IS may be difficult or impossible based on contextual use, newer guidelines from The United Nations provide methods to pragmatically implement human rights ideals within business or corporate contexts that could be adapted for engineers and technologists. In this way, technologists can take into account human rights in the way A/IS are developed, operated, tested, and validated. In short, human rights should be part of the ethical risk assessment of A/IS.

Recommendations

To best respect human rights, society must assure the safety and security of A/IS so that they are designed and operated in a way that benefits humans. Specifically:

- Governance frameworks, including standards and regulatory bodies, should be established to oversee processes which ensure that the use of A/IS does not infringe upon human rights, freedoms, dignity, and privacy, and which ensure traceability. This will contribute to building public trust in A/IS.
- A way to translate existing and forthcoming legal obligations into informed policy and technical considerations is needed. Such a method should allow for diverse cultural norms as well as differing legal and regulatory frameworks.

General Principles

- A/IS should always be subordinate to human judgment and control.
- For the foreseeable future, A/IS should not be granted rights and privileges equal to human rights.

Further Resources

The following documents and organizations are provided both as references and examples of the types of work that can be emulated, adapted, and proliferated regarding ethical best practices around A/IS to best honor human rights:

- [The Universal Declaration of Human Rights](#), 1947.
- N. Wiener, *The Human Use of Human Beings*, New York: Houghton Mifflin, 1954.
- [The International Covenant on Civil and Political Rights](#), 1966.
- [The International Covenant on Economic, Social and Cultural Rights](#), 1966.
- [The International Convention on the Elimination of All Forms of Racial Discrimination](#), 1965.
- [The Convention on the Rights of the Child](#), 1990.
- [The Convention on the Elimination of All Forms of Discrimination against Women](#), 1979.
- [The Convention on the Rights of Persons with Disabilities](#), 2006.
- [The Geneva Conventions and Additional Protocols](#), 1949.
- [IRTF's Research into Human Rights Protocol Considerations](#), 2018.
- [The UN Guiding Principles on Business and Human Rights](#), 2011.
- British Standards Institute BS8611:2016, Robots and Robotic Devices. [Guide to the Ethical Design and Application of Robots and Robotic Systems](#)

General Principles

Principle 2—Well-being

A/IS creators shall adopt increased human well-being as a primary success criterion for development.

Background

For A/IS technologies to demonstrably advance benefit for humanity, we need to be able to define and measure the benefit we wish to increase. But often the only indicators utilized in determining success for A/IS are avoiding negative unintended consequences and increasing productivity and economic growth for customers and society. Today, these are largely measured by gross domestic product (GDP), profit, or consumption levels.

Well-being, for the purpose of *Ethically Aligned Design*, is based on the Organization for Economic Co-operation and Development's (OECD) "[Guidelines on Measuring Subjective Well-being](#)" perspective that, "Being able to measure people's quality of life is fundamental when assessing the progress of societies." There is now widespread acknowledgement that measuring subjective well-being is an essential part of measuring quality of life alongside other social and economic dimensions as identified within [Nassbaum-Sen's capability approach](#) whereby well-being is objectively defined in terms of human capabilities necessary for functioning and flourishing.

Since modern societies will be largely constituted of A/IS users, we believe these considerations to be relevant for A/IS creators.

A/IS technologies can be narrowly conceived from an ethical standpoint. They can be legal, profitable, and safe in their usage, yet not positively contribute to human and environmental well-being. This means technologies created with the best intentions, but without considering well-being, can still have dramatic negative consequences on people's mental health, emotions, sense of themselves, their autonomy, their ability to achieve their goals, and other dimensions of well-being.

Recommendation

A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being metrics as their reference point.

Further Resources

- IEEE P7010™, [Well-being Metric for Autonomous and Intelligent Systems](#).
- [The Measurement of Economic Performance and Social Progress](#) now commonly referred to as "The Stiglitz Report", commissioned by the then President of the French Republic, 2009. From the report: "...the time is ripe for our measurement system to shift emphasis from measuring economic production to measuring

General Principles

- people's well-being ... emphasizing well-being is important because there appears to be an increasing gap between the information contained in aggregate GDP data and what counts for common people's well-being."
- [OECD Guidelines on Measuring Subjective Well-being](#), 2013.
 - [OECD Better Life Index](#), 2017.
 - [World Happiness Reports](#), 2012 – 2018.
 - United Nations [Sustainable Development Goal \(SDG\) Indicators](#), 2018.
 - [Beyond GDP](#), European Commission, 2018. From the site: "The Beyond GDP initiative is about developing indicators that are as clear and appealing as GDP, but more inclusive of environmental and social aspects of progress."
 - [Genuine Progress Indicator](#), State of Maryland (first developed by Redefining Progress), 2015.
 - The International Panel on Social Progress, [Social Justice, Well-Being and Economic Organization](#), 2018.
 - R. Veenhoven, World Database of Happiness, Erasmus University Rotterdam, The Netherlands, Accessed 2018 at: <http://worlddatabaseofhappiness.eur.nl>.
 - Royal Government of Bhutan, [The Report of the High-Level Meeting on Wellbeing and Happiness: Defining a New Economic Paradigm](#), New York: The Permanent Mission of the Kingdom of Bhutan to the United Nations, 2012.

General Principles

Principle 3—Data Agency

A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people’s capacity to have control over their identity.

Background

Digital consent is a misnomer in its current manifestation. Terms and conditions or privacy policies are largely designed to provide legally accurate information regarding the usage of people’s data to safeguard institutional and corporate interests, while often neglecting the needs of the people whose data they process. “Consent fatigue”, the constant request for agreement to sets of long and unreadable data handling conditions, causes a majority of users to simply click and accept terms in order to access the services they wish to use. General obfuscation regarding privacy policies, and scenarios like the [Cambridge Analytica scandal](#) in 2018, demonstrate that even when individuals provide consent, the understanding of the value regarding their data and its safety is out of an individual’s control.

This existing model of data exchange has eroded human agency in the algorithmic age. People don’t know how their data is being used at all times or when predictive messaging is honoring their existing preferences or manipulating them to create new behaviors.

Regulations like the [EU General Data Protection Regulation](#) (GDPR) will help improve this lack of clarity regarding the exchange of personal data. But compliance with existing models of consent is not enough to safeguard people’s agency regarding their personal information. In an era where A/IS are already pervasive in society, governments must recognize that limiting the misuse of personal data is not enough.

Society must also recognize that human rights in the digital sphere don’t exist until individuals globally are empowered with means—including tools and policies—that ensure their dignity through some form of sovereignty, agency, symmetry, or control regarding their identity and personal data. These rights rely on individuals being able to make their choices, outside of the potential influence of biased algorithmic messaging or bad actors. Society also needs to be confident that those who are unable to provide legal informed consent, including minors and people with diminished capacity to make informed decisions, do not lose their dignity due to this.

Recommendation

Organizations, including governments, should immediately explore, test, and implement technologies and policies that let individuals specify their online agent for case-by-case authorization decisions as to who can process what personal data for what purpose. For minors and those with diminished capacity to make informed decisions, current guardianship approaches should be viewed to determine their suitability in this context.

General Principles

The general solution to give agency to the individual is meant to anticipate and enable individuals to own and fully control autonomous and intelligent (as in capable of learning) technology that can evaluate data use requests by external parties and service providers. This technology would then provide a form of “digital sovereignty” and could issue limited and specific authorizations for processing of the individual’s personal data wherever it is held in a compatible system.

Further Resources

The following resources are designed to provide governments and other organizations—corporate, for-profit, not-for-profit, B Corp, or any form of public institution—basic information on services designed to provide user agency and/or sovereignty over their personal data.

- The European Data Protection Supervisor [defines personal information management systems](#) (PIMS) as:
- “...systems that help give individuals more control over their personal data...allowing individuals to manage their personal data in secure, local or online storage systems and share them when and with whom they choose. Providers of online services and advertisers will need to interact with the PIMS if they plan to process individuals’ data. This can enable a human centric approach to personal information and new business models.” For further information and ongoing research regarding PIMS, visit [Ctrl-Shift’s PIMS monthly archive](#).
- IEEE P7006™, [IEEE Standards Project for Personal Data Artificial Intelligence \(AI\) Agent](#) describes the technical elements required to create and grant access to a personalized Artificial Intelligence that will comprise inputs, learning, ethics, rules, and values controlled by individuals.
- IEEE P7012™, [IEEE Standards Project for Machine Readable Personal Privacy Terms](#) is designed to provide individuals with a means to proffer their own terms respecting personal privacy in ways that can be read, acknowledged, and be agreed to by machines operated by others in the networked world.

General Principles

Principle 4—Effectiveness

Creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

Background

The responsible adoption and deployment of A/IS are essential if such systems are to realize their many potential benefits to the well-being of both individuals and societies. A/IS will not be trusted unless they can be shown to be effective in use. Harms caused by A/IS, from harm to an individual through to systemic damage, can undermine the perceived value of A/IS and delay or prevent its adoption.

Operators and other users will therefore benefit from measurement of the effectiveness of the A/IS in question. To be adequate, effective measurements need to be both valid and accurate, as well as meaningful and actionable. And such measurements must be accompanied by practical guidance on how to interpret and respond to them.

Recommendations

1. Creators engaged in the development of A/IS should seek to define metrics or benchmarks that will serve as valid and meaningful gauges of the effectiveness of the system in meeting its objectives, adhering to standards and remaining within risk tolerances. Creators building A/IS should ensure that the results when the defined metrics are applied are readily obtainable by all interested parties, e.g., users, safety certifiers, and regulators of the system.
2. Creators of A/IS should provide guidance on how to interpret and respond to the metrics generated by the systems.
3. To the extent warranted by specific circumstances, operators of A/IS should follow the guidance on measurement provided with the systems, i.e., which metrics to obtain, how and when to obtain them, how to respond to given results, and so on.
4. To the extent that measurements are sample-based, measurements should account for the scope of sampling error, e.g., the reporting of confidence intervals associated with the measurements. Operators should be advised how to interpret the results.
5. Creators of A/IS should design their systems such that metrics on specific deployments of the system can be aggregated to provide information on the effectiveness of the system across multiple deployments. For example, in the case of autonomous vehicles, metrics should be generated both for a specific instance of a vehicle and for a fleet of many instances of the same kind of vehicle.
6. In interpreting and responding to measurements, allowance should be made for variation in the specific objectives and circumstances of a given deployment of A/IS.

General Principles

7. To the extent possible, industry associations or other organizations, e.g., IEEE and ISO, should work toward developing standards for the measurement and reporting on the effectiveness of A/IS.

Further Resources

- R. Dillmann, [KA 1.10 Benchmarks for Robotics Research](#), 2010.
- A. Steinfeld, T.W. Fong, D. Kaber, J. Scholtz, A. Schultz, and M. Goodrich, "[Common Metrics for Human-Robot Interaction](#)", 2006 Human-Robot Interaction Conference, March, 2006.
- R. Madhavan, E. Messina, and E. Tunstel, Eds., [Performance Evaluation and Benchmarking of Intelligent Systems](#), Boston, MA: Springer, 2009.
- *IEEE Robotics & Automation Magazine*, [Special Issue on Replicable and Measurable Robotics Research](#), Volume 22, No. 3, September 2015.
- C. Flanagan, [A Survey on Robotics Systems and Performance Analysis](#), 2011.
- [Transaction Processing Performance Council \(TPC\) Establishes Artificial Intelligence Working Group \(TPC-AI\)](#) tasked with developing industry standard benchmarks for both hardware and software platforms associated with running Artificial Intelligence (AI) based workloads, 2017.

General Principles

Principle 5—Transparency

The basis of a particular A/IS decision should always be discoverable.

Background

A key concern over autonomous and intelligent systems is that their operation must be transparent to a wide range of stakeholders for different reasons, noting that the level of transparency will necessarily be different for each stakeholder. Transparent A/IS are ones in which it is possible to discover how and why a system made a particular decision, or in the case of a robot, acted the way it did. The term “transparency” in the context of A/IS also addresses the concepts of traceability, explainability, and interpretability.

A/IS will perform tasks that are far more complex and have more effect on our world than prior generations of technology. Where the task is undertaken in a non-deterministic manner, it may defy simple explanation. This reality will be particularly acute with systems that interact with the physical world, thus raising the potential level of harm that such a system could cause. For example, some A/IS already have real consequences to human safety or well-being, such as medical diagnosis or driverless car autopilots. Systems such as these are safety-critical systems.

At the same time, the complexity of A/IS technology and the non-intuitive way in which it may operate will make it difficult for users of those systems to understand the actions of the A/IS that they use, or with which they interact. This opacity, combined with the often distributed manner in which the A/IS are developed, will complicate efforts to determine and allocate responsibility when something goes wrong. Thus, lack of transparency increases the risk and magnitude of harm when users do not understand the systems they are using, or there is a failure to fix faults and improve systems following accidents. Lack of transparency also increases the difficulty of ensuring accountability (see Principle 6—Accountability).

Achieving transparency, which may involve a significant portion of the resources required to develop the A/IS, is important to each stakeholder group for the following reasons:

1. For users, what the system is doing and why.
2. For creators, including those undertaking the validation and certification of A/IS, the systems’ processes and input data.
3. For an accident investigator, if accidents occur.
4. For those in the legal process, to inform evidence and decision-making.
5. For the public, to build confidence in the technology.

General Principles

Recommendation

Develop new standards that describe measurable, testable levels of transparency, so that systems can be objectively assessed and levels of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. The mechanisms by which transparency is provided will vary significantly, including but not limited to, the following use cases:

1. For users of care or domestic robots, a “why-did-you-do-that button” which, when pressed, causes the robot to explain the action it just took.
2. For validation or certification agencies, the algorithms underlying the A/IS and how they have been verified.
3. For accident investigators, secure storage of sensor and internal state data comparable to a flight data recorder or black box.

IEEE P7001™, [IEEE Standard for Transparency of Autonomous Systems](#) is one such standard, developed in response to this recommendation.

Further Resources

- C. Cappelli, P. Engiel, R. Mendes de Araujo, and J. C. Sampaio do Prado Leite, “Managing Transparency Guided by a Maturity Model,” *3rd Global Conference on Transparency Research* 1 no. 3, pp. 1–17, Jouy-en-Josas, France: HEC Paris, 2013.
- J.C. Sampaio do Prado Leite and C. Cappelli, “Software Transparency,” *Business & Information Systems Engineering* 2, no. 3, pp. 127–139, 2010.
- A. Winfield, and M. Jirotko, “The Case for an Ethical Black Box,” *Lecture Notes in Artificial Intelligence* 10454, pp. 262–273, 2017.
- R. R. Wortham, A. Theodorou, and J. J. Bryson, “What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems,” *IJCAI-2016 Ethics for Artificial Intelligence Workshop*, New York, 2016.
- Machine Intelligence Research Institute, “[Transparency in Safety-Critical Systems](#),” August 25, 2013.
- M. Scherer, “[Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies](#),” *Harvard Journal of Law & Technology* 29, no. 2, 2015.
- U.K. House of Commons, “Decision Making Transparency,” [Report of the U.K. House of Commons Science and Technology Committee on Robotics and Artificial Intelligence](#), pp. 17-18, September 13, 2016.

General Principles

Principle 6—Accountability

A/IS shall be created and operated to provide an unambiguous rationale for decisions made.

Background

The programming, output, and purpose of A/IS are often not discernible by the general public. Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the manufacture and deployment of these systems to establish responsibility and accountability, and to avoid potential harm. Additionally, manufacturers of these systems must be accountable in order to address legal issues of culpability. It should, if necessary, be possible to apportion culpability among responsible creators (designers and manufacturers) and operators to avoid confusion or fear within the general public.

Accountability and partial accountability are not possible without transparency, thus this principle is closely linked with Principle 5—Transparency.

Recommendations

To best address issues of responsibility and accountability:

1. Legislatures/courts should clarify responsibility, culpability, liability, and accountability for A/IS, where possible, prior to development and deployment so that manufacturers and users understand their rights and obligations.
2. Designers and developers of A/IS should remain aware of, and take into account, the diversity of existing cultural norms among the groups of users of these A/IS.
3. Multi-stakeholder ecosystems including creators, and government, civil, and commercial stakeholders, should be developed to help establish norms where they do not exist because A/IS-oriented technology and their impacts are too new. These ecosystems would include, but not be limited to, representatives of civil society, law enforcement, insurers, investors, manufacturers, engineers, lawyers, and users. The norms can mature into best practices and laws.

General Principles

4. Systems for registration and record-keeping should be established so that it is always possible to find out who is legally responsible for a particular A/IS. Creators, including manufacturers, along with operators, of A/IS should register key, high-level parameters, including:

- Intended use,
- Training data and training environment, if applicable,
- Sensors and real world data sources,
- Algorithms,
- Process graphs,
- Model features, at various levels,
- User interfaces,
- Actuators and outputs, and
- Optimization goals, loss functions, and reward functions.

Further Resources

- B. Shneiderman, "[Human Responsibility for Autonomous Agents](#)," *IEEE Intelligent Systems* 22, no. 2, pp. 60–61, 2007.
- A. Matthias, "[The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata](#)," *Ethics and Information Technology* 6, no. 3, pp. 175–183, 2004.
- A. Hevelke and J. Nida-Rümelin, "[Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis](#)," *Science and Engineering Ethics* 21, no. 3, pp. 619–630, 2015.
- An example of good practice (in relation to Recommendation #3) can be found in [Sciencewise](#)—the U.K. national center for public dialogue in policy-making involving science and technology issues.

General Principles

Principle 7—Awareness of Misuse

Creators shall guard against all potential misuses and risks of A/IS in operation.

Background

New technologies give rise to greater risk of deliberate or accidental misuse, and this is especially true for A/IS. A/IS increases the impact of risks such as hacking, misuse of personal data, system manipulation, or exploitation of vulnerable users by unscrupulous parties. Cases of A/IS hacking have already been widely reported, with [driverless cars](#), for example. The [Microsoft Tay AI chatbot](#) was famously manipulated when it mimicked deliberately offensive users. In an age where these powerful tools are easily available, there is a need for a new kind of education for citizens to be sensitized to risks associated with the misuse of A/IS. The EU's General Data Protection Regulation (GDPR) provides measures to remedy the misuse of personal data.

Responsible innovation requires A/IS creators to anticipate, reflect, and engage with users of A/IS. Thus, citizens, lawyers, governments, etc., all have a role to play through education and awareness in developing accountability structures (see Principle 6), in addition to guiding new technology proactively toward beneficial ends.

Recommendations

1. Creators should be aware of methods of misuse, and they should design A/IS in ways to minimize the opportunity for these.
2. Raise public awareness around the issues of potential A/IS technology misuse in an informed and measured way by:
 - Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS. For example, provide “data privacy warnings” that some smart devices will collect their users’ personal data.
 - Delivering this education in scalable and effective ways, including having experts with the greatest credibility and impact who can minimize unwarranted fear about A/IS.
 - Educating government, lawmakers, and enforcement agencies about these issues of A/IS so citizens can work collaboratively with these agencies to understand safe use of A/IS. For example, the same way police officers give public safety lectures in schools, they could provide workshops on safe use and interaction with A/IS.

Further Resources

- A. Greenberg, “[Hackers Fool Tesla S's Autopilot to Hide and Spoof Obstacles](#),” *Wired*, August 2016.
- C. Wilkinson and E. Weitkamp, [Creative Research and Communication: Theory and Practice](#), Manchester, UK: Manchester University Press, 2016 (in relation to Recommendation #2).
- Engineering and Physical Sciences Research Council, “[Anticipate, Reflect, Engage and Act \(AREA\)](#),” Framework for Responsible Research and Innovation, Accessed 2018.

General Principles

Principle 8—Competence

Creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

Background

A/IS can and often do make decisions that previously required human knowledge, expertise, and reason. Algorithms potentially can make even better decisions, by accessing more information, more quickly, and without the error, inconsistency, and bias that can plague human decision-making. As the use of algorithms becomes common and the decisions they make become more complex, however, the more normal and natural such decisions appear.

Operators of A/IS can become less likely to question and potentially less able to question the decisions that algorithms make. Operators will not necessarily know the sources, scale, accuracy, and uncertainty that are implicit in applications of A/IS. As the use of A/IS expands, more systems will rely on machine learning where actions are not preprogrammed and that might not leave a clear record of the steps that led the system to its current state. Even if those records do exist, operators might not have access to them or the expertise necessary to decipher those records.

Standards for the operators are essential. Operators should be able to understand how

A/IS reach their decisions, the information and logic on which the A/IS rely, and the effects of those decisions. Even more crucially, operators should know when they need to question A/IS and when they need to overrule them.

Creators of A/IS should take an active role in ensuring that operators of their technologies have the knowledge, experience, and skill necessary not only to use A/IS, but also to use it safely and appropriately, towards their intended ends. Creators should make provisions for the operators to override A/IS in appropriate circumstances.

While standards for operator competence are necessary to ensure the effective, safe, and ethical application of A/IS, these standards are not the same for all forms of A/IS. The level of competence required for the safe and effective operation of A/IS will range from elementary, such as “intuitive” use guided by design, to advanced, such as fluency in statistics.

Recommendations

1. Creators of A/IS should specify the types and levels of knowledge necessary to understand and operate any given application of A/IS. In specifying the requisite types and levels of expertise, creators should do so for the individual components of A/IS and for the entire systems.
2. Creators of A/IS should integrate safeguards against the incompetent operation of their systems. Safeguards could include issuing

General Principles

- notifications/warnings to operators in certain conditions, limiting functionalities for different levels of operators (e.g., novice vs. advanced), system shut-down in potentially risky conditions, etc.
3. Creators of A/IS should provide the parties affected by the output of A/IS with information on the role of the operator, the competencies required, and the implications of operator error. Such documentation should be accessible and understandable to both experts and the general public.
 4. Entities that operate A/IS should create documented policies to govern how A/IS should be operated. These policies should include the real-world applications for such A/IS, any preconditions for their effective use, who is qualified to operate them, what training is required for operators, how to measure the performance of the A/IS, and what should be expected from the A/IS. The policies should also include specification of circumstances in which it might be necessary for the operator to override the A/IS.
 5. Operators of A/IS should, before operating a system, make sure that they have access to the requisite competencies. The operator need not be an expert in all the pertinent domains but should have access to individuals with the requisite kinds of expertise.

Further Resources

- S. Barocas and A.D. Selbst, [The Intuitive Appeal of Explainable Machines](#), Fordham Law Review, 2018.
- W. Smart, C. Grimm, and W. Hartzog, ["An Education Theory of Fault for Autonomous Systems"](#), 2017.

General Principles

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The General Principles Committee

- **Alan Winfield** (Founding Chair) – Professor, Bristol Robotics Laboratory, University of the West of England; Visiting Professor, University of York
- **Mark Halverson** (Co-Chair) – Founder and CEO at Precision Autonomy
- **Peet van Biljon** (Co-Chair) – Founder and CEO at BMNP Strategies LLC, advisor on strategy, innovation, and business transformation; Adjunct professor at Georgetown University; Business ethics author
- **Shahar Avin** – Research Associate, Centre for the Study of Existential Risk, University of Cambridge
- **Bijilash Babu** – Senior Manager, Ernst and Young, EY Global Delivery Services India LLP
- **Richard Bartley** – Senior Director - Analyst, Security & Risk Management, Gartner, Toronto, Canada Security Principal Director, Accenture, Toronto, Canada.
- **R. R. Brooks** – Professor, Holcombe Department of Electrical and Computer Engineering, Clemson University
- **Nicolas Economou** – Chief Executive Officer, H5; Chair, Science, Law and Society Initiative at The Future Society Chair, Law Committee, Global Governance of AI Roundtable; Member, Council on Extended Intelligence (CXI)
- **Hugo Giordano** – Engineering Student at Texas A&M University
- **Alexei Grinbaum** – Researcher at CEA (French Alternative Energies and Atomic Energy Commission) and Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA
- **Jia He** – Independent Researcher, Graduate Delft University of Technology in Engineering and Public Policy, project member within United Nations, ICANN, and ITU Executive Director of Toutiao Research (Think Tank), Bytedance Inc.
- **Bruce Hedin** – Principal Scientist, H5
- **Cyrus Hodes** – Advisor AI Office, UAE Prime Minister's Office, Co-founder and Senior Advisor, AI Initiatives@The Future Society; Member, AI Expert Group at the OECD, Member, Global Council on Extended Intelligence; Co-founder and Senior Advisor, The AI Initiative @ The Future Society
- **Nathan F. Hutchins** – Applied Assistant Professor, Department of Electrical and Computer Engineering, The University of Tulsa
- **Narayana GPL. Mandaleeka ("MGPL")** – Vice President & Chief Scientist, Head, Business Systems & Cybernetics Centre, Tata Consultancy Services Ltd.
- **Vidushi Marda** – Programme Officer, ARTICLE 19
- **George T. Matthew** – Chief Medical Officer, North America, DXC Technology

General Principles

- **Nicolas Mialhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs
- **Rupak Rathore** – Principal Consultant at ATCS for Telematics, Connected Car and Internet of Things; Advisor on strategy, innovation and transformation journey management; Senior Member, IEEE
- **Peter Teneriello** – Investment Analyst, Private Equity and Venture Capital, TMRS
- **Niels ten Oever** – Head of Digital, Article 19, Co-chair Research Group on Human Rights Protocol Considerations in the Internet Research Taskforce (IRTF)
- **Alan R. Wagner** – Assistant Professor, Department of Aerospace Engineering, Research Associate, The Rock Ethics Institute, The Pennsylvania State University.

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Classical Ethics in A/IS

We applied classical ethics methodologies to considerations of algorithmic design in autonomous and intelligent systems (A/IS) where machine learning may or may not reflect ethical outcomes that mimic human decision-making. To meet this goal, we drew from classical ethics theories and the disciplines of machine ethics, information ethics, and technology ethics.

As direct control over tools becomes further removed, creators of autonomous systems must ask themselves how cultural and ethical presumptions bias artificially intelligent creations. Such introspection is more necessary than ever because the precise and deliberate design of algorithms in self-sustained digital systems will result in responses based on such design.

By drawing from over two thousand years' worth of classical ethics traditions, we explore established ethics systems, including both philosophical traditions (utilitarianism, virtue ethics, and deontological ethics) and religious and culture-based ethical systems (Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto) and their stance on human morality in the digital age.¹ In doing so, we critique assumptions around concepts such as good and evil, right and wrong, virtue and vice, and we attempt to carry these inquiries into artificial systems' decision-making processes.

Through reviewing the philosophical foundations that define autonomy and ontology, we address the potential for autonomous capacity of artificially intelligent systems, posing questions of morality in amoral systems and asking whether decisions made by amoral systems can have moral consequences. Ultimately, we address notions of responsibility and accountability for the decisions made by autonomous systems and other artificially intelligent technologies.

Classical Ethics in A/IS

Section 1—Definitions for Classical Ethics in Autonomous and Intelligent Systems Research

Issue: Assigning Foundations for Morality, Autonomy, and Intelligence

Background

Classical theories of economy in the Western tradition, starting with Plato and Aristotle, embrace three domains: the individual, the family, and the *polis*. The formation of the individual character (*ethos*) is intrinsically related to the others, as well as to the tasks of administration of work within the family (*oikos*). Eventually, this all expands into the framework of the *polis*, or public space (*poleis*). When we discuss ethical issues of A/IS, it becomes crucial to consider these three traditional economic dimensions, since western classical ethics was developed from this foundation and has evolved in modernity into an individual morality disconnected from economics and politics. This disconnection has been questioned and explored by thinkers such as Adam Smith, Georg W. F. Hegel, Karl Marx, and others. In particular,

Immanuel Kant's ethics located morality within the subject (see: [categorical imperative](#)) and separated morality from the outside world and the consequences of being a part of it. The moral autonomous subject of modernity became thus a worldless isolated subject. This process is important to understand in terms of ethics for A/IS since it is, paradoxically, the kind of autonomy that is supposed to be achieved by intelligent machines as humans evolve into digitally networked beings.

There lies a danger in uncritically attributing classical concepts of anthropomorphic autonomy to machines, including using the term "artificial intelligence" to describe them since, in the attempt to make them "moral" by programming moral rules into their behavior, we run the risk of assuming economic and political dimensions that do not exist, or that are not in line with contemporary human societies. While the concepts of artificial intelligence and autonomy are mainly used metaphorically as technical terms in computer science and technology, general and popular discourse may not share in the same nuanced understanding, and political and societal discourse may become distorted or

Classical Ethics in A/IS

misleading. The question of whether A/IS and the terminology used to describe them will have any kind of impact on our conception of autonomy depends on our policy toward it. For example, the commonly held fear that A/IS will relegate humanity to mere spectators or slaves, whether realistic or not, is informed by our view of, and terminology around, A/IS. Such attitudes are flexible and can be negotiated. As noted above, present human societies are being redefined in terms of digital citizenship via online social networks. The present public debate about the replaceability of human work by “intelligent” machines is a symptom of this lack of awareness of the economic and political dimensions as defined by classical ethics, reducing ethical thinking to the “morality” of a worldless and isolated machine.

There is still value that can be gained by considering how Western ethical traditions can be integrated into either A/IS public awareness campaigns or supplemented in engineering and science education programs, as noted under the issue “Presenting ethics to the creators of A/IS”. Below is a short overview of how four different traditions can add value.

- **Virtue ethics:** Aristotle argues, using the concept of *telos*, or goal, that the ultimate goal of humans is “*eudaimonia*”, roughly translated as “flourishing”. A moral agent achieves “flourishing”—since it is an action, not a state—by constantly balancing factors including social environment, material provisions, friends, family, and one's own self. One cultivates the self through habituation, practicing and strengthening virtuous action as the “golden mean” (a principle of rationality). Such cultivation requires an appropriate

balance between extremes of excess and deficiency, which Aristotle identifies as vices. In the context of A/IS, virtue ethics has two immediate values. First, it provides a model for iterative learning and growth, and moral value informed by context and practice, not just as compliance with a given, static ruleset. Second, it provides to those who develop and implement A/IS a framework to counterbalance tendencies toward excess, which are common in economically-driven environments.

- **Deontological ethics:** As developed by 18th century German philosopher, Immanuel Kant, the basic premise of deontological ethics addresses the concept of duty. Humans have a rational capacity to create and abide by rules that allow for duty-based ethics to emerge. Rules that produce duties are said to have value in themselves without requiring a greater-good justification. Such rules are fundamental to our existence, self-worth, and to creating conditions that allow for peaceful coexistence and interaction, e.g., the duty not to harm others; the duty not to steal. To identify rules that can be universalized and made duties, Kant uses the categorical imperative: “Act only on that maxim through which you can at the same time will that it should become a universal law.” This means the rule must be inherently desirable, doable, valuable, and others must be able to understand and follow it. Rules based merely on personal choice without wider appeal are not capable of universalization. There is mutual reciprocity in rule-making and rule adherence; if you “will” that a rule should become universal law, you not only contribute

Classical Ethics in A/IS

to rule creation but also agree to be bound by the same rule. The rule should be action-guiding, i.e., recommending, prescribing, limiting, or proscribing action. Kant also uses the humanity formulation of the categorical imperative: “Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end.” This produces duties to respect humanity and human dignity, and not to treat either as a means to an end.

- In the context of A/IS, one consideration is to wonder if developers are acting with the best interests of humanity and human dignity in mind. This could possibly be extended to A/IS whereby they are assisting humanity as an instrument of action that has an impact on decision-making capabilities, despite being based on neural machine learning or set protocols. The humanity formulation of “the categorical imperative” has implications for various scenarios. The duty to respect human dignity may require some limitations on the functions and capability of A/IS so that they do not completely replace humans, human functions, and/or “human central thinking activities” such as judgment, discretion, and reasoning. Privacy and safeguarding issues around A/IS assisting humans, e.g., healthcare robots, may require programming certain values so that A/IS do not divulge personal information to third parties, or compromise a human’s physical or mental well-being. It may also involve preventing A/IS from deceiving or manipulating humans.
- Potential benefits and financial incentives from exploiting A/IS may provide ends-means

justifications for their use, while disregarding the treatment of humanity as an end in itself, e.g., cutting back on funding rigorous testing of A/IS before they reach the market and society. Maintaining human agency in human-machine interaction is a manifestation of the duty to respect human dignity. For example, a human has the right to know when they are interacting with A/IS, and may require consent for any A/IS interaction.

- **Utilitarian ethics:** Also called consequentialist ethics, this code of ethics refers to the consequences of one’s decisions and actions. According to the utility principle, the right course of action is the one that maximizes the utility (utilitarianism) or pleasure (hedonism) for the greatest number of people. This ethics theory does, however, warn against superficial and short-term evaluations of utility or pleasure. Therefore, it is the responsibility of the A/IS developers to consider long-term effects. Social justice is paramount in this instance, thus it must be ascertained if the implementation of A/IS will contribute to humanity, or negatively impact employment or other capabilities. Indeed, where it is deemed A/IS can supplement humanity, it should be designed in such a way that the benefits are obvious to all the stakeholders.
- **Ethics of care:** Generally viewed as an instance of feminist ethics, this approach emphasizes the importance of relationships which is context-bound. Relationships are ontologically basic to humanity, according to Nel Noddings, feminist and philosopher of education; to care for other human beings is one of our basic human attributes. For such

Classical Ethics in A/IS

a theory to have relevance in this context, one needs to consider two criteria: 1) the relationship with the other person, or entity, must already exist or must have the potential to exist, and 2) the relationship should have the potential to grow into a caring relationship. Applied to A/IS, an interesting question comes to the foreground: Can one care for humans and their interests in tandem with non-human entities? If one expects A/IS to be beneficial to humanity, as in the instance of robots assisting with care of the elderly, then can one deduce the possibility of humans caring for A/IS? If that possibility exists, do principles of social justice become applicable to A/IS?

Recommendations

By returning to classical ethics foundations, expand the discussion on ethics in A/IS to include a critical assessment of anthropomorphic presumptions of ethics and moral rules for A/IS. Keep in mind that machines do not, in terms of classical autonomy, comprehend the moral or legal rules they follow. They move according to their programming, following rules that are designed by humans to be moral.

Expand the discussion on ethics for A/IS to include an exploration of the classical foundations of economy, outlined above, as potentially influencing current views and assumptions around machines achieving isolated autonomy.

Further Resources

- J. Bielby, Ed., "[Digital Global Citizenship](#)," International Review of Information Ethics, vol. 23, pp. 2-3, Nov. 2015.
- O. Bendel, "Towards Machine Ethics," in Technology Assessment and Policy Areas of Great Transitions: Proceedings from the PACITA 2013 Conference in Prague, PACITA 2013, Prague, March 13-15, 2013, T. Michalek, L. Hebáková, L. Hennen, C. Scherz, L. Nierling, J. Hahn, Eds. Prague: Technology Centre ASCR, 2014. pp. 321-326.
- O. Bendel, "[Considerations about the Relationship between Animal and Machine Ethics](#)," AI & Society, vol. 31, no. 1, pp. 103-108, Feb. 2016.
- N. Berberich and K. Diepold, "[The Virtuous Machine - Old Ethics for New Technology?](#)" arXiv:1806.10322 [cs.AI], June 2018.
- R. Capurro, M. Eldred, and D. Nagel, Digital Whoness: Identity, Privacy and Freedom in the Cyberworld. Berlin: Walter de Gruyter, 2013.
- D. Chalmers, "[The Singularity: A Philosophical Analysis](#)," Journal of Consciousness Studies, vol. 17, pp. 7-65, 2010.
- D. Davidson, "Representation and Interpretation," in Modelling the Mind, K. A. M. Said, W. H. Newton-Smith, R. Viale, and K. V. Wilkes, Eds. New York: Oxford University Press, 1990, pp. 13-26.
- N. Noddings, Caring: A Relational Approach to Ethics and Moral Education. Oakland, CA: University of California Press, 2013.
- O. Ulgen, "Kantian Ethics in the Age of Artificial Intelligence and Robotics," QIL, vol. 43, pp. 59-83, Oct. 2017.

Classical Ethics in A/IS

- O. Ulgen, “The Ethical Implications of Developing and Using Artificial Intelligence and Robotics in the Civilian and Military Spheres,” House of Lords Select Committee, Sept. 6, 2017, UK.
- O. Ulgen, “Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an ‘Elementary Consideration of Humanity’?” in *How International Law Works in Times of Crisis*, I. Ziemele and G. Ulrich, Eds. Oxford: Oxford University Press, 2018.

Issue: The Distinction between Agents and Patients

Background

Of particular concern when understanding the relationship between human beings and A/IS is the uncritically applied anthropomorphic approach toward A/IS that many industry and policymakers are using today. This approach erroneously blurs the distinction between moral agents and moral patients, i.e., subjects, otherwise understood as a distinction between “natural” self-organizing systems and artificial, non-self-organizing devices. As noted above, A/IS cannot, by definition, become autonomous in the sense that humans or living beings are autonomous. With that said, autonomy in machines, when critically defined, designates how machines act and operate independently in certain contexts through a consideration of implemented order generated by laws and rules. In this sense, A/IS can, by definition, qualify as

autonomous, especially in the case of genetic algorithms and evolutionary strategies. However, attempts to implant true morality and emotions, and thus accountability, i.e., autonomy, into A/IS blurs the distinction between agents and patients and may encourage anthropomorphic expectations of machines by human beings when designing and interacting with A/IS.

Thus, an adequate assessment of expectations and language used to describe the human-A/IS relationship becomes critical in the early stages of its development, where analyzing subtleties is necessary. Definitions of autonomy need to be clearly drawn, both in terms of A/IS and human autonomy. On one hand, A/IS may in some cases manifest seemingly ethical and moral decisions, resulting for all intents and purposes in efficient and agreeable moral outcomes. Many human traditions, on the other hand, can and have manifested as fundamentalism under the guise of morality. Such is the case with many religious moral foundations, where established cultural mores are neither questioned nor assessed. In such scenarios, one must consider whether there is any functional difference between the level of autonomy in A/IS and that of assumed agency—the ability to choose and act—in humans via the blind adherence to religious, traditional, or habitual mores. The relationship between assumed moral customs, the ethical critique of those customs, and the law are important distinctions.

The above misunderstanding in definitions of autonomy arises in part because of the tendency for humans to shape artificial creations in their own image, and our desire to lend our human

Classical Ethics in A/IS

experience to shaping a morphology of artificially intelligent systems. This is not to say that such terminology cannot be used metaphorically, but the difference must be maintained, especially as A/IS begin to resemble human beings more closely. It is possible for terms like “artificial intelligence” or “morality of machines” to be used as metaphors without resulting in misunderstanding. This is how language works and how humans try to understand their natural and artificial environment.

However, the critical difference between human autonomy and autonomous systems involves questions of free will, predetermination, and being (ontology). The questions of critical ontology currently being applied to machines are not new questions to ethical discourse and philosophy; they have been thoroughly applied to the nature of human *being* as well. John Stuart Mill, for example, is a determinist and claims that human actions are predicated on predetermined laws. He does, however, argue for a reconciliation of human free will with determinism through a theory of compatibility. Millian ethics provides a detailed and informed foundation for defining autonomy that could serve to help overcome general assumptions of anthropomorphism in A/IS and thereby address the uncertainty therein (Mill, 1999).

Recommendations

When addressing the nature of “autonomy” in autonomous systems, it is recommended that the discussion first consider free will, civil liberty, and society from a Millian perspective in order to better grasp definitions of autonomy and to address general assumptions of anthropomorphism in A/IS.

Further Resources

- R. Capurro, “[Toward a Comparative Theory of Agents](#).” *AI & Society*, vol. 27, no. 4, pp. 479-488, Nov. 2012.
- W. J. King and J. Ohya, “The Representation of Agents: Anthropomorphism, Agency, and Intelligence,” in *Conference Companion on Human Factors in Computing Systems*. Vancouver: ACM, 1996, pp. 289-290.
- W. Hofkirchner, “[Does Computing Embrace Self-Organisation?](#)” in *Information and Computation: Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation*, G. Dodig-Crnkovic and M. Burgin, Eds. London: World Scientific, 2011, pp. 185-202.
- [International Center for Information Ethics, 2018.](#)
- J. S. Mill, *On Liberty*. London: Longman, Roberts & Green, 1869.
- P. P. Verbeek, *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. University Park, PA: Pennsylvania State University Press, 2005.

Classical Ethics in A/IS

Issue: The Need for an Accessible, Classical Ethics Vocabulary

Background

Philosophers and ethicists are trained in vocabulary relating to philosophical concepts and terminology. There is an intrinsic value placed on these concepts when discussing ethics and A/IS, since the layered meaning behind the terminology used is foundational to these discussions and is grounded in a subsequent entrenchment of values. Unfortunately, using philosophical terminology in cross-disciplinary instances, i.e., a conversation between technologists and policymakers, is often ineffective since not everyone has the education to be able to encompass the abstracted layers of meaning contained in philosophical terminology.

However, not understanding a philosophical definition does not detract from the necessity of its utility. While ethical and philosophical theories should not be over-simplified for popular consumption, being able to adequately translate the essence of the rich history of ethics will go a long way in supporting a constructive dialogue on ethics and A/IS. With access and accessibility concerns intricately linked with education in communities, as well as secondary and tertiary institutions, society needs to take a vested interest in creating awareness for government officials, rural communities, and school teachers. Creating a more “user-friendly” vocabulary raises awareness on the necessity and application of classical ethics to digital societies.

Identifying terms that will be intelligible to all relevant audiences is pragmatic, but care should be taken not to dilute or misrepresent concepts that are familiar to moral philosophy and ethics. One way around this is to engage in applied ethics; illustrate how a particular concept would work in the A/IS context or scenario. Another way is to understand whether terminology used across different disciplines actually has the same or similar meaning and effect which can be expressed accordingly.

Recommendations

Support and encourage the efforts of groups raising awareness for social and ethics committees, whose roles are to support ethics dialogue within their organizations, seeking approaches that are both aspirational and values-based. A/IS technologists should engage in cross-disciplinary exchanges whereby philosophy scholars and ethicists attend and present in non-philosophical courses. This will both raise awareness and sensitize non-philosophical scholars and practitioners to the vocabulary.

Further Resources

- R. T. Ames, *Confucian Role Ethics: A Vocabulary*. Hong Kong: Chinese University Press, 2011.
- R. Capurro, “[Towards an Ontological Foundation of Information Ethics](#),” *Ethics and Information Technology*, vol. 8, no. 4, pp. 175-186, 2006.
- S. Mattingly-Jordan, R. Day, B. Donaldson, P. Gray, and L. M. Ingram, “[Ethically Aligned Design, First Edition Glossary](#),” Prepared for The IEEE Global Initiative for Ethically Aligned Design, Feb. 2019.

Classical Ethics in A/IS

- B. M. Lowe, *Emerging Moral Vocabularies: The Creation and Establishment of New Forms of Moral and Ethical Meanings*. Lanham, MD: Lexington Books, 2006.
- D. J. Flinders, "[In Search of Ethical Guidance: Constructing a Basis for Dialogue](#)," *International Journal of Qualitative Studies in Education*, vol. 5, no. 2, pp. 101-115, 1992.
- G. S. Saldanha, "[The Demon in the Gap of Language: Capurro, Ethics and Language in Divided Germany](#)," in *Information Cultures in the Digital Age*. Wiesbaden, Germany: Springer Fachmedien, 2016, pp. 253-268.
- J. Van Den Hoven and G. J. Lokhorst, "Deontic Logic and Computer Supported Computer Ethics," *Metaphilosophy*, vol. 33, no. 3, pp. 376-386, April 2002.

Issue: Presenting Ethics to the Creators of Autonomous and Intelligent Systems

Background

The question arises as to whether or not classical ethics theories can be used to produce meta-level orientations to data collection and data use in decision-making. Keeping in mind that the task of philosophical ethics should be to examine good and evil, ethics should examine values, not prescribe them. Laws, which arise from ethics, are entrenched mores that have been critically assessed to prescribe.

The key is to embed ethics into engineering in a way that does not make ethics a servant, but instead a partner in the process. In addition to an ethics-in-practice approach, providing students and engineers with the tools necessary to build a similar orientation into their inventions further entrenches ethical design practices. In the abstract, this is not so difficult to describe, but is very difficult to encode into systems. This problem can be addressed by providing students with job aids such as checklists, flowcharts, and matrices that will help them select and use a principal ethical framework, and then exercise use of those devices with steadily more complex examples. In such an iterative process, students will start to determine for themselves what examples do not allow for perfectly clear decisions, and, in fact, require some interaction between frameworks. Produced outcomes such as videos, essays, and other formats—such as project-based learning activities—allow for a didactic strategy which proves effective in artificial intelligence ethics education.

The goal is to provide students a means to use ethics in a manner analogous to how they are being taught to use engineering principles and tools. In other words, the goal is to help engineers tell the story of what they are doing.

- Ethicists should use information flows and consider at a meta-level what information flows do and what they are supposed to do.
- Engineers should then build a narrative that outlines the iterative process of ethical considerations in their design. Intentions are part of the narrative and provide a base to reflect back on those intentions.

Classical Ethics in A/IS

- The process then allows engineers to better understand their assumptions and adjust their intentions and design processes accordingly. They can only get to these by asking targeted questions.

This process, one with which engineers are quite familiar, is basically Kantian and Millian ethics in play.

The aim is to produce what is referred to in the computer programming lexicon as a *macro*. A macro is code that takes other code as its input(s) and produces unique outputs. This macro is built using the Western ethics tradition of virtue ethics.

This further underscores the importance of education and training on ethical considerations relating to A/IS. Such courses should be developed and presented to students of engineering, A/IS, computer science, and other relevant fields. These courses do not add value *a posteriori*, but should be embedded from the beginning to allow for absorption of the underlying ethical considerations as well as allowing for critical thinking to come to fruition once the students graduate. There are various approaches that can be considered on a tertiary level:

- Parallel (information) ethics program that is presented together with the science program during the course of undergraduate and postgraduate study;
- Embedded (information) ethics modules within the science program, i.e., one module per semester;
- Short (information) ethics courses specifically designed for the science program that can be attended by the current students, alumni, or professionals. These will function as either introductory, refresher, or specialized courses.

Courses can also be blended to include students and/or practitioners from diverse backgrounds rather than the more traditional practice of homogenous groups, such as engineering students, continuing education programs directed at a specific specialization, and the like.

Recommendations

Find ways to present ethics where the methodologies used are familiar to engineering students. As engineering is taught as a collection of techno-science, logic, and mathematics, embedding ethical sensitivity into these objective and non-objective processes is essential. Curricula development is crucial in each approach. In addition to research articles and best practices, it is recommended that engineers and practitioners come together with social scientists and philosophers to develop case studies, interactive virtual reality gaming, and additional course interventions that are relevant to students.

Further Resources

- T. W. Bynum and S. Rogerson, *Computer Ethics and Professional Responsibility*. Malden, MA: Wiley-Blackwell, 2003.
- E. G. Seebauer and R. L. Barry, *Fundamentals of Ethics for Scientists and Engineers*. New York: Oxford University Press, 2001.

Classical Ethics in A/IS

- C. Whitbeck, "[Teaching Ethics to Scientists and Engineers: Moral Agents and Moral Problems](#)," Science and Engineering Ethics, vol. 1, no. 3, pp. 299-308, Sept. 1995.
- B. Zevenbergen, et al. "[Philosophy Meets Internet Engineering: Ethics in Networked Systems Research](#)," GTC Workshop Outcomes Paper. Oxford: Oxford Internet Institute, University of Oxford, 2015.
- M. Alvarez, "[Teaching Information Ethics](#)," International Review of Information Ethics, vol. 14, pp. 23-28, Dec. 2010.
- P. P. Verbeek, [Moralizing Technology: Understanding and Designing the Morality of Things](#). Chicago, IL: University of Chicago Press, 2011.
- K. A. Joyce, K. Darfler, D. George, J. Ludwig, and K. Unsworth, "[Engaging STEM Ethics Education](#)," Engaging Science, Technology, and Society, vol. 4, no. 1-7, 2018.

Issue: Accessing Classical Ethics by Corporations and Companies

Background

Many companies, from startups to tech giants, understand that ethical considerations in tech design are increasingly important, but are not sure how to incorporate ethics into their tech design agenda. How can ethical considerations in tech design become an integrated part of the agenda of companies, public projects, and research consortia? Corporate workshops and exercises will need to go beyond

opinion-gathering exercises to embed ethical considerations into structures, environments, training, and development.

As it stands, classical ethics is not accessible enough to corporate endeavors in ethics, and as such, are not applicable to tech projects. There is often, but not always, a big discrepancy between the output of engineers, lawyers, and philosophers when dealing with computer science issues; there is also a large difference in how various disciplines approach these issues. While this is not true in all cases—and there are now several interdisciplinary approaches in robotics and machine ethics as well as a growing number of scientists that hold double and interdisciplinary degrees—there remains a vacuum for the wider understanding of classical ethics theories in the interdisciplinary setting. Such an understanding includes that of the philosophical language used in ethics and the translation of that language across disciplines.

If we take, for instance, the terminology and usage of the concept of "trust" in reference to technology, the term "trust" has specific philosophical, legal, and engineering connotations. It is not an abstract concept. It is attributable to humans, and relates to claims and actions people make. Machines, robots, and algorithms lack the ability to make claims and so cannot be attributed with trust. They cannot determine whether something is trustworthy or not. Software engineers may refer to "trusting" the data, but this relates to the data's authenticity and veracity to ensure software performance. In the context of A/IS, "trust" means "functional reliability"; it means there is confidence in the technology's predictability, reliability, and security against hackers or impersonators of authentic users.

Classical Ethics in A/IS

Recommendations

In order to achieve multicultural, multidisciplinary, and multi-sectoral dialogues between technologists, philosophers, and policymakers, a nuanced understanding in philosophical and technical language, which is critical to digital society from Internet of Things (IoT), privacy, and cybersecurity to issues of Internet governance, must be translated into norms and made available to technicians and policymakers who may not understand the nuances of the terminology in philosophical, legal, and engineering contexts. It is therefore recommended that the translation of the critical-thinking terminology of philosophers, policymakers, and other stakeholders on A/IS be translated into norms accessible to technicians.

Further Resources

- A. Bhimani, "[Making Corporate Governance Count: The Fusion of Ethics and Economic Rationality](#)," Journal of Management & Governance, vol. 12, no. 2, pp. 135-147, June 2008.
- A. B. Carroll, "A History of Corporate Social Responsibility," in The Oxford Handbook of Corporate Social Responsibility, A. Chisanthi, R. Mansell, D. Quah, and R. Silverstone, Eds. Oxford, U.K.: Oxford University Press, 2008.
- W. Lazonick, "Globalization of the ICT Labor Force," in The Oxford Handbook of Information and Communication Technologies, A. Chisanthi, R. Mansell, D. Quah, and R. Silverstone, Eds. Oxford, U.K.: Oxford University Press, 2006.

- IEEE P7000™, [IEEE Standards Project for Model Process for Addressing Ethical Concerns During System Design](#) will provide engineers and technologists with an implementable process aligning innovation management processes, IT system design approaches, and software engineering methods to minimize ethical risk for their organizations, stakeholders and end users.

Issue: The Impact of Automated Systems on the Workplace

Background

The impact of A/IS on the workplace and the changing power relationships between workers and employers requires ethical guidance. Issues of data protection and privacy via big data in combination with the use of autonomous systems by employers are increasing, where decisions made via aggregate algorithms directly impact employment prospects. The uncritical use of A/IS in the workplace, and its impact on employee-employer relations, is of utmost concern due to the high chance of error and biased outcome.

The concept of responsible research and innovation (RRI) is a growing area, particularly within the EU. It offers potential solutions to workplace bias and is being adopted by several research funders, such as the Engineering and Physical Sciences Research Council (EPSRC), who include RRI core principles in their mission statement. RRI is an umbrella concept that draws on classical ethics theory to provide tools to address ethical concerns from the outset of a project, from the design stage onwards.

Classical Ethics in A/IS

Quoting Rene Von Schomberg, science and technologies studies specialist and philosopher, “Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, **sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society).**”²

When RRI methodologies are used in the ethical considerations of A/IS design, especially in response to the potential bias of A/IS in the workplace, theoretical deficiencies are then often exposed that would not otherwise have been exposed, allowing room for improvement in design at the development stage rather than from a retroactive perspective. RRI in design increases the chances of both relevance and strength in ethically aligned design.

This emerging and exciting new concept aims to also push the boundaries to incorporate relevant stakeholders whose influence in responsible research is on a global stage. While this concept initially focuses on the workplace setting, success will only be achieved through the active involvement from private companies of industry, AI Institutes, and those who are at the forefront in A/IS design. Responsible research and innovation will be achieved through careful research and innovation governance that will ensure research purpose, process, and outcomes that are acceptable, sustainable, and even desirable. It will be incumbent on RRI experts to engage at a level where private companies will feel empowered

and embrace this concept as both practical to implement and enact.

Recommendations

It is recommended, through the application of RRI as founded in classical ethics theory, that research in A/IS design utilize available tools and approaches to better understand the design process, addressing ethical concerns from the very beginning of the design stage of the project, thus maintaining a stronger, more efficient methodological accountability throughout.

Further Resources

- M. Burget, E. Bardone, and M. Pedaste, “Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review,” *Science and Engineering Ethics*, vol. 23, no. 1, pp. 1-9, 2016.
- European Commission Communication, “[Artificial Intelligence for Europe](#),” COM 237, April, 2018.
- R. Von Schomberg, “Prospects for Technology Assessment in a Framework of Responsible Research and Innovation,” in *Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode*. Wiesbaden, Germany: Springer VS, 2011, pp. 39-61.
- B. C. Stahl, G. Eden, M. Jirotko, M. Coeckelbergh, “From Computer Ethics to Responsible Research and Innovation in ICT: The Transition of Reference Discourses Informing Ethics-Related Research in Information Systems,” *Information & Management*, vol. 51, no. 6, pp. 810-818, September 2014.

Classical Ethics in A/IS

- B. C. Stahl, M. Obach, E. Yaghmaei, V. Ikonen, K. Chatfield, and A. Brem, "[The Responsible Research and Innovation \(RRI\) Maturity Model: Linking Theory and Practice](#)," Sustainability, vol. 9, no. 6, June 2017.
- IEEE P7005™, [Standards Project for Transparent Employer Data Governance](#) is

designed to provide organizations with a set of clear guidelines and certifications guaranteeing they are storing, protecting, and utilizing employee data in an ethical and transparent way.

Section 2—Classical Ethics from Globally Diverse Traditions

Issue: The Monopoly on Ethics by Western Ethical Traditions

Background

As human creators, our most fundamental values are imposed on the systems we design. It becomes incumbent on the global community to recognize which sets of values guide the design, and whether or not A/IS will generate problematic, i.e., discriminatory, consequences without consideration of non-Western values. There is an urgent need to broaden traditional ethics in its contemporary form of "responsible innovation" (RI) beyond the scope of "Western" ethical foundations, such as utilitarianism, deontology, and virtue ethics. There is also a need to include other traditions of ethics in RI, such as those inherent to Buddhism, Confucianism, and Ubuntu traditions.

However, this venture poses problematic assumptions even before the issue above can be explored. In classifying Western values, we group together thousands of years of independent and disparate ideas originating from the Greco-Roman philosophical tradition with their Christian-infused cultural heritage and then the break from that heritage with the Enlightenment. What is it that one refers to by the term "Western ethics"? Does one refer to philosophical ethics (ethics as a scientific discipline) or is the reference to Western morality?

The "West", however it may be defined, is an individualistic society, arguably more so than much of the rest of the world, and thus, in some aspects, should be even less collectively defined than "Eastern" ethical traditions. Suggest instead: If one is referring to Western values, one must designate which values and to whom they belong. Additionally, there is a danger in the field of intercultural information ethics, however

Classical Ethics in A/IS

unconsciously or instinctively propagated, to not only group together all Western traditions under a single banner, but to negatively designate any and all Western influence in global exchange to representing an abusive collective of colonial-influenced ideals. Just because there exists a monopoly of influence by one system over another does not mean that said monopoly is devoid of value, even for systems outside itself. In the same way that culturally diverse traditions have much to offer Western tradition(s), so, too, do they have much to gain from them.

In order to establish mutually beneficial connections in addressing globally diverse traditions, it is of critical importance to first properly distinguish between subtleties in Western ethics as a discipline and morality as its object or subject matter. It is also important to differentiate between philosophical or scientific ethics and theological ethics. As noted above, the relationship between assumed moral customs, the ethical critique of those customs, and the law is an established methodology in scientific communities. Western and Eastern philosophy are very different, just like Western and Eastern ethics. Western philosophical ethics use scientific methods such as the logical, discursive, and dialectical approach (models of normative ethics) alongside the analytical and hermeneutical approaches. The Western tradition is not about education and teaching of social and moral values, but rather about the application of fundamentals, frameworks, and explanations. However, several contemporary globally relevant community mores are based in traditional and theological moral systems, requiring a conversation around how best to collaborate in

the design and programming of ethics in A/IS amidst differing ethical traditions.

While experts in Intercultural Information Ethics, such as Pak-Hang Wong, highlight the dangers of the dominance of “Western” ethics in A/IS design, noting specifically the appropriation of ethics by liberal democratic values to the exclusion of other value systems, it should be noted that those same liberal democratic values are put in place and specifically designed to accommodate such differences. However, while the accommodation of differences are, in theory, accounted for in dominant liberal value systems, the reality of the situation reveals a monopoly of, and a bias toward, established Western ethical value systems, especially when it comes to standardization. As Wong notes:

Standardization is an inherently value-laden project, as it designates the normative criteria for inclusion to the global network. Here, one of the major adverse implications of the introduction of value-laden standard(s) of responsible innovation (RI) appears to be the delegitimization of the plausibility of RI based on local values, especially when those values come into conflict with the liberal democratic values, as the local values (or, the RI based on local values) do not enable scientists and technology developers to be recognized as members of the global network of research and innovation (Wong, 2016).

It does, however, become necessary for those who do not work within the parameters of accepted value monopolies to find alternative methods of accommodating different value systems. Liberal values arose out of conflicts

Classical Ethics in A/IS

of cultural and subcultural differences and are designed to be accommodating enough to include a rather wide range of differences.

RI enables policymakers, scientists, technology developers, and the public to better understand and respond to the social, ethical, and policy challenges raised by new and emerging technologies. Given the historical context from which RI emerges, it should not be surprising that the current discourse on RI is predominantly based on liberal democratic values. Yet, the bias toward liberal democratic values will inevitably limit the discussion of RI, especially in the cases where liberal democratic values are not taken for granted. Against this background, it is important to recognize the problematic consequences of RI solely grounded on, or justified by, liberal democratic values.

In addition, many non-Western ethics traditions, including the Buddhist and Ubuntu traditions highlighted below, view “relationship” as a foundationally important concept to ethical discourse. One of the key parameters of intercultural information ethics and RI research must be to identify main commonalities of “relationship” approaches from different cultures and how to operationalize them for A/IS to complement classical methodologies of deontological and teleological ethics. Different cultural perceptions of time may influence “relationship” approaches and impact how A/IS are perceived and integrated, e.g., technology as part of linear progress in the West; inter-generational needs and principles of respect and benevolence in Chinese culture determining current and future use of technology.

Recommendations

In order to enable a cross-cultural dialogue of ethics in technology, discussions on ethics and A/IS must first return to normative foundations of RI to address the notion of “responsible innovation” from a range of value systems not predominant in Western classical ethics. Together with acknowledging differences, a special focus on commonalities in the intercultural understanding of the concept of “relationship” must complement the process.

Further Resources

- J. Bielby, “Comparative Philosophies in Intercultural Information Ethics,” *Confluence: Journal of World Philosophies*, vol. 2, 2016.
- W. B. Carlin and K. C. Strong, “A Critique of Western Philosophical Ethics: Multidisciplinary Alternatives for Framing Ethical Dilemmas,” *Journal of Business Ethics*, vol. 14, no. 5, pp. 387-396, May 1995.
- C. Ess, “[“Lost in translation”?: Intercultural dialogues on privacy and information ethics \(introduction to special issue on privacy and data privacy protection in Asia\)](#),” *Ethics and Information Technology*, vol. 7, no. 1, pp. 1-6, March 2005.
- S. Hongladarom, “[Intercultural Information Ethics: A Pragmatic Consideration](#),” in *Information Cultures in the Digital Age*. Wiesbaden, Germany: Springer Fachmedien, 2016, pp. 191-206.
- L. G. Rodríguez and M. Á. P. Álvarez, *Ética Multicultural y Sociedad en Red*. Madrid: Fundación Telefónica, 2014.

Classical Ethics in A/IS

- P. H. Wong, "[What Should We Share?: Understanding the Aim of Intercultural Information Ethics](#)," ACM SIGCAS Computers and Society, vol. 39, no. 3 pp. 50-58, Dec. 2009.
- S. A. Wilson, "[Conformity, Individuality, and the Nature of Virtue: A Classical Confucian Contribution to Contemporary Ethical Reflection](#)," The Journal of Religious Ethics, vol. 23, no. 2, pp. 263-289, 1995.
- P. H. Wong, "[Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?](#)" Journal of Responsible Innovation, vol. 3, no. 2, pp. 154-168, July 2016.
- R. B. Zeuschner, Classical Ethics, East and West: Ethics from a Comparative Perspective. Boston, MA: McGraw-Hill, 2000.
- S. Mattingly-Jordan, "[Becoming a Leader in Global Ethics](#)," IEEE, 2017.

Issue: The Application of Classical Buddhist Ethical Traditions to A/IS Design

Background

According to Buddhism, the field of ethics is concerned with behaving in such a way that the subject ultimately realizes the goal of liberation. The question, "How should I act?" is answered straightforwardly; one should act in such a way that one realizes liberation (nirvana)

in the future, achieving what in Buddhism is understood as "supreme happiness". Thus Buddhist ethics are clearly goal-oriented. In the Buddhist tradition, people attain liberation when they no longer endure any unsatisfactory conditions, when they have attained the state where they are completely free from any passions, including desire, anger, and delusion—to name the traditional three, that ensnare one's self against freedom. In order to attain liberation, one engages oneself in mindful behavior (ethics), concentration (meditation), and what is deemed in Buddhism as "wisdom", a term that remains ambiguous in Western scientific approaches to ethics.

Thus ethics in Buddhism are concerned exclusively with how to attain the goal of liberation, or freedom. In contrast to Western ethics, Buddhist ethics are not concerned with theoretical questions on the source of normativity or what constitutes the good life. What makes an action a "good" action in Buddhism is always concerned with whether the action leads, eventually, to liberation or not. In Buddhism, there is no questioning why liberation is a good thing. It is simply assumed. Such an assumption places Buddhism, and ethical reflection from a Buddhist perspective, in the camp of mores rather than scientifically led ethical discourse, and it is approached as an ideology or a worldview.

While it is critically important to consider, understand, and apply accepted ideologies such as Buddhism in A/IS, it is both necessary to differentiate the methodology from Western ethics, and respectful to Buddhist tradition, not to require that it be considered in a scientific

Classical Ethics in A/IS

context. Such assumptions put it at odds with the Western foundation of ethical reflection on mores. From a Buddhist perspective, one does not ask why supreme happiness is a good thing; one simply accepts it. The relevant question in Buddhism is not about methodological reflection, but about how to attain liberation from the necessity for such reflection.

Thus, Buddhist ethics contain potential for conflict with Western ethical value systems which are founded on ideas of questioning moral and epistemological assumptions. Buddhist ethics are different from, for example, utilitarianism, which operates via critical analysis toward providing the best possible situation to the largest number of people, especially as it pertains to the good life. These fundamental differences between the traditions need to be, first and foremost, mutually understood and then addressed in one form or another when designing A/IS that span cultural contexts.

The main difference between Buddhist and Western ethics is that Buddhism is based upon a metaphysics of relation. Buddhist ethics emphasizes how *action* leads to achieving a *goal*, or in the case of Buddhism, the final goal. In other words, an action is considered a good one when it contributes to the realization of the goal. It is relational when the value of an action is relative to whether or not it leads to the goal, the goal being the reduction and eventual cessation of suffering. In Buddhism, the self is constituted through the relationship between the synergy of bodily parts and mental activities. In Buddhist analysis, the self does not actually exist as a self-subsisting entity. Liberation, or nirvana, consists in realizing that what is known to be the

self actually consists of nothing more than these connecting episodes and parts. To exemplify the above, one can draw from the concept of privacy as often explored via intercultural information ethics. The Buddhist perspective understands privacy as a protection, not of self-subsisting individuals, because such do not exist ultimately speaking, but of certain values that are found to be necessary for a well-functioning society to prosper in the globalized world.

The secular formulation of the supreme happiness mentioned above is that of the reduction of the experience of suffering, or reduction of the metacognitive state of suffering. Such a state is the result of lifelong discipline and meditation aimed at achieving proper relationships with others and with the world. This notion of the reduction of suffering is something that can resonate well with certain Western traditions, such as epicureanism ataraxia, i.e., freedom from fear through reason and discipline, and versions of consequentialist ethics that are more focused on the reduction of harm. It also encompasses the concept of phronesis or practical wisdom from virtue ethics.

Relational ethical boundaries promote ethical guidance that focuses on creativity and growth rather than solely on mitigation of consequence and avoidance of error. If the goal of the reduction of suffering can be formulated in a way that is not absolute, but collaboratively defined, this leaves room for many philosophies and related approaches as to how this goal can be accomplished. Intentionally making space for ethical pluralism is one potential antidote to dominance of the conversation by liberal thought, with its legacy of Western colonialism.

Classical Ethics in A/IS

Recommendations

In considering the nature of interactions between human and autonomous systems, the above notion of “proper relationships” through Buddhist ethics can provide a useful platform that results in ethical statements formulated in a relational way, instead of an absolutist way. It is recommended as an additional methodology, along with Western-value methodologies, to address human/computer interactions.

Further Resources

- R. Capurro, [“Intercultural Information Ethics: Foundations and Applications,”](#) Journal of Information, Communication & Ethics in Society, vol. 6, no. 2, pp. 116-126, 2008.
- C. Ess, [“Ethical Pluralism and Global Information Ethics,”](#) Ethics and Information Technology, vol. 8, no. 4, pp. 215-226, Nov. 2006.
- S. Hongladarom, [“Intercultural Information Ethics: A Pragmatic Consideration,”](#) in Information Cultures in the Digital Age, K. M. Bielby, Ed. Wiesbaden, Germany: Springer Fachmedien Wiesbaden, 2016, pp. 191-206.
- S. Hongladarom, J. Britz, [“Intercultural Information Ethics,”](#) International Review of Information Ethics, vol. 13, pp. 2-5, Oct. 2010.
- M. Nakada, “Different Discussions on Roboethics and Information Ethics Based on Different Contexts (Ba). Discussions on Robots, Informatics and Life in the Information Era in Japanese Bulletin Board Forums and Mass Media,” Proceedings

Cultural Attitudes towards Communication and Technology, pp. 300-314, 2010.

- M. Mori, The Buddha in the Robot. Suginami-ku, Japan: Kosei Publishing, 1989.

Issue: The Application of Ubuntu Ethical Traditions to A/IS Design

Background

In his article, “African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu”, Thaddeus Metz frames the following question: “What does a sub-Saharan ethic focused on the good of community, interpreted philosophically as a moral theory, entail for the duties of various agents with respect to the news/opinion media?” (Metz, 2015, 1). In applying that question to A/IS, it reads: “If an ethic focused on the good of community, interpreted philosophically as a moral theory, is applied to A/IS, what would the implications be on the duties of various agents?” Agents, in this regard, would therefore be the following:

- Members of the A/IS research community
- A/IS programmers/computer scientists
- A/IS end-users
- A/IS themselves

Classical Ethics in A/IS

Ubuntu is a sub-Saharan philosophical tradition. Its basic tenet is that a person is a person through other persons. It develops further in the notions of caring and sharing as well as identity and belonging, whereby people experience their lives as bound up with their community. A person is defined in relation to the community since the sense of being is intricately linked with belonging. Therefore, community exists through shared experiences and values. It is a commonly held maxim in the Ubuntu tradition that, “to be is to belong to a community and participate.” As the saying goes, *motho ke motho ka batho babang*, or, “a person is a person because of other people.”

Very little research, if any at all, has been conducted in light of Ubuntu ethics and A/IS, but its focus will be within the following moral domains:

1. Among the members of the A/IS research community
2. Between the A/IS community/programmers/ computer scientists and the end-users
3. Between the A/IS community/programmers/ computer scientists and A/IS
4. Between the end-users and A/IS
5. Between A/IS and A/IS

Considering a future where A/IS will become more entrenched in our everyday lives, one must keep in mind that an attitude of sharing one's experiences with others and caring for their well-being will be impacted. Also, by trying to ensure solidarity within one's community, one

must identify factors and devices that will form part of their lifeworld. If so, will the presence of A/IS inhibit the process of partaking in a community, or does it create more opportunities for doing so? One cannot classify A/IS as only a negative or disruptive force; it is here to stay and its presence will only increase. Ubuntu ethics must come to grips with, and contribute to, the body of knowledge by establishing a platform for mutual discussion and understanding. Ubuntu, as collective human dignity, may offer a way of understanding the impact of A/IS on humankind, e.g., the need for human moral and legal agency; human life and death decisions to be taken by humans rather than A/IS.

Such analysis fleshes out the following suggestive comments of Desmond Tutu, renowned former chair of South Africa's Truth and Reconciliation Commission, when he says of Africans, “(We say) a person is a person through other people... I am human because I belong” (Tutu, 1999). As Tutu notes, “Harmony, friendliness, and community are great goods. Social harmony is for us the *summum bonum*—the greatest good. Anything that subverts or undermines this sought-after good is to be avoided” (2015:78).

In considering the above, it is fair to state that community remains central to Ubuntu. In situating A/IS within this moral domain, they will have to adhere to the principles of community, identity, and solidarity with others. On the other hand, they will also need to be cognizant of, and sensitive toward, the potential for community-based ethics to exclude individuals on the basis that they do not belong or fail to meet communitarian standards. For example,

Classical Ethics in A/IS

would this mean the excluded individual lacks personhood and as a consequence would not be able to benefit from community-based A/IS initiatives? How would community-based A/IS programming avoid such biases against individuals?

While virtue ethics question the goal or purpose of A/IS and deontological ethics question the duties, the fundamental question asked by Ubuntu would be, "How does A/IS affect the community in which it is situated?" This question links with the initial question concerning the duties of the various moral agents within the specific community. Motivation becomes very important, because if A/IS seek to detract from community, they will be detrimental to the identity of this community when it comes to job losses, poverty, lacks in education, and lacks in skills training. However, should A/IS seek to supplement the community by means of ease of access, support systems, and more, then it cannot be argued that they will be detrimental. In between these two motivators is a safeguarding issue about how to avoid excluding individuals from accessing community-based A/IS initiatives. It therefore becomes imperative that whoever designs the systems must work closely both with ethicists and the target community, audience, or end-user to ascertain whether their needs are identified and met.

Recommendations

It is recommended that a concerted effort be made toward the study and publication of literature addressing potential relationships between Ubuntu and other instances of African ethical traditions and A/IS value design. A/IS

designers and programmers must work closely with the end-users and target communities to ensure their design objectives, products, and services are aligned with the needs of the end-users and target communities.

Further Resources

- D. W. Lutz, "[African Ubuntu Philosophy and Global Management](#)," *Journal of Business Ethics*, vol. 84, pp. 313-328, Oct. 2009.
- T. Metz, "[African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu](#)," *Journal of Media Ethics: Exploring Questions of Media Morality*, vol. 30 no. 2, pp. 74-90, April 2015.
- T. Metz, "[Ubuntu as a moral theory and human rights in South Africa](#)," *African Human Rights Law Journal*, vol. 11, no. 2, pp. 532-559, 2011.
- R. Nicolson, *Persons in Community: African Ethics in a Global Culture*. Scottsville, South Africa: University of KwaZulu-Natal Press, 2008.
- A. Shutte, *Ubuntu: An Ethic for a New South Africa*. Dorpspruit, South Africa: Cluster Publications, 2001.
- D. Tutu, *No Future without Forgiveness*. London: Rider, 1999.
- O. Ulgen, "Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?" in *How International Law Works in Times of Crisis*, I. Ziemele and G. Ulrich, Eds. Oxford: Oxford University Press, 2018, pp. 242-272.

Classical Ethics in A/IS

Issue: The Application of Shinto-Influenced Traditions to A/IS Design

Background

Alongside the burgeoning African Ubuntu reflections on A/IS, other indigenous techno-ethical reflections boast an extensive engagement. One such tradition is Japanese Shinto indigenous spirituality, or, *Kami no michi*, often cited as the catalyst for Japanese robot and autonomous systems culture, a culture that naturally stems from the traditional Japanese concept of *karakuri ningyo* (automata). Popular Japanese artificial intelligence, robot, and video-gaming culture can be directly connected to indigenous Shinto tradition, from the existence of *kami* (spirits) to puppets and automata.

The relationship between A/IS and a human being is a personal relationship in Japanese culture and, one could argue, a very natural one. The phenomenon of “relationship” in Japan between humans and automata stands out as unique to technological relationships in world cultures, since the Shinto tradition is arguably the only animistic and naturalistic tradition that can be directly connected to contemporary digital culture and A/IS. From the Shinto perspective, the existence of A/IS, whether manifested through robots or other technological autonomous systems, is as natural to the world as rivers, forests, and thunderstorms. As noted by Spyros G. Tzafestas, author of *Roboethics: A Navigating Overview*, “Japan’s harmonious feeling

for intelligent machines and robots, particularly for humanoid ones,” (Tzafestas, 2015, 155) colors and influences technological development in Japan, especially robot culture.

The word “Shinto” can be traced to two Japanese concepts: *Shin*, meaning spirit, and *to*, the philosophical path. Along with the modern concept of the android, which can be traced back to three sources—the first, to its Greek etymology that combines andras (“άνδρας”), or man, and gynoids/gyni (“γυνή”), or woman; the second, via automatons and toys as per U.S. patent developers in the 1800s; and the third to Japan, where both historical and technological foundations for android development have dominated the market since the 1970s—Japanese Shinto-influenced technology culture is perhaps the most authentic representation of the human-automaton interface.

Shinto tradition is an animistic religious tradition, positing that everything is created with, and maintains, its own spirit (*kami*) and is animated by that spirit—an idea that goes a long way to defining autonomy in robots from a Japanese viewpoint. This includes, on one hand, everything that Western culture might deem natural, including rivers, trees, and rocks, and on the other hand, everything artificially (read: *artfully*) created, including vehicles, homes, and automata (robots). Artifacts are as much a part of nature in Shinto as animals, and they are considered naturally beautiful rather than falsely artificial.

A potential conflict between Western and Japanese concepts of nature and artifact arises when the two traditions are compared

Classical Ethics in A/IS

and contrasted, especially in the exploration of artificial intelligence. While in Shinto, the artifact as “artificial” represents creation and authentic being, with implications for defining autonomy, the same artifact is designated as secondary and often times unnatural, false, and counterfeit in Western ethical philosophical tradition, dating back to Platonic and Christian ideas of separation of form and spirit. In both traditions, culturally presumed biases define our relationships with technology. While disparate in origin and foundation, both Western classical ethics traditions and Shinto ethical influences in modern A/IS have similar goals and outlooks for ethics in A/IS, goals that are centered in “relationship”.

Recommendations

Where Japanese culture leads the way in the synthesis of traditional value systems and technology, we recommend that people involved with efforts in A/IS ethics explore the Shinto paradigm as representative, though not necessarily as directly applicable, to global efforts in understanding and applying traditional and classical ethics methodologies to A/IS.

Further Resources

- R. M. Geraci, "Spiritual Robots: Religion and Our Scientific View of the Natural World," *Theology and Science*, vol. 4, no. 3, pp. 229-246, 2006.
- D. F. Holland-Minkley, "[God in the Machine: Perceptions and Portrayals of Mechanical Kami in Japanese Anime.](#)" Ph.D. dissertation, University of Pittsburgh, Pittsburgh, PA, 2010.
- C. B. Jensen and A. Blok, "[Techno-Animism in Japan: Shinto Cosmograms, Actor-Network Theory, and the Enabling Powers of Non-Human Agencies,](#)" *Theory, Culture & Society*, vol. 30, no. 2, pp. 84-115, March 2013.
- F. Kaplan, "[Who Is Afraid of the Humanoid? Investigating Cultural Differences in the Acceptance of Robots,](#)" *International Journal of Humanoid Robotics*, vol. 1, no. 3, pp. 465-480, 2004.
- S. G. Tzafestas, *Roboethics: A Navigating Overview*. Cham, Switzerland: Springer, 2015.
- G. Veruggio and K. Abney, "22 Roboethics: The Applied Ethics for a New Science," in *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press, 2011, p. 347.

Classical Ethics in A/IS

Section 3—Classical Ethics for a Technical World

Issue: Maintaining Human Autonomy

Background

A/IS present the possibility for a digitally networked intellectual capacity that imitates, matches, and supersedes human intellectual capacity, including, among other things, general skills, discovery, and computing functions. In addition, A/IS can potentially acquire functionality in areas traditionally captured under the rubric of what we deem unique human and social ability. While the larger question of ethics and A/IS looks at the implications of the influence of autonomous systems in these areas, the pertinent issue is the possibility of autonomous systems imitating, influencing, and then determining the norms of human autonomy. This is done through the eventual negation of independent human thinking and decision-making, where algorithms begin to inform through targeted feedback loops what it is we *are* and what it is we should decide. Thus, how can the academic rigor of traditional ethics speak to the question of maintaining human autonomy in light of algorithmic decision-making?

How will A/IS influence human autonomy in ways that may or may not be advantageous to the good life, and perhaps—even if advantageous—may be detrimental at the same time? How do these systems affect human autonomy and decision-making through the use of algorithms when said algorithms tend to inform (“in-form”) via targeted feedback loops?

Consider, for example, Google’s autocomplete tool, where algorithms attempt to determine one’s search parameters via the user’s initial keyword input, offering suggestions based on several criteria including search patterns. In this scenario, autocomplete suggestions influence, in real-time, the parameters the user phrases their search by, often reforming the user’s perceived notions of what it was they were looking for in the first place, versus what they might have actually originally intended.

Targeted algorithms also inform, as per emerging IoT, applications that monitor the user’s routines and habits in the analog world. Consider for example that our bioinformation is, or soon will be, available for interpretation by autonomous systems. What happens when autonomous systems can inform the user in ways the user is not even aware of, using one’s bioinformation in targeted advertising campaigns that seek to influence the user in real-time feedback loops based on the user’s biological reactions such as

Classical Ethics in A/IS

pupil dilation, body temperature, and emotional reaction, whether positive or negative, to that very same advertising, using information about our being to in-form and re-form our being? On the other hand, it becomes important not to adopt dystopian assumptions concerning autonomous machines threatening human autonomy.

The tendency to think only in negative terms presupposes a case for interactions between autonomous machines and human beings, a presumption not necessarily based in evidence. Ultimately, the behavior of algorithms rests solely in their design, and that design rests solely in the hands of those who designed them. Perhaps more importantly, however, is the matter of choice in terms of how the user chooses to interact with the algorithm. Users often don't know when an algorithm is interacting with them directly or their data which acts as a proxy for their identity. Should there be a precedent for the A/IS user to know when they are interacting with an algorithm? What about consent?

The responsibility for the behavior of algorithms remains with the designer, the user, and a set of well-designed guidelines that guarantee the importance of human autonomy in any interaction. As machine functions become more autonomous and begin to operate in a wider range of situations, any notion of those machines working for or against human beings becomes contested. Does the machine work *for* someone in particular, or for particular groups but not others? Who decides on the parameters? Is it the machine itself? Such questions become key factors in conversations around ethical standards.

Recommendations

A two-step process is recommended to maintain human autonomy in A/IS. The creation of an ethics-by-design methodology is the first step to addressing human autonomy in A/IS, where a critically applied ethical design of autonomous systems preemptively considers how and where autonomous systems may or may not dissolve human autonomy. The second step is the creation of a pointed and widely applied education curriculum that spans grade school through university, one based on a classical ethics foundation that focuses on providing choice and accountability toward digital being as a priority in information and knowledge societies.

Further Resources

- B. van den Berg and J. de Mul, "Remote Control. Human Autonomy in the Age of Computer-Mediated Agency," in *Law, Human Agency and Autonomic Computing: The Philosophy of Law Meets the Philosophy of Technology*, M. Hildebrandt and A. Rouvroy, Eds. London: Routledge, 2011, pp. 46-63.
- L. Costa, "[A World of Ambient Intelligence](#)," in *Virtuality and Capabilities in a World of Ambient Intelligence*. Cham, Switzerland: Springer International, 2016, pp. 15-41.
- P. P. Verbeek, "[Subject to Technology on Autonomic Computing and Human Autonomy](#)," in *The Philosophy of Law Meets the Philosophy of Technology: Autonomic Computing and Transformations of Human Agency*, M. Hildebrandt and A. Rouvroy, Eds. New York: Routledge, 2011.

Classical Ethics in A/IS

- D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "[Algorithmic Impact Assessments: A practical Framework for Public Agency Accountability](#)," AI NOW, April 2018.
- A. Chaudhuri, "[Philosophical Dimensions of Information and Ethics in the Internet of Things \(IoT\) Technology](#)," EDPACS, vol. 56, no. 4, pp. 7-18, Nov. 2017.

Issue: Implications of Cultural Migration in A/IS

Background

In addition to developing an understanding of A/IS via different cultures, it is crucial to understand how A/IS are shaped and reshaped—how they affect and are affected by—human mobility and cultural diversity through active immigration. The effect of human mobility on state systems reliant on A/IS impacts the State structure itself, and thus the systems that the structure relies on, in the end influencing everything from democracy to citizenship. Where the State, through A/IS, invests in and gathers big data through mechanisms for registration and identification of people, mainly immigrants, human mobility becomes a foundational component in a system geared toward the preservation of human dignity.

Traditional national concerns reflect two information foundations: information produced for human rights and information produced for national sovereignty. In the second foundation, State borders are considered the limits from which political governance is defined in terms of

security. The preservation of national sovereignty depends on the production and domination of knowledge. In the realm of migratory policies, knowledge is created to measure people in transit: collecting, treating, and transferring information about territory and society.

Knowledge organization has been the paramount pillar of scientific thought and scientific practice since the beginning of written civilization. Any scientific and technological development has only been possible through information policies that include the establishment of management processes to systematize them, and the codification of language. For the Greeks, this process was closely associated with the concept of *arete*, or the excellence of one's self in politics as congregated in the *polis*. The notion of *polis* is as relevant as ever in the digital age with the development of digital technologies and the discussions around morality in A/IS. Where the systematization of knowledge is potentially freely created, the advent of the Internet and its flows are difficult to control. Ethical issues about the production of information are becoming paramount to our digital society.

The advancement of the fields of science and technology has not been followed by innovations in the political community, and the technical community has repeatedly tabled academic discussions about the hegemony of technocracy over policy issues, restricting the space of the policy arena and valorizing excessively technic solutions for human problems. This monopoly alters conceptions of morality, relocating the locus of the Kantian "Categorical Imperative", causing the tension among different social and political contexts to become more pervasive.

Classical Ethics in A/IS

Current global migration dynamics have been met by unfavorable public opinion based in ideas of crisis and emergency, a response vastly disproportionate to what statistics have shown to be the reality. In response to these views, A/IS are currently designed and applied to measure, calculate, identify, register, systematize, normalize, and frame both human rights and security policies. This is largely no different of a process than what has been practiced since the period of colonialism. It includes the creation and implementation of a set of ancient and new technologies. Throughout history, mechanisms have been created firstly to identify and select individuals who share certain biological heritage, and secondly to individuals and social groups, including biological characteristics.

Information is only possible when materialized as an infrastructure supported by ideas in action as a “communicative act”, which Habermas (1968) identifies in Hegel’s work, converging three elements in human-in-the-world relationships: symbol, language, and labor. Information policies reveal the importance and the strength in which technologies influence economic, social, cultural, identity, and ethnic interactions.

Traditional mechanisms used to control migration, such as the passport, are associated with globally established walls and fences. The more intense human mobility becomes, the more amplified are the discourses to discourage it, restricting human migrations, and deepening the need for an ethics related to conditions of citizenship. Together with the building of walls, other remote technologies are developed to monitor and surveil borders, buildings, and streets, also impacting ideas and

moral presumptions of citizenship. Closed Circuit Television(CCTV), Unmanned Aerial Vehicles (UAVs), and satellites allow data transference in real time to databases, cementing the backbone that A/IS draws from, often with bias as per the expectations of developed countries. This centrality of data sources for A/IS expresses a divide between developed and underdeveloped countries, particularly as relevant to the refugee.

Information is something that links languages, habits, customs, identification, and registration technologies. It provokes a reshaping of the immigrants’ and refugees’ citizenship and their value as people in terms of their citizenship, as they seek forms of surviving in, and against, the restrictions imposed by A/IS for surveillance and monitoring in an enlarged and more complex cosmopolis.

An understanding of the impact of A/IS on migration and mobile populations, as used in state systems, is a critical first step to consider if systems are to become truly autonomous and intelligent, especially beyond the guidance of human deliberation. Digital technology systems used to register and identify human mobility, including refugees and other displaced populations, are not autonomous in the intelligent sense, and are dependent on the biases of worldviews around immigration. In this aspect, language is the locus where this dichotomy has to be considered to understand the diversity of morals when there are contacts among different cultures.

Classical Ethics in A/IS

Recommendations

Is it recommended that the State become a proactive player in the globalized processes of A/IS for migrant and mobile populations, introducing a series of mechanisms that limit the segregation of social spaces and groups, and consider the biases inherent in surveillance for control.

Further Resources

- I. About and V. Denis, *Histoire de l'identification des personnes*. Paris: La Découverte, 2010.
- I. About, J. Brown, G. Lonergan, *Identification and Registration Practices in Transnational Perspective: People, Papers and Practices*. London: Palgrave Macmillan, 2013, pp. 1-13.
- D. Bigo, "Security and Immigration: Toward a Critique of the Governmentality of Unease," in *Alternatives*, Special Issue, no. 27. pp. 63-92, 2002.
- R. Capurro, "[Citizenship in the Digital Age](#)," in *Information Ethics, Globalization and Citizenship*, T. Samek and L. Schultz, Eds. Jefferson NC: McFarland, 2017, pp. 11-30.
- R. Capurro, "[Intercultural Information Ethics](#)," in *Localizing the Internet: Ethical Aspects in Intercultural Perspective*, R. Capurro, J. Frühbauer, and T. Hausmanninger, Eds. Munich: Fink, 2007, pp. 21-38.
- UN High Commissioner for Refugees (UNHCR), [Policy on the Protection of Personal Data of Persons of Concern to UNHCR](#), May 2015.

Issue: Applying Goal-Directed Behavior (Virtue Ethics) to Autonomous and Intelligent Systems

Background

Initial concerns regarding A/IS also include questions of function, purpose, identity, and agency, a continuum of goal-directed behavior with function being the most primitive expression. How can classical ethics act as a regulating force in autonomous technologies as goal-directed behavior transitions from being externally set by operators to being internally set? The question is important not just for safety reasons, but for mutual productivity. If autonomous systems are to be our trusted, creative partners, then we need to be confident that we possess mutual anticipation of goal-directed action in a wide variety of circumstances.

A virtue ethics approach has merits for accomplishing this even without having to posit a "character" in an autonomous technology, since it places emphasis on habitual, iterative action focused on achieving excellence in a chosen domain or in accord with a guiding purpose. At points on the goal-directed continuum associated with greater sophistication, virtue ethics become even more useful by providing a framework for prudent decision-making that is in keeping with the autonomous system's purpose, but allows for creativity in how to achieve the purpose in a way that still allows for a degree of predictability. An ethics approach that does not rely on a decision

Classical Ethics in A/IS

to refrain from transgressing, but instead to prudently pursue a sense of purpose informed by one's identity, might provide a greater degree of insight into the behavior of the system.

Recommendations

Program autonomous systems to be able to recognize user behavior for the purposes of predictability, traceability, and accountability and to hold expectations, as an operator and co-collaborator, whereby both user and system mutually recognize the decisions of the autonomous system as virtue ethics-based.

Further Resources

- M. A. Boden, Ed. *The Philosophy of Artificial Life*. Oxford, U.K.: Oxford University Press, 1996.
- C. Castelfranchi, "Modelling Social Action for AI Agents," *Artificial Intelligence*, vol. 103, no.1-2, pp. 157-182, 1998.
- W. D. Christensen and C. A. Hooker, "Anticipation in Autonomous Systems: Foundations for a Theory of Embodied Agents," *International Journal of Computing Anticipatory Systems*, vol. 5, pp. 135-154, Dec. 2000.
- K. G. Coleman, "Android Arete: Toward a Virtue Ethic for Computational Agents," *Ethics and Information Technology*, vol. 3, no. 4, pp. 247-265, 2001.
- J. G. Lennox, "Aristotle on the Biological Roots of Virtue," *Biology and the Foundations of Ethics*, J. Maienschein and M. Ruse, Eds. Cambridge, U.K.: Cambridge University Press, 1999, pp. 405-438.
- L. Muehlhauser and L. Helm, "The Singularity and Machine Ethics," in *Singularity Hypotheses*, A. H. Eden, J. H. Moor, J. H. Soraker, and E. Steinhart, Eds. Berlin: Springer, 2012, pp. 101-126.
- D. Vernon, G. Metta, and G. Sandini, "[A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents](#)," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 151-180, April 2007.

Issue: A Requirement for Rule-Based Ethics in Practical Programming

Background

Research in machine ethics focuses on simple moral machines. It is deontological ethics and [teleological ethics](#) that are best suited to the kind of practical programming needed for such machines, as these ethical systems are abstractable enough to encompass ideas of non-human agency, whereas most modern ethics approaches are far too human-centered to properly accommodate the task.

In the deontological model, duty is the point of departure. Duty can be translated into rules. It can be distinguished into rules and metarules. For example, a rule might take the form "Don't lie!", whereas a metarule would take the form of Kant's categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

Classical Ethics in A/IS

A machine can follow simple rules. Rule-based systems can be implemented as formal systems, also referred to as “axiomatic systems”, and in the case of machine ethics, a set of rules is used to determine which actions are morally allowable and which are not. Since it is not possible to cover every situation by a rule, an [inference engine](#) is used to deduce new rules from a small set of simple rules called axioms by combining them. The morality of a machine comprises the set of rules that is deducible from the axioms.

Formal systems have an advantage since properties such as decidability and consistency of a system can be effectively examined. If a formal system is decidable, every rule is either morally allowable or not, and the “unknown” is eliminated. If the formal system is consistent, one can be sure that no two rules can be deduced that contradict each other. In other words, the machine never has moral doubt about an action and never encounters a deadlock.

The disadvantage of using formal systems is that many of them work only in closed worlds like computer games. In this case, what is not known is assumed to be false. This is in drastic conflict with real world situations, where rules can conflict and it is impossible to take into account the totality of the environment. In other words, consistent and decidable formal systems that rely on a closed world assumption can be used to implement an ideal moral framework for a machine, yet they are not viable for real world tasks.

One approach to avoiding a closed world scenario is to utilize self-learning algorithms, such as case-

based reasoning approaches. Here, the machine uses “experience” in the form of similar cases that it has encountered in the past or uses cases which are collected in databases.

In the context of the *teleological model*, the consequences of an action are assessed. The machine must know the consequences of an action and what the action’s consequences mean for humans, for animals, for things in the environment, and, finally, for the machine itself. It also must be able to assess whether these consequences are good or bad, or if they are acceptable or not, and this assessment is not absolute. While a decision may be good for one person, it may be bad for another; while it may be good for a group of people or for all of humanity, it may be bad for a minority of people. An implementation approach that allows for the consideration of potentially contradictory subjective interests may be realized by decentralized reasoning approaches such as agent-based systems. In contrast to this, centralized approaches may be used to assess the overall consequences for all involved parties.

Recommendations

By applying the classical methodologies of deontological and teleological ethics to machine learning, rules-based programming in A/IS can be supplemented with established praxis, providing both theory and a practicality toward consistent and determinable formal systems.

Classical Ethics in A/IS

Further Resources

- C. Allen, I. Smit, and W. Wallach, "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149-155, 2005.
- O. Bendel, [*Die Moral in der Maschine: Beiträge zu Roboter-und Maschinenethik*](#). Heise Medien, 2016.
- O. Bendel, Oliver, *Handbuch Maschinenethik*. Wiesbaden, Germany: Springer VS, 2018.
- M. Fisher, L. Dennis, and M. Webster, "[Verifying Autonomous Systems](#)," *Communications of the ACM*, vol. 56, no. 9, pp. 84-93, Sept. 2013.
- B. M. McLaren, "[Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions](#)," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 29-37, July 2006.
- M. A. Perez Alvarez, "[Tecnologías de la Mente y Exocerebro o las Mediaciones del Aprendizaje](#)," 2015.
- E. L. Rissland and D. B. Skalak, "Combining Case-Based and Rule-Based Reasoning: A Heuristic Approach." *Proceedings of the 11th International Joint Conference on Artificial Intelligence, IJCAI 1989*, Detroit, MI, August 20-25, 1989, San Francisco, CA: Morgan Kaufmann Publishers Inc., 1989. pp. 524-530.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Classical Ethics in A/IS Committee

- **Jared Bielby** (Chair) – President, Netizen Consulting Ltd; Chair, International Center for Information Ethics; editor, *Information Cultures in the Digital Age*
- **Soraj Hongladarom** (Co-chair) – President at The Philosophy and Religion Society of Thailand
- **Miguel Á. Pérez Álvarez** – Professor of Technology in Education, Colegio de Pedagogía, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México
- **Oliver Bendel** – Professor of Information Systems, Information Ethics and Machine Ethics, University of Applied Sciences and Arts Northwestern Switzerland FHNW
- **Dr. John T. F. Burgess** – Assistant Professor / Coordinator for Distance Education, School of Library and Information Studies, The University of Alabama
- **Rafael Capurro** – Founder, International Center for Information Ethics
- **Corinne Cath-Speth** – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Dr. Paola Di Maio** – Center for Technology Ethics, ISTCS.org UK and NCKU Taiwan
- **Robert Donaldson** – Independent Computer Scientist, BMRILLC, Hershey, PA

Classical Ethics in A/IS

- **Rachel Fischer** – Research Officer: African Centre of Excellence for Information Ethics, Information Science Department, University of Pretoria, South Africa.
- **Dr. D. Michael Franklin** – Assistant Professor, Kennesaw State University, Marietta Campus, Marietta, GA
- **Wolfgang Hofkirchner** – Associate Professor, Institute for Design and Technology Assessment, Vienna University of Technology
- **Dr. Tae Wan Kim** – Associate Professor of Business Ethics, Tepper School of Business Carnegie Mellon University
- **Kai Kimppa** – University Research Fellow, Information Systems, Turku School of Economics, University of Turku
- **Sara R. Mattingly-Jordan** – Assistant Professor Center for Public Administration & Policy, Virginia Tech
- **Dr Neil McBride** – Reader in IT Management, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Bruno Macedo Nathansohn** – Perspectivas Filosóficas em Informação (Perfil-i); Brazilian Institute of Information in Science and Technology (IBICT)
- **Marie-Therese Png** – PhD Student, Oxford Internet Institute, PhD Intern, DeepMind Ethics & Society
- **Derek Poitras** – Independent Consultant, Object Oriented Software Development
- **Samuel T. Segun** – PhD Candidate, Department of Philosophy, University of Johannesburg. Fellow, Philosophy Node of the Centre for Artificial Intelligence Research (CAIR) at the University of Pretoria and Research fellow at the Conversational School of Philosophy (CSP)
- **Dr. Ozlem Ulgen** – Reader in International Law and Ethics, School of Law, Birmingham City University
- **Kristene Unsworth** – Assistant Professor, The College of Computing & Informatics, Drexel University
- **Dr. Xiaowei Wang** – Associate professor of Philosophy, Renmin University of China
- **Dr Sara Wilford** – Senior Lecturer, Research Fellow, School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University
- **Pak-Hang Wong** – Research Associate, Department of Informatics, University of Hamburg
- **Bendert Zevenbergen** – Oxford Internet Institute, University of Oxford & Center for Information Technology Policy, Princeton University

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Endnotes

¹ This edition of “Classical Ethics in A/IS” does not (and could not) aspire to universal coverage of all of the world’s traditions in the space available to us. Future editions will touch on several other traditions, including Judaism and Islam.

² R. Von Schomberg, “Prospects for Technology Assessment in a Framework of Responsible Research and Innovation” in *Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode*. Wiesbaden, Germany: Springer VS, 2011, pp. 39-61.

Well-being

Prioritizing ethical and responsible artificial intelligence has become a widespread goal for society. Important issues of transparency, accountability, algorithmic bias, and value systems are being directly addressed in the design and implementation of autonomous and intelligent systems (A/IS). While this is an encouraging trend, a key question still facing technologists, manufacturers, and policymakers alike is how to assess, understand, measure, monitor, safeguard, and improve the well-being impacts of A/IS on humans. Finding the answer to this question is further complicated when A/IS are within a holistic and interconnected framework of well-being in which individual well-being is inseparable from societal, economic, and environmental systems.

For A/IS to demonstrably advance well-being, we need consistent and multidimensional indicators that are easily implementable by the developers, engineers, and designers who are building our future. This chapter is intended for such developers, engineers, and designers—referred to in this chapter as “A/IS creators”. Those affected by A/IS are referred to as “A/IS stakeholders”.

A/IS technologies affect human agency, identity, emotion, and ecological systems in new and profound ways. Traditional metrics of success are not equipped to ensure A/IS creators can avoid unintended consequences or benefit from unexpected innovation in the algorithmic age. A/IS creators need expanded ways to evaluate the impact of their products, services, or systems on human well-being. These evaluations must also be done with an understanding that human well-being is deeply linked to the well-being of society, economies, and ecosystems.

Today, A/IS creators largely measure success using metrics including profit, gross domestic product (GDP), consumption levels, and occupational safety. While important, these metrics fail to encompass the full spectrum of well-being impacts on individuals and society, such as psychological, social, and environmental factors. Where the priority given to these factors is not equal to that given to fiscal metrics of success, A/IS creators risk causing or contributing to negative and irreversible harms to our people and our planet.

When A/IS creators are not aware that well-being indicators, in addition to traditional metrics, can provide guidance for their work, they are also missing out on innovation that can increase well-being and societal value. For instance, while it is commonly recognized that autonomous vehicles will save lives when safely deployed, a topic of less frequent discussion is how self-

Well-being

driving cars also have the potential to help the environment by [reducing greenhouse gas emissions and increasing green space](#). Autonomous vehicles can also positively impact well-being by increasing work-life balance and enhancing the quality of time spent during commutes.

Unless A/IS creators are made aware of the existence of alternative measures of progress, the value they provide, and the way they can be incorporated into A/IS work, technology and society will continue to rely upon traditional metrics of success. In an era where innovation is defined by holistic prosperity, alternative measures are needed more now than ever before. The 2009 [Report by the Commission on the Measurement of Economic Performance and Social Progress](#) which contributed substantially to the worldwide movement of governments using wider measures of well-being, states, “What we measure affects what we do; and if our measurements are flawed, decisions may be distorted.”

We believe that A/IS creators can profoundly increase human and environmental flourishing by prioritizing well-being metrics as an outcome in all A/IS system designs—now and for the future. *The primary intended audience for this chapter is A/IS creators who are unfamiliar with the term “well-being” as it is used in the field of positive psychology and well-being studies. Our initial goal is to provide a broad introduction to qualitative and quantitative metrics and applications of well-being to educate and inspire A/IS creators. We do not prioritize or advocate for any specific indicator or methodology. For further elaboration on the definition of well-being, please see the first Issue listed in Section 1.*

This chapter is divided into two main sections:

- [The Value of Well-being Metrics for A/IS Creators](#)
- [Implementing Well-being Metrics for A/IS Creators](#)

The following resources are available online to provide readers with an introduction to existing well-being metrics and tools currently in use:

- [The State of Well-being Metrics](#)
- [The Happiness Screening Tool for Business Product Decisions](#)
- [Additional Resources: Standards Development Models and Frameworks](#)

Well-being

Section 1—The Value of Well-being Metrics for A/IS Creators

Well-being metrics provide a broader perspective for A/IS creators than they normally might be familiar with in evaluating their products. This broader perspective unlocks greater opportunities to assure a positive impact of A/IS on human well-being, while minimizing the risk of unintended negative outcomes. This section defines well-being, discusses the value of well-being metrics to A/IS creators, and notes how similar frameworks like sustainability and human rights can be complemented by incorporating well-being metrics.

Definition of Well-being

For the purposes of *Ethically Aligned Design*, the term “well-being” refers to an evaluation of the general quality of life of an individual and the state of external circumstances. The conception of well-being encompasses the full spectrum of personal, social, and environmental factors that enhance human life and on which human life depend. The concept of well-being shall be considered distinct from moral or legal evaluation.

Issue: There is ample and robust science behind well-being metrics and their use by international and national institutions. However, A/IS creators are often unaware that well-being metrics exist, or that they can be used to plan, develop, and evaluate technology.

Background

The concept of well-being refers to an evaluation of the general goodness of the state of an individual or community and is distinct from moral or legal evaluation. A well-being evaluation takes into account major aspects of a person’s life, such as their happiness, success in their goals, and their overall positive functioning in their environment. There is now a thriving area of scientific research into the psychological, social, behavioral, economic, and environmental determinants of human well-being.

Well-being

The term “well-being” is defined and used in various ways across different contexts and fields. For example: economists identifying economic welfare with levels of consumption and economic vitality, psychologists highlighting subjective experience, and sociologists emphasizing living, labor, political, social, and environmental conditions. We do not take a stand on any specific measure of well-being. The metrics listed below are an incomplete list and provided as a starting point for further inquiry. Among these are subjective well-being indicators, measures of quality of life, social progress and capabilities, and many more.

There is now sufficient consensus among scientists that well-being can be reliably measured. Well-being measures differ in the number and the intricacy of indicators they employ. Short questionnaires of life satisfaction have emerged as particularly popular, although they do not reflect all aspects of well-being. While recognizing a scope for differences across well-being indicators, we note that the richest conception of well-being encompasses the full spectrum of personal, social, and environmental goods that enhance human life.

We encourage A/IS creators to consider the wide range of available indicators and select those most relevant and revealing for particular stages of the A/IS technology’s life cycle and the particular context for the technology’s use and evaluation. That is, measures of well-being that may be well-suited to wealthy, industrialized nations may be less applicable in low- and middle-income countries, and vice versa.

Among the most important and recognized aspects of well-being are (in alphabetical order):

- Community: Belonging, Crime & Safety, Discrimination & Inclusion, Participation, Social Support
- Culture: Identity, Values
- Economy: Economic Policy, Equality & Environment, Innovation, Jobs, Sustainable Natural Resources & Consumption & Production, Standard of Living
- Education: Formal Education, Lifelong Learning, Teacher Training
- Environment: Air, Biodiversity, Climate Change, Soil, Water
- Government: Confidence, Engagement, Human Rights, Institutions
- Human Settlements: Energy, Food, Housing, Information & Communication Technology, Transportation
- Physical Health: Health Status, Risk Factors, Service Coverage
- Psychological Health: Affect (feelings), Flourishing, Mental Illness & Health, Satisfaction with Life
- Work: Governance, Time Balance, Workplace Environment

Well-being

In an effort to provide a basic orientation to well-being metrics, information about well-being indicators can be segmented into four categories:

1. Subjective or survey-based indicators

Survey-based well-being indicators, subjective well-being (SWB) indicators, and multidimensional measurements of aspects of well-being, are being used by national institutions, international institutions, and governments to better understand levels of psychological well-being within countries and aspects of a country's population. These indicators are also being used to understand people's satisfaction in specific domains of life. Examples of surveys that include survey-based well-being indicators and SWB indicators include the [European Social Survey](#), [Bhutan's Gross National Happiness Indicators](#), well-being surveys created by [The UK Office for National Statistics](#), and many more.

Survey-based metrics are also employed in the field of positive psychology and in the [World Happiness Report](#). The data are employed by researchers to understand the causes, consequences, and correlates of well-being. Data gathered from surveys tend to address concerns, such as day-to-day experience, overall satisfaction with life, and perceived flourishing. The findings of these researchers provide crucial and necessary guidance because they often diverge from and complement the understanding of traditional conditions, such as economic growth.

2. Objective indicators

Objective indicators of quality of life have typically incorporated areas such as income, consumption, health, education, crime, housing, etc. These indicators have been used to understand

conditions that support the well-being of countries and populations, and to measure the societal and environmental impact of companies. They are in use by organizations like the OECD with their [Better Life Index](#), which also includes survey-based well-being indicators and SWB indicators, and the United Nations with their [Sustainable Development Goals Indicators](#) (formerly the Millennium Development Goals). For business, the [Global Reporting Initiative](#), [SDG Compass](#), and [B-Corp](#) provide broad indicator sets.

3. Composite indicators (indices that aggregate multiple metrics)

Aggregate metrics combine subjective and/or objective metrics to produce one measure reflecting both objective aspects of quality of life and people's subjective evaluation of these. Examples of this are the [UN's Human Development Index](#), the [Social Progress Index](#), and the [United Kingdom's Office of National Statistics Measures of National Well-being](#). Some subjective and objective indicators are also composite indicators, such as Bhutan's Gross National Happiness Index and the OECD's Better Life Index.

4. Social media sourced data

Social media can be used to measure the well-being of a geographic region or demographic group, based on sentiment analysis of publicly available data. Examples include the [the Hedonometer](#) and the [World Well-being Project](#).

Well-being

Recommendation

A/IS creators should prioritize learning about well-being concepts, scientific learnings, research findings, and well-being metrics as potential determinants for how they create, deploy, market, and monitor their technologies, and ensuring their stakeholders learn the same. This process can be expedited if Standards Development Organizations (SDOs), such as the IEEE Standards Association, or other institutions such as the Global Reporting Initiative (GRI) or B-Corp, create certifications, guidelines, and standards that for the use of holistic, well-being metrics for A/IS in the public and private sectors.

Further Resources

- The IEEE P7010™ Standards Project for [Well-being Metric for Autonomous/Intelligent Systems](#), was formed with the aim of identifying well-being metrics for applicability to A/IS today and in the future. All are welcome to join the working group.
- On 11 April 2017, IEEE hosted a dinner debate at the European Parliament in Brussels to discuss how the world's top metric of value, gross domestic product, must move [Beyond GDP](#) to holistically measure how intelligent and autonomous systems can hinder or improve human well-being.
- [Prioritizing Human Well-being in the Age of Artificial Intelligence](#) (Report)
- [Prioritizing Human Well-being in the Age of Artificial Intelligence](#) (Video)

Issue: Increased awareness and application of well-being metrics by A/IS creators can create greater value, safety, and relevance to corporate communities and other organizations in the algorithmic age.

Background

While many organizations in the private and public sectors are increasingly aware of the need to incorporate well-being measures as part of their efforts, the reality is that bottom line, quarterly-driven shareholder growth remains a dominant goal and metric. Short term growth is often the priority in the private sector and public sector. As long as organizations exist in a larger societal system which prioritizes financial success, these companies will remain under pressure to deliver financial results that do not fully incorporate societal and environmental impacts, measurements, or priorities.

Rather than focus solely on the negative aspects of how A/IS could harm humans and environments, we seek to explore how the implementation of well-being metrics can help A/IS to have a measurable, positive impact on human well-being as well as on systems and organizations. Incorporation of well-being goals and measures beyond what is strictly required can benefit both private sector organizations' brands and public sector organizations' stability and reputation, as well as help realize financial

Well-being

savings, innovation, trust, and many other benefits. For instance, a companion robot outfitted to support seniors in assisted living situations might traditionally be launched with a technology development model that was popularized by Silicon Valley known as “move fast and break things”. The A/IS creator who rushed to bring the robot to market faster than the competition and who was unaware of well-being metrics, may have overlooked critical needs of the seniors. The robot might actually hurt the senior instead of helping by exacerbating isolation or feelings of loneliness and helplessness. While this is a hypothetical scenario, it is intended to demonstrate the value of linking A/IS design to well-being indicators.

By prioritizing largely fiscal metrics of success, A/IS devices might fail in the market because of limited adoption and subpar reception. However, if during use of the A/IS product, success were measured in terms of relevant aspects of well-being, developers and researchers could be in a better position to attain funding and public support. Depending on the intended use of the A/IS product, well-being measures that could be used extend to emotional levels of calm or stress; psychological states of thriving or depression; behavioral patterns of engagement in community or isolation; eating, exercise and consumption habits; and many other aspects of human well-being. The A/IS product could significantly improve quality of life guided by metrics from trusted sources, such as the [World Health Organization](#), [European Social Survey](#), and [Sustainable Development Goal Indicators](#).

Thought leaders in the corporate arena have recognized the multifaceted need to utilize metrics beyond fiscal indicators.

PricewaterhouseCoopers defines “[total impact](#)” as a “holistic view of social, environmental, fiscal and economic dimensions—the big picture”. Other thought-leading organizations in the public sector, such as the OECD, demonstrate the desire for business leaders to incorporate metrics of success beyond fiscal indicators for their efforts, exemplified in their 2017 workshop, [Measuring Business Impacts on People’s Well-Being](#). The [B-Corporation movement](#) has created a new legal status for “a new type of company that uses the power of business to solve social and environmental problems”. Focusing on increasing stakeholder value versus shareholder returns alone, B-Corps are defining their brands by provably aligning their efforts with wider measures of well-being.

Recommendations

A/IS creators should work to better understand and apply well-being metrics in the algorithmic age. Specifically:

- A/IS creators should work directly with experts, researchers, and practitioners in well-being concepts and metrics to identify existing metrics and combinations of indicators that would bring support a “triple bottom line”, i.e., accounting for economic, social, and environmental impacts, approach to well-being. However, well-being metrics should only be used with consent, respect for privacy, and with strict standards for collection and use of these data.
- For A/IS to promote human well-being, the well-being metrics should be chosen in collaboration with the populations most affected by those systems—the A/IS

Well-being

stakeholders—including both the intended end-users or beneficiaries and those groups whose lives might be unintentionally transformed by them. This selection process should be iterative and through a learning and continually improving process. In addition, “metrics of well-being” should be treated as vehicles for learning and potential mid-course corrections. The effects of A/IS on human well-being should be monitored continuously throughout their life cycles, by A/IS creators and stakeholders, and both A/IS creators and stakeholders should be prepared to significantly modify, or even roll back, technology that is shown to reduce well-being, as defined by affected populations.

- A/IS creators in the business or academic, engineering, or policy arenas are advised to review the additional resources on standards development models and frameworks at the end of this chapter to familiarize themselves with existing indicators relevant to their work.

Further Resources

- PricewaterhouseCoopers (PwC). [Managing and Measuring Total Impact: A New Language for Business Decisions](#), 2017.
- World Economic Forum. [The Inclusive Growth and Development Report 2017](#), Geneva, Switzerland: World Economic Forum, January 16, 2017.
- [OECD Guidelines on Measuring Subjective Well-being](#), 2013.
- National Research Council. [Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience](#). DC: The National Academies Press, 2013.

Issue: A/IS creators have opportunities to safeguard human well-being by ensuring that A/IS does no harm to earth’s natural systems or that A/IS contributes to realizing sustainable stewardship, preservation, and/or restoration of earth’s natural systems. A/IS creators have opportunities to prevent A/IS from contributing to the degradation of earth’s natural systems and hence losses to human well-being.

Background

It is unwise, and in truth impossible, to separate the well-being of the natural environment of the planet from the well-being of humanity. A range of studies, from the [historic](#) to more [recent](#), prove that ecological collapse endangers human existence. Hence, the concept of well-being should encompass planetary well-being. Moreover, biodiversity and ecological integrity have intrinsic merit beyond simply their instrumental value to humans.

Technology has a long history of contributing to ecological degradation through its role in expanding the scale of resource extraction and environmental pollution, for example, the immense power needs of network computing, which leads to [climate change](#), [water scarcity](#), [soil degradation](#), [species extinction](#), [deforestation](#),

Well-being

[biodiversity loss](#), and destruction of ecosystems which in turn threatens humankind in the long run. These and other costs are often considered externalities and often do not figure into decisions or plans. At the same time, there are many examples, such as photovoltaics and smart grid technology that present potential ways to restore earth's ecosystems if undertaken within a systems approach aimed at sustainable economic and environmental development.

Environmental justice [research](#) demonstrates that the negative environmental impacts of technology are commonly concentrated on the middle class and working poor, as well as those suffering from abject poverty, fleeing disaster zones, or otherwise lacking the resources to meet their needs. Ecological impact can thus exacerbate the economic and sociological effects of wealth disparities on human well-being by concentrating environmental injustice onto those who are less well off. Moreover, [well-being research findings](#) indicate that unfair economic and social inequality has a dampening effect on everyone's well-being, regardless of economic or social class.

In these respects, A/IS are no exception; they can be used in ways that either help or harm the ecological integrity of the planet. It may be fair to say that ecological health and human well-being will, increasingly, depend upon A/IS creators. It is imperative that A/IS creators and stakeholders find ways to use A/IS to do no harm and to reduce the environmental degradation associated with economic growth—while simultaneously identifying applications to restore the ecological health of the planet and thereby safeguarding the well-being of humans. For A/IS to reduce environmental degradation and promote well-

being, it is required that not only A/IS creators act along such lines, but also that a systems approach is taken by all A/IS stakeholders to find solutions that safeguard human well-being with the understanding that human well-being is inextricable from healthy social, economic, and environmental systems.

Recommendations

A/IS creators need to recognize and prioritize the stewardship of the Earth's natural systems to promote human and ecological well-being. Specifically:

- Human well-being should be defined to encompass ecological health, access to nature, safe climate and natural environments, biosystem diversity, and other aspects of a healthy, sustainable natural environment.
- A/IS systems should be designed to use, support, and strengthen existing ecological sustainability standards with a certification or similar system, e.g., [LEED](#), [Energy Star](#), or [Forest Stewardship Council](#). This directs automation and machine intelligence to follow the principle of doing no harm and to safeguard environmental, social, and economic systems.
- A/IS creators should prioritize doing no harm to the Earth's natural systems, both intended and unintended harm.
- A committee should be convened to issue findings on ways in which A/IS can be used by business, NGOs, and governmental agencies to promote stewardship and restoration of natural systems while reducing the harmful impact of economic development on ecological sustainability and environmental justice.

Well-being

Further Resources

- D. Austin and M. Macauley. "[Cutting Through Environmental Issues: Technology as a double-edged sword.](#)" The Brookings Institution, Dec. 2001 [Online]. Available: <https://www.brookings.edu/articles/cutting-through-environmental-issues-technology-as-a-double-edged-sword/>. [Accessed Dec. 1, 2018].
- J. Newton, [Well-being and the Natural Environment: An Overview of the Evidence](#). August 20, 2007.
- P. Dasgupta, [Human Well-Being and the Natural Environment](#). Oxford, U.K.: Oxford University Press, 2001.
- R. Haines-Young and M. Potschin. "[The Links Between Biodiversity, Ecosystem Services and Human Well-Being](#)," in *Ecosystem Ecology: A New Synthesis*, D. Raffaelli, and C. Frid, Eds. Cambridge, U.K.: Cambridge University Press, 2010.
- S. Hart, [Capitalism at the Crossroads: Next Generation Business Strategies for a Post-Crisis World](#). Upper Saddle River, NJ: Pearson Education, 2010.
- United Nations Department of Economic and Social Affairs. "[Call for New Technologies to Avoid Ecological Destruction](#)." Geneva, Switzerland, July 5, 2011.
- Pope Francis. [Encyclical Letter Laudato Si' of the Holy Father Francis On the Care for Our Common Home](#). May 24, 2015.
- "[Environment](#)," The 14th Dalai Lama. Accessed Dec. 9, 2018. <https://www.dalailama.com/messages/environment>.
- Why Islam.org, Environment and Islam, 2018.

Issue: Human rights law is related to, but distinct from, the pursuit of well-being. Incorporating a human-rights framework as an essential basis for A/IS creators means A/IS creators honor existing law as part of their well-being analysis and implementation.

Background

International human rights law has been firmly established for decades in order to protect various guarantees and freedoms as enshrined in charters such as the United Nations' [Universal Declaration of Human Rights](#) and the Council of Europe's [Convention on Human Rights](#). In 2018, the [Toronto Declaration](#) on machine learning standards was released, calling on both governments and technology companies to ensure that algorithms respect basic principles of equality and non-discrimination. The Toronto Declaration sets forth an obligation to prevent machine learning systems from discriminating, and in some cases violating, existing human rights law.

Well-being initiatives are typically undertaken for the sake of public interest. However, any metric, including well-being metrics, can be misused to justify human rights violations. Encampment and mistreatment of refugees and ethnic cleansing undertaken to preserve a nation's culture (an aspect of well-being) is one example. Imprisonment or assassination of journalists or researchers to ensure the stability

Well-being

of a government is another. The use of well-being metrics to justify human rights violations is an unconscionable perversion of the nature of any well-being metric. It should be noted that these same practices happen today in relation to GDP. For instance, in 2012, according to the [International Labour Organization](#) (ILO), approximately 21 million people are victims of forced labor (slavery), representing 9% to 56% of GDP income for various countries. These clear human rights violations, from sex trafficking and use of children in armies, to indentured farming or manufacturing labor, can increase a country's GDP while obviously harming human well-being.

Well-being metrics are designed to measure the efficacy of efforts related to individual and societal flourishing. Well-being as a value complements justice, equality, and freedom. Well-designed application of well-being considerations by A/IS creators should not displace other issues of human rights or ethical methodologies, but rather complement them.

Recommendation

A human rights framework should represent the floor, and not the ceiling, for the standards to which A/IS creators must adhere. Developers and users of well-being metrics should be aware these metrics will not always adequately address human rights.

Further Resources

- United Nations [Universal Declaration of Human Rights](#), 1948.
- Council of Europe's [Convention on Human Rights](#), 2018.
- International Labor Organization (ILO) [Declaration on Fundamental Principles and Rights at Work](#), 1998.
- The regularly updated [University of Minnesota Human Rights Library](#) provides a wealth of material on human rights laws, its history, and the organizations engaged in promoting them.
- The [Oxford Human Rights Hub](#) reports on how and why technologies surrounding artificial intelligence raise human rights issues.

Well-being

Section 2—Implementing Well-being Metrics for A/IS Creators

A key challenge for A/IS creators in realizing the benefits of well-being metrics is how to best incorporate them into their work. This section explores current best thinking on how to make this happen.

Issue: How can A/IS creators incorporate well-being into their work?

Background

Without practical ways of incorporating well-being metrics to guide, measure, and monitor impact, A/IS will likely lack fall short of its potential to avoid harm and promote well-being. Incorporating well-being thinking into typical organizational processes of design, prototyping, marketing, etc., suggests a variety of adaptations.

Organizations and A/IS creators should consider clearly defining the type of A/IS product or service that they are developing, including articulating its intended stakeholders and uses. By defining typical uses, possible uses, and finally unacceptable uses of the technology, creators will help to spell out the context of well-being. This can help to identify possible harms and risks given the different possible uses and end users, as well as intended and unintended positive consequences.

Additionally, internal and external stakeholders should be extensively consulted to ensure that impacts are thoroughly considered through an iterative and learning stakeholder engagement process. After consultation, A/IS creators should select appropriate well-being indicators based on the possible scope and impact of their A/IS product or service. These well-being indicators can be drawn from mainstream sources and models and adapted as necessary. They can be used to engage in pre-assessment of the intended user population, projection of possible impacts, and post-assessment. Development of a well-being indicator measurement plan and relevant data infrastructure will support a robust integration of well-being. A/IS models can also be trained to explicitly include well-being indicators as subgoals.

Data and discussions on well-being impacts can be used to suggest improvements and modifications to existing A/IS products and services throughout their lifecycle. For example, a [team seeking to increase the well-being](#) of people using wheelchairs found that when provided the opportunity to use a smart wheelchair, some users were delighted with the opportunity for more mobility, while others felt it would decrease their opportunities for social contact, increase their sense of isolation, and lead to an overall decrease in their well-being. Therefore, even though a product modification may increase well-being according to one indicator or set of

Well-being

A/IS stakeholders, it does not mean that this modification should automatically be adopted.

Finally, organizational processes can be modified to incorporate the above strategies. Appointment of an organizational lead person for well-being impacts, e.g., a well-being lead, ombudsman, or officer can help to facilitate this effort.

Recommendation

A/IS creators should adjust their existing development, marketing, and assessment cycles to incorporate well-being concerns throughout their processes. This includes identification of an A/IS lead ombudsperson or officer; identification of stakeholders and end users; determination of possible uses, harm and risk assessment; robust stakeholder engagement; selection of well-being indicators; development of a well-being indicator measurement plan; and ongoing improvement of A/IS products and services throughout the lifecycle.

Further Resources

- [Peter Senge and the Learning Organization](#) - (synopsis) Purdue University
- Stakeholder Engagement: A Good Practice Handbook for Companies Doing Business in Emerging Markets. International Finance Corporation, May 2007.
- [Global Reporting Initiative](#)
- [GNH Certification](#), Centre for Bhutan and GNH Studies, 2018.

- J. Helliwell, R. Layard, and J. Sachs, Eds., "The Objective Benefits of Subjective Well-Being," in [World Happiness Report](#) 2013. New York: UN Sustainable Development Solutions Network, pp. 54-79, 2013.
- [Global Happiness and Well-being Policy Report](#) by the Global Happiness Council, 2018.

Issue: How can A/IS creators influence A/IS goals to ensure well-being, and what can A/IS creators learn or borrow from existing models in the well-being and other arenas?

Background

Another way to incorporate considerations of well-being is to include well-being measures in the development, goal setting, and training of the A/IS systems themselves.

Identified metrics of well-being could be formulated as auxiliary objectives of the A/IS. As these auxiliary well-being objectives will be only a subset of the intended goals of the system, the architecture will need to balance multiple objectives. Each of these auxiliary objectives may be expressed as a goal, set of rules, set of values, or as a set of preferences, which can be weighted and combined using established methodologies from intelligent systems engineering.

Well-being

For example, an educational A/IS tool could not only optimize learning outcomes, but also incorporate measures of student social and emotional education, learning, and thriving.

A/IS-related data relates both to the individual—through personalized algorithms, in conjunction with affective sensors measuring and influencing emotion, and other aspects of individual well-being—and to society as large data sets representing aggregate individual subjective and objective data. As the exchange of this data becomes more widely available via establishing tracking methodologies, the data can be aligned within A/IS products and services to increase human well-being. For example, robots like [Pepper](#) are equipped to share data regarding their usage and interaction with humans to the cloud. This allows almost instantaneous innovation, as once an action is validated as useful for one Pepper robot, all other Pepper units (and ostensibly their owners) benefit as well. As long as this data exchange happens with the predetermined consent of the robots' owners, this innovation in real time model can be emulated for the large-scale aggregation of information relating to existing well-being metrics.

A/IS creators can also help to operationalize well-being metrics by providing stakeholders with reports on the expected or actual outcomes of the A/IS and the values and objectives embedded in the systems. This transparency will help creators, users, and third parties assess the state of well-being produced by A/IS and make improvements in A/IS. In addition, A/IS creators should consider allowing end users to layer on their own preferences, such as allowing users

to limit their use of an A/IS product if it leads to increased sustained stress levels, sustained isolation, development of unhealthy habits, or other decreases to well-being.

Incorporating well-being goals and metrics into broader organizational values and processes would support the use of well-being metrics as there would be institutional support. A key factor in industrial, corporate, and societal progress is cross-dissemination of concepts and models from one industry or field to another. To date, a number of successful concepts and models exist in the fields of sustainability, economics, industrial design and manufacturing, architecture and urban development, and governmental policy. These concepts and models can provide a foundation for building a metrics standard and the use of well-being metrics by A/IS creators, from conception and design to marketing, product updates, and improvements to the user experience.

Recommendation

Create technical standards for representing goals, metrics, and evaluation guidelines for well-being metrics and their precursors and components within A/IS that include:

- Ontologies for representing technological requirements.
- A testing framework for validating adherence to well-being metrics and ethical principles such as [IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems](#).

Well-being

- The exploration of models and concepts listed above as well as others as a basis for a well-being metrics standard for A/IS creators. (See page 191, [Additional Resources: Additional Resources: Standards Development Models and Frameworks](#))
- The development of a well-being metrics standard for A/IS that encompasses an understanding of well-being as holistic and interlinked to social, economic, and ecological systems.
- 2017), H. Trautmann, G. Rudolph, K. Klamroth, O. Schütze, M. Wiecek, Y. Jin, and C. Grimme, Eds., Vol. 10173. Springer-Verlag, Berlin, Heidelberg, 406-421, 2017.
- [PositiveSocialImpact](#): Empowering people, organizations and planet with information and knowledge to make a positive impact to sustainable development, 2017.
- D.K. Ura, Bhutan's [Gross National Happiness Policy Screening Tool](#).

Further Resources

- A.F.T Winfield, C. Blum, and W. Liu. "[Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection](#)," in Advances in Autonomous Robotics Systems. Springer, 2014, pp. 85–96
- R. A. Calvo, and D. Peters. [Positive Computing: Technology for Well-Being and Human Potential](#). Cambridge MA: MIT Press, 2014.
- Y. Collette, and P. Slarry. [Multiobjective Optimization: Principles and Case Studies](#) (Decision Engineering Series). Berlin, Germany: Springer, 2004. doi: 10.1007/978-3-662-08883-8.
- J. Greene, et al. "[Embedding Ethical Principles in Collective Decision Support Systems](#)," in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 4147–4151. Palo Alto, CA: AAAI Press, 2016.
- L. Li, I. Yevseyeva, V. Basto-Fernandes, H. Trautmann, N. Jing, and M. Emmerich, "[Building and Using an Ontology of Preference-Based Multiobjective Evolutionary Algorithms](#)." In 9th International Conference on Evolutionary Multi-Criterion Optimization—Volume 10173 (EMO

Issue: Decision processes for determining relevant well-being indicators through stakeholder deliberations need to be established.

Background

A/IS stakeholder involvement is necessary to determine relevant well-being indicators, for a number of reasons:

- "Well-being" will be defined differently by different groups affected by A/IS. The most relevant indicators of well-being may vary according to country, with concerns of wealthy nations being different than those of low- and middle-income countries. Indicators may vary based on geographical region or unique circumstances. The indicators may also be different across social groups, including gender, race, ethnicity, and disability status.
- Common indicators of well-being include satisfaction with life, healthy life expectancy,

Well-being

economic standard of living, trust in government, social support, perceived freedom to make life decisions, income equality, access to education, and poverty rates. Applying them in particular settings necessarily requires judgment, to ensure that assessments of well-being are in fact meaningful in context and reflective of the life circumstances of the diverse groups in question.

- Not all aspects of well-being are easily quantifiable. The importance of hard-to-quantify aspects of well-being is most likely to become apparent through interaction with those more directly affected by A/IS in specific settings.
- Engineers and corporate employees frequently misunderstand stakeholders' needs and expectations, especially when the stakeholders are very different from them in terms of educational and cultural background, social location, and/or economic status.

The processes through which stakeholders become involved in determining relevant well-being indicators will affect the quality of the indicators selected and assessed. Stakeholders should be empowered to define well-being, assess the appropriateness of existing indicators and propose new ones, and highlight context-specific factors that bear on issues of well-being, whether or not the issues have been recognized previously or are amenable to measurement. Interactive, open-ended discussions or deliberations among a wide variety of stakeholders and system designers are more likely to yield robust, widely-shared understandings of well-being and how to measure it in context. Closed-ended or over-determined methods for soliciting stakeholder input are likely to miss relevant information that system designers have not anticipated.

A process of stakeholder engagement and deliberation is one model for collective decision-making. Parties in such deliberation come together as equals. Their goal is to set aside their immediate, personal interests in order to think together about the common good. Participants in a stakeholder engagement and deliberation learn from one another's perspectives and experiences.

In the real world, stakeholder engagement and deliberation may run into the following challenges:

- Individuals with more education, power, or higher social status may—intentionally or unintentionally—dominate the discussion, undermining their ability to learn from less powerful participants.
- Topics may be preemptively ruled “out of bounds”, to the detriment of collective problem-solving. An example would be if, in a deliberation on well-being and A/IS, participants were told that worries about the costs of health insurance were unrelated to A/IS and thus could not be discussed.
- Engineers and scientists may claim authority over technical issues and be willing to deliberate only on social issues, obscuring the ways that technical and social issues are intertwined.
- Less powerful groups may be unable to keep more powerful ones “at the table” when discussions get contentious, and vice versa.
- Participants may not agree on who can legitimately be involved in the conversation. For example, the consensual spirit of deliberation is often used as a justification for excluding activists and others who already hold a position on the issue.

Well-being

Stakeholder engagement and deliberative processes can be effective when:

- Their design is guided by experts or practitioners who are experienced in deliberation models.
- Deliberations are facilitated by individuals sensitive to issues of power and are skilled in mediating deliberation sessions.
- Less powerful actors participate with the help of allies who can amplify their voices.
- More powerful actors participate with an awareness of their own power and make a commitment to listen with humility, curiosity, and open-mindedness.
- Deliberations are convened by institutions or individuals who are trusted and respected by all parties and who hold all actors accountable for participating constructively.

Ethically aligned design of A/IS would be furthered by thoughtfully constructed, context-specific deliberations on well-being and the best indicators for assessing it.

Recommendation

Appoint a lead team or person, “leads”, to facilitate stakeholder engagement and to serve as a resource for A/IS creators who use stakeholder-based processes to establish well-being indicators. Specifically:

- Leads should solicit and collect lessons learned from specific applications of stakeholder engagement and deliberation in order to continually refine its guidance.
- When determining well-being indicators, the leads should enlist the help of experts in public

participation and deliberation. With expert guidance, facilitators can provide guidance for how to: take steps to mitigate the effects of unequal power in deliberative processes; incorporate appropriately trained facilitators and coaching participants in deliberations; recognize and curb disproportionate influence by more-powerful groups; use techniques to maximize the voices of less-powerful groups.

- Leads should use their convening power to bring together A/IS creators and stakeholders, including critics of A/IS, for deliberations on well-being indicators, impacts, and other considerations for specific contexts and settings. Leads’ involvement would help bring actors to the table with a balance of power and encourage all actors to remain in conversation until robust, mutually agreeable definitions are found.

Further Resources

- D. E. Booher and J. E. Innes. Planning with Complexity: An Introduction to Collaborative Rationality for Public Policy. London: Routledge, 2010.
- J. A. Leydens and J. C. Lucena. Engineering Justice: Transforming Engineering Education and Practice. Wiley-IEEE Press, 2018.
- G. Ottinger. [Assessing Community Advisory Panels: A Case Study from Louisiana’s Industrial Corridor](#). Center for Contemporary History and Policy, 2008.
- [Expert and Citizen Assessment of Science and Technology \(ECAST\) Network](#)

Well-being

Issue: There are insufficient mechanisms to foresee and measure negative impacts, and to promote and safeguard positive impacts of A/IS.

Background

A/IS technologies present great opportunity for positive change in every aspect of society. However, they can—by design or unintentionally—cause harm as well. While it is important to consider and make sense of possible benefits, harms, and trade-offs, it is extremely challenging to foresee all of the relevant, direct, and secondary impacts.

However, it is prudent to review case studies of similar products and the impacts they have had on well-being, as well as to consider possible types of impacts that could apply. Issues to consider include:

- Economic and labor impacts, including labor displacement, unemployment, and inequality,
- Accountability, transparency, and explainability,
- Surveillance, privacy, and civil liberties,
- Fairness, ethics, and human rights,
- Political manipulation, deception, “nudging”, and propaganda,
- Human physical and psychological health,
- Environmental impacts,
- Human dignity, autonomy, and human vs. A/IS roles,
- Security, cybersecurity, and autonomous weapons, and
- Existential risk and super intelligence.

While this is a partial list, it is important to be aware of and reflect on possible and actual cases. For example:

- A prominent concern related to A/IS is of labor displacement and economic and social impacts at an individual and a systems level. A/IS technologies designed to replicate human tasks, behavior, or emotion have the potential to increase or decrease human well-being. These systems could complement human work and increase productivity, wages, and leisure time; or they could be used to supplement and displace human workers, leading to unemployment, inequality, and social strife. It is important for A/IS creators to think about possible uses of their technology and whether they want to encourage or design in restrictions in light of these impacts.
- Another example relates to manipulation. Sophisticated manipulative technologies utilizing A/IS can restrict the fundamental freedom of human choice by manipulating humans who consume content without them recognizing the extent of the manipulation. Software platforms are moving from targeting and customizing content to much more powerful and potentially harmful “persuasive computing” that leverages psychological data and methods. While these approaches may be effective in encouraging use of a product, they may come at significant psychological and social costs.
- A/IS may deceive and harm humans by posing as humans. With the increased ability of artificial systems to meet the Turing test, an intelligence test for a computer that allows a human to distinguish human intelligence from artificial intelligence, there is a significant risk

Well-being

that unscrupulous operators will abuse the technology for unethical commercial or outright criminal purposes. Without taking action to prevent it, it is highly conceivable that A/IS will be used to deceive humans by pretending to be another human being in a plethora of situations and via multiple mediums.

A potential entry point for exploring these unintended consequences is computational sustainability.

[Computational-Sustainability.org](https://www.computational-sustainability.org/) defines the term as an “interdisciplinary field that aims to apply techniques from computer science, information science, operations research, applied mathematics, and statistics for balancing environmental, economic, and societal needs for sustainable development”. [The Institute of Computational Sustainability](#) states that the intent of computational sustainability is provide “computational models for a sustainable environment, economy, and society”. Examples of applied computational sustainability can be seen in the [Stanford University Engineering Department’s course in computational sustainability presentation](#). Computational sustainability technologies designed to increase social good could also be tied to existing well-being metrics.

Recommendation

- To avoid potential negative, unintended consequences, and secure and safeguard positive impacts, A/IS creators, end-users, and stakeholders should be aware of possible

well-being impacts when designing, using, and monitoring A/IS systems. This includes being aware of existing cases and possible areas of impact, measuring impacts on well-being outcomes, and developing regulations to promote beneficent uses of A/IS. Specifically:

- A/IS creators should protect human dignity, autonomy, rights, and well-being of those directly and indirectly affected by the technology. As part of this effort, it is important to include multiple stakeholders, minorities, marginalized groups, and those often without power or a voice in consultation.
- Policymakers, regulators, monitors, and researchers should consider issuing guidance on areas such as A/IS labor and the proper role of humans vs. A/IS in work transparency, trust, and explainability; manipulation and deception; and other areas that emerge.
- Ongoing literature review and analysis should be performed by research and other communities to curate and aggregate information on positive and negative A/IS impacts, along with demonstrated approaches to realize positive ones and ameliorate negative ones.
- A/IS creators working toward computational sustainability should integrate well-being concepts, scientific findings, and indicators into current computational sustainability models. They should work with well-being experts, researchers, and practitioners to conduct research and develop and apply models in A/IS development that prioritize and increase human well-being.

Well-being

- Cross-pollination should be developed between computational sustainability and well-being professionals to ensure integration of well-being into computational sustainability frameworks, and vice versa. Where feasible and reasonable, do the same for conceptual models such as doughnut economics and systems thinking.

Further Resources

- [AI Safety Research](#) by The Future of Life Institute
- D. Helbing, et al. "[Will Democracy Survive Big Data and Artificial Intelligence?](#)" *Scientific American*, February 25, 2017.
- J. L. Schenker, "[Can We Balance Human Ethics with Artificial Intelligence?](#)" *Techonomy*, January 23, 2017.
- M. Bulman, "[EU to Vote on Declaring Robots To Be 'Electronic Persons'.](#)" *Independent*, January 14, 2017.
- N. Nevejan, for the European Parliament. "[European Civil Law Rules in Robotics.](#)" October 2016.
- University of Oxford. "Social media manipulation rising globally, new report warns," <https://phys.org/news/2018-07-social-media-globally.html>. July 20, 2018.
- "[The AI That Pretends To Be Human,](#)" *LessWrong* blog post, February 2, 2016.
- C. Chan, "[Monkeys Grieve When Their Robot Friend Dies.](#)" *Gizmodo*, January 11, 2017.
- Partnership on AI, "AI, Labor, and the Economy" Working Group launches in New York City," <https://www.partnershiponai.org/aile-wg-launch/>. April 25, 2018.
- C.Y. Johnson, "[Children can be swayed by robot peer pressure, study says,](#)" *The Washington Post*, August 15, 2018. [Online]. Available: www.WashingtonPost.com. [Accessed 2018].

Further Resources for Computational Sustainability

- Stanford Engineering Department, [Topics in Computational Sustainability Course Presentation](#), 2016.
- Computational Sustainability, [Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society Project Summary](#).
- C. P. Gomes, "[Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society](#)" in *The Bridge: Linking Engineering and Society*. Washington, DC: National Academy of Engineering of the National Academies, 2009.
- S.J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. "[Computational rationality: A converging paradigm for intelligence in brains, minds, and machines,](#)" *Science* vol. 349, no. 6245, pp. 273–278, July 2015.
- [ACM Fairness, Accountability and Transparency Conference](#)

Well-being

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Well-being Committee

- **John C. Havens** (Co-Chair) – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Laura Musikanski** (Co-Chair) – Executive Director at The Happiness Alliance—home of The Happiness Initiative & Gross National Happiness Index
- **Liz Alexander** – PhD Futurist
- **Anna Alexandrova** – Senior Lecturer in Philosophy of Science at Cambridge University and Fellow of Kings College
- **Christina Berkley** – Executive Coach to leaders in exponential technologies, cutting-edge science, and aerospace
- **Catalina Butnaru** – UK AI Ambassador for global community City.AI, and Founder of HAI, the first methodology for applications of AI in cognitive businesses
- **Celina Beatriz** – Project Director at the Institute for Technology & Society of Rio de Janeiro (ITS Rio)
- **Peet van Biljon** – Founder and CEO at BMNP Strategies LLC, advisor on strategy, innovation, and business transformation;
- Adjunct faculty at Georgetown University; Business ethics author
- **Amy Blankson** – Author of [The Future of Happiness](#) and Founder of TechWell, a research and consulting firm that aims to help organizations to create more positive digital cultures
- **Marc Böhlen** – Professor, University at Buffalo, Emerging Practices in Computational Media. www.realtechsupport.org
- **Rafael A. Calvo** – Professor and ARC Future Fellow at The University of Sydney. Co-author of Positive Computing: Technology for Well-Being and Human Potential
- **Rumman Chowdhury** – PhD Senior Principal, Artificial Intelligence, and Strategic Growth Initiative Responsible AI Lead, Accenture
- **Dr. Aymee Coget** – CEO and Founder of Happiness For HumanKind
- **Danny W. Devriendt** – Managing director of Mediabrands Dynamic (IPG) in Brussels, and the CEO of the Eye of Horus, a global think-tank for communication-technology related topics
- **Eimear Farrell** – Eimear Farrell, independent expert/consultant on technology and human rights (formerly at OHCHR)
- **Danit Gal** – Project Assistant Professor, Keio University; Chair, IEEE Standard P7009 on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems

Well-being

- **Marek Havrda** – PhD Strategy Advisor, GoodAI
- **Andra Keay** – Managing Director of Silicon Valley Robotics, cofounder of Robohub
- **Dr. Peggy Kern** – Senior Lecturer, Centre for Positive Psychology at the University of Melbourne's Graduate School of Education
- **Michael Lennon** – Senior Fellow, Center for Excellence in Public Leadership, George Washington University; Co-Founder, Govpreneur.org; Principal, CAIPP.org (Consortium for Action Intelligence and Positive Performance); Member, [Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems](#) Committee
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of “Artificial Intelligence: Foundations of Computational Agents”
- **Richard Mallah** – Director of AI Project, Future of Life Institute
- **Fabrice Murtin** – Senior Economist, OECD Statistics and Data Directorate
- **Gwen Ottinger** – Associate Professor, Center for Science, Technology, and Society and Department of Politics, Drexel University; Director, [Fair Tech Collective](#)
- **Eleonore Pauwels** – Research Fellow on AI and Emerging Cybertechnologies, United Nations University (NY) and Director of the AI Lab, Woodrow Wilson International Center for Scholars (DC)
- **Venerable Tenzin Priyadarshi** – MIT Media Lab, Director, Ethics Initiative
- **Gideon Rosenblatt** – Writer, focused on work and the human experience in an era of machine intelligence, at [The Vital Edge](#)
- **Daniel Schiff** – PhD Student, Georgia Institute of Technology; Chair, Sub-Group for Autonomous and Intelligent Systems Implementation, [IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems](#)
- **Madalena Sula** – Undergraduate student of Electrical and Computer Engineering, University of Thessaly, Greece, x-PR Manager of IEEE Student Branch of University of Thessaly, Data Scientist & Business Analyst in a multinational company
- **Vincent Siegerink** – Analyst, OECD Statistics and Data Directorate
- **Andy Townsend** – Emerging and Disruptive Technology, PwC UK
- **Andre Uhl** – Research Associate, Director's Office, MIT Media Lab
- **Ramón Villasante** – Founder of [PositiveSocialImpact](#). Software designer, engineer, CTO & CPO in EdTech for sustainable development, social impact and innovation
- **Sarah Villeneuve** – Policy Analyst; Member, [IEEE P7010™ Standards Project for Well-being Metric for Autonomous and Intelligent Systems](#).

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Affective Computing

Affect is a core aspect of intelligence. Drives and emotions, such as excitement and depression, are used to coordinate action throughout intelligent life, even in species that lack a nervous system. Emotions are one mechanism that humans evolved to accomplish what needs to be done in the time available with the information at hand—to satisfy. Emotions are not an impediment to rationality; arguably they are integral to rationality in humans. Humans create and respond to both positive and negative emotional influence as they coordinate their actions with other individuals to create societies. Autonomous and intelligent systems (A/IS) are being designed to simulate emotions in their interactions with humans in ways that will alter our societies.

A/IS should be used to help humanity to the greatest extent possible in as many contexts as are appropriate. While A/IS have tremendous potential to effect positive change, there is also potential that artifacts used in society could cause harm either by amplifying, altering, or even dampening human emotional experience. Even rudimentary versions of synthetic emotions, such as those already in use within nudging systems, have already altered the perception of A/IS by the general public and public policy makers.

This chapter of *Ethically Aligned Design* addresses issues related to emotions and emotion-like control in interactions between humans and design of A/IS. We have put forward recommendations on a variety of topics: considering how affect varies across human cultures; the particular problems of artifacts designed for caring and private relationships; considerations of how intelligent artifacts may be used for “nudging”; how systems can support human flourishing; and appropriate policy interventions for artifacts designed with inbuilt affective systems.

Document Sections

- [Section 1—Systems Across Cultures](#)
- [Section 2—When Systems Care](#)
- [Section 3—System Manipulation/Nudging/Deception](#)
- [Section 4—Systems Supporting Human Potential](#)
- [Section 5—Systems with Synthetic Emotions](#)

Affective Computing

Section 1—Systems Across Cultures

Issue: Should affective systems interact using the norms for verbal and nonverbal communication consistent with the norms of the society in which they are embedded?

Background

Individuals around the world express intentions differently, including the ways that they make eye contact, use gestures, or interpret silence. These particularities are part of an individual's and a society's culture and are incorporated into their affective systems in order to convey the intended message. To ensure that the emotional systems of autonomous and intelligent systems foster effective communication within a specific culture, an understanding of the norms/values of the community where the affective system will be deployed is essential.

Recommendations

1. A well-designed affective system will have a set of essential norms, specific to its intended cultural context of use, in its knowledge base. Research has shown that A/IS technologies can use at least five types of cues to simulate social interactions.
2. These include: physical cues such as simulated facial expressions, psychological cues such as simulated humor or other emotions, use of language, use of social dynamics like taking turns, and through social roles such as acting as a tutor or medical advisor. Further examples are listed below:
 - a. Well-designed affective systems will use language with affective content carefully and within the contemporaneous expectations of the culture. An example is small talk. Although small talk is useful for establishing a friendly rapport in many communities, some communities see people that use small talk as insincere and hypocritical. Other cultures may consider people that do not use small talk as unfriendly, uncooperative, rude, arrogant, or ignorant. Additionally, speaking with proper vocabulary, grammar, and sentence structure may contrast with the typical informal interactions between individuals. For example, the latest trend, TV show, or other media may significantly influence what is viewed as appropriate vocabulary and interaction style.
 - b. Well-designed affective systems will recognize that the amount of personal space (proxemics) given by individuals in an important part of culturally specific

Affective Computing

human interaction. People from varying cultures maintain, often unknowingly, different spatial distances between themselves to establish smooth communication. Crossing these limits may require explicit or implicit consent, which A/IS must learn to negotiate to avoid transmitting unintended messages.

- c. Eye contact is an essential component for culturally sensitive social interaction. For some interactions, direct eye contact is needed but for others it is not essential and may even generate misunderstandings. It is important that A/IS be equipped to recognize the role of eye contact in the development of emotional interaction.
- d. Hand gestures and other non-verbal communication are very important for social interaction. Communicative gestures are culturally specific and thus should be used with caution in cross-cultural situations. The specificity of physical communication techniques must be acknowledged in the design of functional affective systems. For instance, although a “thumbs-up” sign is commonly used to indicate approval, in some countries this gesture can be considered an insult.
- e. Humans use facial expressions to detect emotions and facilitate communication. Facial expressions may not be universal across cultures, however, and A/IS trained with a dataset from one culture may not be readily usable in another

culture. Well-developed A/IS will be able to recognize, analyze, and even display facial expressions essential for culturally specific social interaction.

3. Engineers should consider the need for cross-cultural use of affective systems. Well-designed systems will have options innate to facilitate flexibility in cultural programming. Mechanisms to enable and disable culturally specific “add-ons” should be considered an essential part of A/IS development.

Further Resources

- G. Cotton, “[Gestures to Avoid in Cross-Cultural Business: In Other Words, ‘Keep Your Fingers to Yourself!’](#)” *Huffington Post*, June 13, 2013.
- “[Paralanguage Across Cultures](#),” Sydney, Australia: Culture Plus Consulting, 2016.
- G. Cotton, [Say Anything to Anyone, Say Anything to Anyone, Anywhere: 5 Keys to Successful Cross-Cultural Communication](#). Hoboken, NJ: Wiley, 2013.
- D. Elmer, [Cross-Cultural Connections: Stepping Out and Fitting In Around the World](#). Westmont, IL: InterVarsity Press, 2002.
- B. J. Fogg, [Persuasive Technology](#). *Ubiquity*, December 2, 2002.
- A. McStay, *Emotional AI: The Rise of Empathic Media*. London: Sage, 2018.
- M. Price, “[Facial Expressions—Including Fear—May Not Be as Universal as We Thought](#).” *Science*, October 17, 2016.

Affective Computing

Issue: It is presently unknown whether long-term interaction with affective artifacts that lack cultural sensitivity could alter human social interaction.

Background

Systems that do not have cultural knowledge incorporated into their knowledge base may or may not interact effectively with humans for whom emotion and culture are significant. Given that interaction with A/IS may affect individuals and societies, it is imperative that we carefully evaluate mechanisms to promote beneficial affective interaction between humans and A/IS. Humans often use mirroring in order to understand and develop their norms for behavior. Certain machine learning approaches also address improving A/IS interaction with humans through mirroring human behavior. Thus, we must remember that learning via mirroring can go in both directions and that interacting with machines has the potential to impact individuals' norms, as well as societal and cultural norms. If affective artifacts with enhanced, different, or absent cultural sensitivity interact with impressionable humans this could alter their responses to social and cultural cues and values. The potential for A/IS to exert cultural influence in powerful ways, at scale, is an area of substantial concern.

Recommendations

1. Collaborative research teams must research the effects of long-term interaction of people with affective systems. This should be done using multiple protocols, disciplinary approaches, and metrics to measure the modifications of habits, norms, and principles as well as careful evaluation of the downstream cultural and societal impacts.
2. Parties responsible for deploying affective systems into the lives of individuals or communities should be trained to detect the influence of A/IS, and to utilize mitigation techniques if A/IS effects appear to be harmful. It should always be possible to shut down harmful A/IS.

Further Resources

- T. Nishida and C. Faucher, Eds., [Modelling Machine Emotions for Realizing Intelligence: Foundations and Applications](#). Berlin, Germany: Springer-Verlag, 2010.
- D. J. Pauleen, et al. "Cultural Bias in Information Systems Research and Practice: Are You Coming from the Same Place I Am?" *Communications of the Association for Information Systems*, vol. 17, pp. 1–36, 2006. J. Bielby, "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2015.
- J. Bryson, ["Why Robot Nannies Probably Won't Do Much Psychological Damage."](#) A commentary on an article by N. Sharkey

Affective Computing

and A. Sharkey, *The Crying Shame of Robot Nannies*. *Interaction Studies*, vol. 11, no. 2 pp. 161–190, July 2010.

- A. Sharkey, and N. Sharkey, "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine*, vol.18, no. 1, pp. 32–38, March 2011.

Issue: When affective systems are deployed across cultures, they could adversely affect the cultural, social, or religious values of the community in which they interact.

Background

Some philosophers argue that there are no universal ethical principles and that ethical norms vary from society to society. Regardless of whether universalism or some form of ethical relativism is true, affective systems need to respect the values of the cultures within which they are embedded. How systems should effectively reflect the values of the designers or the users of affective systems is not a settled discussion. There is general agreement that developers of affective systems should acknowledge that the systems should reflect the values of those with whom the systems are interacting. There is a high likelihood that when spanning different groups, the values imbued by the developer will be different from the operator or customer of that affective system, and that

end-user values should be actively considered. Differences between affective systems and societal values may generate conflict situations producing undesirable results, e.g., gestures or eye contact being misunderstood as rude or threatening. Thus, affective systems should adapt to reflect the values of the community and individuals where they will operate in order to avoid misunderstanding.

Recommendations

Assuming that well-designed affective systems have a minimum subset of configurable norms incorporated in their knowledge base:

1. Affective systems should have capabilities to identify differences between the values they are designed with and the differing values of those with whom the systems are interacting.
2. Where appropriate, affective systems will adapt accordingly over time to better fit the norms of their users. As societal values change, there needs to be a means to detect and accommodate such cultural change in affective systems.
3. Those actions undertaken by an affective system that are most likely to generate an emotional response should be designed to be easily changed in appropriate ways by the user without being easily hacked by actors with malicious intentions. Similar to how software today externalizes the language and vocabulary to be easily changeable based on location, affective systems should externalize some of the core aspects of their actions.

Affective Computing

Further Resources

- J. Bielby, "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2015.
- M. Velasquez, C. Andre, T. Shanks, and M. J. Meyer. "[Ethical Relativism](#)." Markkula Center for Applied Ethics, Santa Clara, CA: Santa Clara University, August 1, 1992.
- Culture reflects the moral values and ethical norms governing how people should behave and interact with others. "[Ethics, an Overview](#)." Boundless Management.
- T. Donaldson, "[Values in Tension: Ethics Away from Home Away from Home](#)." *Harvard Business Review*. September– October 1996.

Section 2—When Systems Care

Issue: Are moral and ethical boundaries crossed when the design of affective systems allows them to develop intimate relationships with their users?

Background

There are many robots in development or production designed to focus on intimate care of children, adults, and the elderly². While robots capable of participating fully in intimate relationships are not currently available, the potential use of such robots routinely captures the attention of the media. It is important that professional communities, policy makers, and the general public participate in development of guidelines for appropriate use of A/IS in this area. Those guidelines should acknowledge

fundamental human rights to highlight potential ethical benefits and risks that may emerge, if and when affective systems interact intimately with users.

Among the many areas of concern are the representation of care, embodiment of caring A/IS, and the sensitivity of data generated through intimate and caring relationships with A/IS. The literature suggests that there are some potential benefits to individuals and to society from the incorporation of caring A/IS, along with duly cautionary notes concerning the possibility that these systems could negatively impact human-to-human intimate relations³.

Recommendations

As this technology develops, it is important to monitor research into the development of intimate relationships between A/IS and humans. Research should emphasize any technical and

Affective Computing

normative developments that reflect use of A/IS in positive and therapeutic ways while also creating appropriate safeguards to mitigate against uses that contribute to problematic individual or social relationships:

1. Intimate systems must not be designed or deployed in ways that contribute to stereotypes, gender or racial inequality, or the exacerbation of human misery.
2. Intimate systems must not be designed to explicitly engage in the psychological manipulation of the users of these systems unless the user is made aware they are being manipulated and consents to this behavior. Any manipulation should be governed through an opt-in system.
3. Caring A/IS should be designed to avoid contributing to user isolation from society.
4. Designers of affective robotics must publicly acknowledge, for example, within a notice associated with the product, that these systems can have side effects, such as interfering with the relationship dynamics between human partners, causing attachments between the user and the A/IS that are distinct from human partnership.
5. Commercially marketed A/IS for caring applications should not be presented to be a person in a legal sense, nor marketed as a person. Rather its artifactual, that is, authored, designed, and built deliberately, nature should always be made as transparent as possible, at least at point of sale and in available documentation, as noted in Section 4, Systems Supporting Human Potential.
6. Existing laws regarding personal imagery need to be reconsidered in light of caring A/IS. In addition to other ethical considerations, it will also be necessary to establish conformance with local laws and mores in the context of caring A/IS systems.

Further Resources

- M. Boden, J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden and T. Sorrell, Principles of robotics: regulating robots in the real world. *Connection Science*, vol. 29, no. 2, pp. 124-129, April 2017.
- J. J. Bryson, M. E. Diamantis, and T. D. Grant, "Of, For, and By the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence & Law*, vol. 25, no. 3, pp. 273–291, Sept. 2017.
- M. Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," in *Robot Ethics: The Ethical and Social Implications of Robotics*, P. Lin, K. Abney, and G. Bekey, Eds., pp. 205. Cambridge, MA: MIT Press, 2011.

Affective Computing

Section 3— System Manipulation/ Nudging/Deception

Issue: Should affective systems be designed to nudge people for the user's personal benefit and/or for the benefit of others?

Background

Manipulation can be defined as an exercise of influence by one person or group, with the intention to attempt to control or modify the actions of another person or group. Thaler and Sunstein (2008) call the tactic of subtly modifying behavior a “nudge⁴”. Nudging mainly operates through the affective elements of a human rational system. Making use of a nudge might be considered appropriate in situations like teaching children, treating drug dependency, and in some healthcare settings. While nudges can be deployed to encourage individuals to express behaviors that have community benefits, a nudge could have unanticipated consequences for people whose backgrounds were not well considered in the development of the nudging system⁵. Likewise, nudges may encourage behaviors with unanticipated long-term effects, whether positive or negative, for the individual and/or society. The effect of A/IS nudging a person, such as potentially eroding or encouraging individual liberty, or expressing behaviors that are for the benefit others, should be well characterized in the design of A/IS.

Recommendations

1. Systematic analyses are needed that examine the ethics and behavioral consequences of designing affective systems to nudge human beings prior to deployment.
2. The user should be empowered, through an explicit opt-in system and readily available, comprehensible information, to recognize different types of A/IS nudges, regardless of whether they seek to promote beneficial social manipulation or to enhance consumer acceptance of commercial goals. The user should be able to access and check facts behind the nudges and then make a conscious decision to accept or reject a nudge. Nudging systems must be transparent, with a clear chain of accountability that includes human agents: data logging is required so users can know how, why, and by whom they were nudged.
3. A/IS nudging must not become coercive and should always have an opt-in system policy with explicit consent.
4. Additional protections against unwanted nudging must be put in place for vulnerable populations, such as children, or when informed consent cannot be obtained. Protections against unwanted nudging should be encouraged when nudges alter long-term behavior or when consent alone may not be a sufficient safeguard against coercion or exploitation.

Affective Computing

5. Data gathered which could reveal an individual or groups' susceptibility to a nudge or their emotional reaction to a nudge should not be collected or distributed without opt-in consent, and should only be retained transparently, with access restrictions in compliance with the highest requirements of data privacy and law.

Further Resources

- R. Thaler, and C. R. Sunstein, *Nudge: Improving Decision about Health, Wealth and Happiness*, New Haven, CT: Yale University Press, 2008.
- L. Bovens, "The Ethics of Nudge," in *Preference change: Approaches from Philosophy, Economics and Psychology*, T. Grüne-Yanoff and S. O. Hansson, Eds., Berlin, Germany: Springer, 2008 pp. 207–219.
- S. D. Hunt and S. Vitell. "A General Theory of Marketing Ethics." *Journal of Macromarketing*, vol.6, no. 1, pp. 5-16, June 1986.
- A. McStay, [Empathic Media and Advertising: Industry, Policy, Legal and Citizen Perspectives \(the Case for Intimacy\)](#), Big Data & Society, pp. 1-11, December 2016.
- J. de Quintana Medina and P. Hermida Justo, "Not All Nudges Are Automatic: Freedom of Choice and Informative Nudges." Working paper presented to the European Consortium for Political Research, Joint Session of Workshops, 2016 Behavioral Change and Public Policy, Pisa, Italy, 2016.
- M. D. White, [The Manipulation of Choice. Ethics and Libertarian Paternalism](#). New York: Palgrave Macmillan, 2013
- C.R. Sunstein, *The Ethics of Influence: Government in the Age of Behavioral Science*. New York: Cambridge, 2016
- M. Scheutz, "[The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?](#)" *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 424–433, Sept. 2012.
- A. Grinbaum, R. Chatila, L. Devillers, J.-G. Ganascia, C. Tessier and M. Dauchet. "[Ethics in Robotics Research: CERNA Recommendations](#)," *IEEE Robotics and Automation Magazine*, vol. 24, no. 3, pp. 139–145, Sept. 2017.
- "Designing Moral Technologies: Theoretical, Practical, and Ethical Issues" Conference July 10–15, 2016, Monte Verità, Switzerland.

Affective Computing

Issue: Governmental entities may potentially use nudging strategies, for example to promote the performance of charitable acts. Does the practice of nudging for the benefit of society, including nudges by affective systems, raise ethical concerns?

Background

A few scholars have noted a potentially controversial practice of the future: allowing a robot or another affective system to nudge a user for the good of society⁶. For instance, if it is possible that a well-designed robot could effectively encourage humans to perform charitable acts, would it be ethically appropriate for the robot to do so? This design possibility illustrates just one behavioral outcome that a robot could potentially elicit from a user.

Given the persuasive power that an affective system may have over a user, ethical concerns related to nudging must be examined. This includes the significant potential for misuse.

Recommendations

1. As more and more computing devices subtly and overtly influence human behavior, it is important to draw attention to whether it is ethically appropriate to pursue this type of design pathway in the context of governmental actions.
2. There needs to be transparency regarding who the intended beneficiaries are, and whether any form of deception or manipulation is going to be used to accomplish the intended goal.

Further Resources

- J. Borenstein and R. Arkin, "[Robotic Nudges: Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being Just Human Being.](#)" *Science and Engineering Ethics*, vol. 22, no. 1, pp. 31–46, Feb. 2016.
- J. Borenstein and R. Arkin. "[Nudging for Good: Robots and the Ethical Appropriateness of Nurturing Empathy and Charitable Behavior.](#)" *AI and Society*, vol. 32, no. 4, pp. 499–507, Nov. 2016.

Issue: Will A/IS nudging systems that are not fully relevant to the sociotechnical context in which they are operating cause behaviors with adverse unintended consequences?

Background

A well-designed nudging or suggestion system will have sophisticated enough technical capabilities for recognizing the context in which it is applying nudging actions. Assessment of the context requires perception of the scope or impact of the actions to be taken, the consequences of incorrectly or incompletely

Affective Computing

applied nudges, and acknowledgement of the uncertainties that may stem from long term consequences of a nudge⁷.

Recommendations

1. Consideration should be given to the development of a system of technical licensing ("permits") or other certification from governments or non-governmental organizations (NGOs) that can aid users to understand the nudges from A/IS in their lives.
2. User autonomy is a key and essential consideration that must be taken into account when addressing whether affective systems should be permitted to nudge human beings.
3. Design features of an affective system that nudges human beings should include the ability to accurately distinguish between users, including detecting characteristics such as whether the user is an adult or a child.
4. Affective systems with nudging strategies should incorporate a design system of evaluation, monitoring, and control for unintended consequences.

Further Resources

- J. Borenstein and R. Arkin, "[Robotic Nudges: Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being Just Human Being](#)," *Science and Engineering Ethics*, vol. 22, no. 1, pp. 31–46, 2016.
- R. C. Arkin, M. Fujita, T. Takagi, and R. Hasegawa, "[An Ethological and Emotional Basis for Human- Robot Interaction](#)," *Robotics and Autonomous Systems*, vol. 42, no. 3–4 pp.191–201, March 2003.
- S. Omohundro "[Autonomous Technology and the Greater Human Good](#)," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 26, no. 3, pp. 303–315, 2014.

Issue: When, if ever, and under which circumstances, is deception performed by affective systems acceptable?

Background

Deception is commonplace in everyday human-human interaction. According to Kantian ethics, it is never ethically appropriate to lie, while utilitarian frameworks indicate that it can be acceptable when deception increases overall happiness. Given the diversity of views on ethics and the appropriateness of deception, should affective systems be designed to deceive? Does the non-consensual nature of deception restrict the use of A/IS in contexts in which deception may be required?

Affective Computing

Recommendations

It is necessary to develop recommendations regarding the acceptability of deception performed by A/IS, specifically with respect to when and under which circumstances, if any, it is appropriate.

1. In general, deception may be acceptable in an affective agent when it is used for the benefit of the person being deceived, not for the agent itself. For example, deception might be necessary in search and rescue operations or for elder- or child-care.
2. For deception to be used under any circumstance, a logical and reasonable justification must be provided by the designer, and this rationale should be certified by an external authority, such as a licensing body or regulatory agency.

Further Resources

- R. C. Arkin, "Robots That Need to Mislead: Biologically-inspired Machine Deception." *IEEE Intelligent Systems* 27, no. 6, pp. 60–75, 2012.
- J. Shim and R. C. Arkin, "Other-Oriented Robot Deception: How Can a Robot's Deceptive Feedback Help Humans in HRI?" *Eighth International Conference on Social Robotics (ICSR 2016)*, Kansas, MO., November 2016.
- J. Shim and R. C. Arkin, "The Benefits of Robot Deception in Search and Rescue: Computational Approach for Deceptive Action Selection via Case-based Reasoning." *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR 2015)*, West Lafayette, IN, October 2015.
- J. Shim and R. C. Arkin, "A Taxonomy of Robot Deception and its Benefits in HRI." *Proceedings of IEEE Systems, Man and Cybernetics Conference*, Manchester England, October 2013.

Affective Computing

Section 4—Systems Supporting Human Potential

Issue: Will extensive use of A/IS in society make our organizations more brittle by reducing human autonomy within organizations, and by replacing creative, affective, empathetic components of management chains?

Background

If human workers are replaced by A/IS, the possibility of corporations, governments, employees, and customers discovering new equilibria outside the scope of what the organizations' past leadership originally foresaw may be unduly limited. A lack of empathy based on shared needs, abilities, and disadvantages between organizations and customers causes disequilibria between the individuals and corporations and governments that exist to serve them. Opportunities for useful innovation may therefore be lost through automation. Collaboration requires enough commonality of collaborating intelligences to create empathy—the capacity to model the other's goals based on one's own.

According to scientists within several fields, autonomy is a psychological need. Without it, humans fail to thrive, create, and innovate.

Ethically aligned design should support, not hinder, human autonomy or its expression.

Recommendations

1. It is important that human workers' interaction with other workers not always be intermediated by affective systems (or other technology) which may filter out autonomy, innovation, and communication.
2. Human points of contact should remain available to customers and other organizations when using A/IS.
3. Affective systems should be designed to support human autonomy, sense of competence, and meaningful relationships as these are necessary to support a flourishing life.
4. Even where A/IS are less expensive, more predictable, and easier to control than human employees, a core network of human employees should be maintained at every level of decision-making in order to ensure preservation of human autonomy, communication, and innovation.
5. Management and organizational theorists should consider appropriate use of affective and autonomous systems to enhance their business models and the efficacy of their workforce within the limits of the preservation of human autonomy.

Affective Computing

Further Resources

- J. J. Bryson, "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems*, C. Misselhorn, Ed., pp. 281–306, Springer, 2015.
- D. Peters, R.A. Calvo, and R.M. Ryan, "[Designing for Motivation, Engagement and Wellbeing in Digital Experience](#)," *Frontiers in Psychology—Human Media Interaction*, vol. 9, pp 797, 2018.

Issue: Does the increased access to personal information about other members of our society, facilitated by A/IS, alter the human affective experience? Does this access potentially lead to a change in human autonomy?

Background

Theoretical biology tells us that we should expect increased communication—which A/IS facilitate—to increase group-level investment⁸. Extensive use of A/IS could change the expression of individual autonomy and in its place increase group-based identities. Examples of this sort of social alteration may include:

1. Changes in the scope of monitoring and control of children's lives by parents.
2. Decreased willingness to express opinions for fear of surveillance or long-term consequences of past expressions being used in changed temporal contexts.

3. Utilization of customers or other end users to perform basic corporate business processes such as data entry as a barter for lower prices or access, resulting potentially in reduced tax revenues.
4. Changes to the expression of individual autonomy could alter the diversity, creativity, and cohesiveness of a society. It may also alter perceptions of privacy and security, and social and legal liability for autonomous expressions.

Recommendations

1. Organizations, including governments, must put a high value on individuals' privacy and autonomy, including restricting the amount and age of data held about individuals specifically.
2. Education in all forms should encourage individuation, the preservation of autonomy, and knowledge of the appropriate uses and limits to A/IS⁹.

Further Resources

- J. J. Bryson, "Artificial Intelligence and Pro-Social Behavior," in *Collective Agency and Cooperation in Natural and Artificial Systems*, C. Misselhorn, Ed., pp. 281–306, Springer, 2015.
- M. Cooke, "A Space of One's Own: Autonomy, Privacy, Liberty," *Philosophy & Social Criticism*, Vol. 25, no. 1, pp. 22–53, 1999.
- D. Peters, R.A. Calvo, R.M. Ryan, "Designing for Motivation, Engagement and Wellbeing in Digital Experience" *Frontiers in Psychology – Human Media Interaction*, vol. 9, pp 797, 2018.

Affective Computing

- J. Roughgarden, M. Oishi and E. Akçay, "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection." *Science* 311, no. 5763, pp. 965–969, 2006.

Issue: Will use of A/IS adversely affect human psychological and emotional well-being in ways not otherwise foreseen?

Background

A/IS may be given unprecedented access to human culture and human spaces—both physical and intellectual. A/IS may communicate via natural language, may move with humanlike form, and may express humanlike identity, but they are not, and should not be regarded as, human. Incorporation of A/IS into daily life may affect human well-being in ways not yet anticipated. Incorporation of A/IS may alter patterns of trust and capability assessment between humans, and between humans and A/IS.

Recommendations

1. Vigilance and robust, interdisciplinary, on-going research on identifying situations where A/IS affect human well-being, both positively and negatively, is necessary. Evidence of correlations between the increased use of A/IS and positive or negative individual or social outcomes must be explored.
2. Design restrictions should be placed on the systems themselves to avoid machine decisions that may alter a person's life in unknown ways. Explanations should be available on demand in systems that may affect human well-being.

Further Resources

- K. Kamewari, M. Kato, T. Kanda, H. Ishiguro and K. Hiraki. "Six-and-a-Half-Month-Old Children Positively Attribute Goals to Human Action and to Humanoid-Robot Motion," *Cognitive Development*, vol. 20, no. 2, pp. 303–320, 2005.
- R.A. Calvo and D. Peters, *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA: MIT Press, 2014.

Affective Computing

Section 5—Systems with Synthetic Emotions

Issue: Will deployment of synthetic emotions into affective systems increase the accessibility of A/IS? Will increased accessibility prompt unforeseen patterns of identification with A/IS?

Background

Deliberately constructed emotions are designed to create empathy between humans and artifacts, which may be useful or even essential for human-A/IS collaboration. Synthetic emotions are essential for humans to collaborate with the A/IS but can also lead to failure to recognize that synthetic emotions can be compartmentalized and even entirely removed. Potential consequences for humans include different patterns of bonding, guilt, and trust, whether between the human and A/IS or between other humans. There is no coherent sense in which A/IS can be made to suffer emotional loss, because any such affect, even if possible, could be avoided at the stage of engineering, or reengineered. As such, it is not possible to allocate moral agency or responsibility in the senses that have been developed for human emotional bonding and thus sociality.

Recommendations

1. Commercially marketed A/IS should not be persons in a legal sense, nor marketed as persons. Rather their artifactual (authored, designed, and built deliberately) nature should always be made as transparent as possible, at least at point of sale and in available documentation.
2. Some systems will, due to their application, require opaqueness in some contexts, e.g., emotional therapy. Transparency in such systems should be available to inspection by responsible parties but may be withdrawn for operational needs.

Further Resources

- R. C. Arkin, P. Ulam and A. R. Wagner, "Moral Decision-making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust and Deception," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571–589, 2012.
- R. Arkin, M. Fujita, T. Takagi and R. Hasegawa. "An Ethological and Emotional Basis for Human-Robot Interaction," *Robotics and Autonomous Systems*, vol.42, no. 3–4, pp.191–201, 2003.
- R. C. Arkin, "Moving up the Food Chain: Motivation and Emotion in Behavior-based Robots," in *Who Needs Emotions: The Brain Meets the Robot*, J. Fellous and M. Arbib., Eds., New York: Oxford University Press, 2005.

Affective Computing

- M. Boden, J. Bryson, D. Caldwell, et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science*, vol. 29, no. 2, pp. 124–129, 2017.
- J. J. Bryson, M. E. Diamantis and T. D. Grant. "Of, For, and By the People: The Legal Lacuna of Synthetic Persons," *Artificial Intelligence & Law*, vol. 25, no. 3, pp. 273–291, Sept. 2017.
- J. Novikova, and L. Watts, "Towards Artificial Emotions to Assist Social Coordination in HRI," *International Journal of Social Robotics*, vol. 7, no. 1, pp. 77–88, 2015.
- M. Scheutz, "The Affect Dilemma for Artificial Agents: Should We Develop Affective Artificial Agents?" *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 424–433, 2012.
- A. Sharkey and N. Sharkey. "Children, the Elderly, and Interactive Robots." *IEEE Robotics & Automation Magazine*, vol. 18, no. 1, pp. 32–38, 2011.

Affective Computing

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Affective Computing Committee

- **Ronald C. Arkin** (Founding Co-Chair) – Regents' Professor & Director of the Mobile Robot Laboratory; College of Computing Georgia Institute of Technology
- **Joanna J. Bryson** (Co-Chair) – Reader (Associate Professor), University of Bath, Intelligent Systems Research Group, Department of Computer Science
- **John P. Sullins** (Co-Chair) – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Genevieve Bell** – Intel Senior Fellow Vice President, Corporate Strategy Office Corporate Sensing and Insights
- **Jason Borenstein** – Director of Graduate Research Ethics Programs, School of Public Policy and Office of Graduate Studies, Georgia Institute of Technology
- **Cynthia Breazeal** – Associate Professor of Media Arts and Sciences, MIT Media Lab; Founder & Chief Scientist of Jibo, Inc.
- **Joost Broekens** – Assistant Professor Affective Computing, Interactive Intelligence group; Department of Intelligent Systems, Delft University of Technology
- **Rafael Calvo** – Professor & ARC Future Fellow, School of Electrical and Information Engineering, The University of Sydney
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions' - member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Jonathan Gratch** – Research Professor of Computer Science and Psychology, Director for Virtual Human Research, USC Institute for Creative Technologies
- **Mark Halverson** – Founder and CEO at Human Ecology Holdings and Precision Autonomy
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*

Affective Computing

- **Noreen Herzfeld** – Reuter Professor of Science and Religion, St. John’s University
- **Chihyung Jeon** – Assistant Professor, Graduate School of Science and Technology Policy, Korea Advanced Institute of Science and Technology (KAIST)
- **Preeti Mohan** – Software Engineer at Microsoft and Computational Linguistics Master’s Student at the University of Washington
- **Bjoern Niehaves** – Professor, Chair of Information Systems, University of Siegen
- **Rosalind Picard** – Rosalind Picard, (Sc.D, FIEEE) Professor, MIT Media Laboratory, Director of Affective Computing Research; Faculty Chair, MIT Mind+Hand+Heart; Co-founder & Chief Scientist, Empatica Inc.; Co-founder, Affectiva Inc.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow
- **Matthias Scheutz** – Professor, Bernard M. Gordon Senior Faculty Fellow, Tufts University School of Engineering
- **Robert Sparrow** – Professor, Monash University, Australian Research Council “Future Fellow”, 2010-15.
- **Cherry Tom** – Emerging Technologies Intelligence Manager, IEEE Standards Association

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Affective Computing

Endnotes

¹ See B. J. Fogg, [Persuasive technology](#). *Ubiquity*, December: 2, 2002.

² See S. Turkle, W. Taggart, C.D. Kidd, and O. Daste, "Relational artifacts with children and elders: the complexities of cybercompanionship, *Connection Science*, vol. 18, no. 4, 2006.

³ A discussion of intimate robots for therapeutic and personal use is outside of the scope of *Ethically Aligned Design, First Edition*. For further treatment, among others, see J. P. Sullins, "Robots, Love, and Sex: The Ethics of Building a Love Machine." *IEEE Transactions on Affective Computing* 3, no. 4 (2012): 398–409.

⁴ See R. Thaler, and C. R. Sunstein. *Nudge: Improving Decision about Health, Wealth and Happiness*, New Haven, CT: Yale University Press, 2008.

⁵ See J. de Quintana Medina and P. Hermida Justo. "[Not All Nudges Are Automatic: Freedom of Choice and Informative Nudges](#)." Working paper presented to the European Consortium for Political Research, Joint Session of Workshops, 2016 Behavioral Change and Public Policy, Pisa, Italy, 2016; and M. D. White, [The Manipulation of Choice. Ethics and Libertarian Paternalism](#). New York: Palgrave Macmillan, 2013.

⁶ See, for example, J. Borenstein and R. Arkin. "[Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being](#)." *Science and Engineering Ethics*, vol. 22, no. 1 (2016): 31–46.

⁷ See S. Omohundro, "[Autonomous Technology and the Greater Human Good](#)." *Journal of Experimental and Theoretical Artificial Intelligence* 26, no. 3 (2014): 303–315.

⁸ See J. Roughgarden, M. Oishi, and E. Akçay. "Reproductive Social Behavior: Cooperative Games to Replace Sexual Selection." *Science* 311, no. 5763 (2006): 965–969.

⁹ See the Well-being chapter of this *Ethically Aligned Design, First Edition*.

Personal Data and Individual Agency

Regulations like the [General Data Protection Regulation](#) (GDPR) and the [California Consumer Privacy Act](#) (CCPA) of 2018 are helping to improve personal data protection. But legal compliance is not enough to mitigate the ethical implications and core challenges to human agency embodied by algorithmically driven behavioral tracking or persuasive computing. The core of the issue is one of parity.

Humans cannot respond on an individual basis to every algorithm tracking their behavior without technological tools supported by policy allowing them to do so. Individuals may provide consent without fully understanding specific terms and conditions agreements. But they are also not equipped to fully recognize how the nuanced use of their data to inform personalized algorithms affects their choices at the risk of eroding their agency.

Here we understand agency as an individual's ability to influence and shape their life trajectory as determined by their cultural and social contexts. Agency in the digital arena enables an individual to make informed decisions where their own terms and conditions can be recognized and honored at an algorithmic level.

To strengthen individual agency, governments and organizations must test and implement technologies and policies that let individuals create, curate, and control their online agency as associated with their identity. Data transactions should be moderated and case-by-case authorization decisions from the individual as to who can process what personal data for what purpose.

Specifically, we recommend governments and organizations:

- **Create:** Provide every individual with the means to create and project their own terms and conditions regarding their personal data that can be read and agreed to at a machine-readable level.
- **Curate:** Provide every individual with a personal data or algorithmic agent which they curate to represent their terms and conditions in any real, digital, or virtual environment.
- **Control:** Provide every individual access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data.

Three sections of this chapter reflect these core ideals regarding human agency.

A fourth section addresses issues surrounding personal data and individual agency relating to children.

Personal Data and Individual Agency

Section 1—Create

To retain agency in the algorithmic era, each individual must have the means to create and project their own terms and conditions regarding their personal data. These must be readable and usable by both humans and machines.

Issue: What would it mean for a person to have individually controlled terms and conditions for their personal data?

Background

Part of providing individually controlled terms and conditions for personal data is to help each person consider what their preferences are regarding their data versus dictating how they need to share it. While questions along these lines are framed in light of a person's privacy, their preferences also reveal larger values for individuals. The ethical issue is whether A/IS act in accordance with these values.

This process of investigating one's values to identify these preferences is a powerful step towards regaining data agency. The point is not only that a person's data are protected, but also that by curating these answers they become educated about how important their information is in the context of how it is shared.

Most individuals also believe controlling their personal data only happens on the sites or social networks to which they belong and have no idea of the consequences of how that data may be used by others in the future. Agreeing to most standard terms and conditions on these sites largely means users consent to give up control of their personal data rather than play a meaningful role in defining and curating its downstream use.

The scope of how long one should or could control the downstream use of their data can be difficult to calculate as consent-based models of personal data have trained users to release rights on any claims for use of their data which are entirely provided to the service, manufacturer, and their partners. However, models like YouTube's [Content ID](#) provide a form of precedent for thinking about how an individual's data could be technically protected where it is considered as an asset they could control and copyright. Here is language from [YouTube's site about the service](#): "Copyright owners can use a system called Content ID to easily identify and manage their content on YouTube. Videos uploaded to YouTube are scanned against a database of files that have been submitted to us by content owners." In this sense, the question of how long or how far downstream one's personal data should be protected takes on the same logic of how long a corporation's intellectual property or copyrights could be protected based on initial legal terms set.

Personal Data and Individual Agency

One challenge is how to define use of data that can affect the individual directly, versus use of aggregated data. For example, an individual subway user's travel card, tracking their individual movements, should be protected from uses that identify or profile that individual to make inferences about his/her likes or location generally. But data provided by a user could be included in an overall travel system's management database, aggregated into patterns for scheduling and maintenance as long as the individual-level data are deleted. Where users have predetermined via their terms and conditions that they are willing to share their data for these travel systems, they can meaningfully articulate how to share their information.

Under current business models, it is common for people to consent to the sharing of discrete data like credit card transaction data, answers to test questions, or how many steps they walk. However, once aggregated these data and the associated insights may lead to complex and sensitive conclusions being drawn about individuals. This end use of the individual's data may not have been part of the initial sharing agreement. This is why models for terms and conditions created for user control typically alert people via onscreen or other warning methods when their predetermined preferences are not being honored.

Recommendation

Individuals should be provided tools that produce machine-readable terms and conditions that are dynamic in nature and serve to protect their data and honor their preferences for its use.

Specifically:

- Personal data access and consent should be managed by the individual using their curated terms and conditions that provide notification and an opportunity for consent at the time data are exchanged, versus outside actors being able to access personal data without an individual's awareness or control.
- Terms should be presented in a way that allows a user to easily read, interpret, understand, and choose to engage with any A/IS. Consent should be both conditional and dynamic, where "dynamic" means downstream uses of a person's data must be explicitly called out, allowing them to cancel a service and potentially rescind or "kill" any data they have shared with a service to date via the use of a "Smart Contract" or specific conditions as described in mutual terms and conditions between two parties at the time of exchange.
- For further information on these issues, please see the following section in regard to algorithmic agents and their application.

Further Resources

- [IEEE P7012™ - IEEE Standards Project for Machine Readable Personal Privacy Terms](#). This approved standardization project (currently in development) directly honors the goals laid out in Section One of this document.
- [The Personalized Privacy Assistant Project](#) Carnegie Mellon University. <https://privacyassistant.org>, 2019.

Personal Data and Individual Agency

- M. Orcutt, "[Personal AI Privacy Watchdog Could Help You Regain Control of Your Data](#)" MIT Technology Review, May 11, 2017.
- M. Hintze, [Privacy Statements: Purposes, Requirements, and Best Practices](#), Cambridge, U.K.: Cambridge University Press, 2017.
- D. J. Solove, "Privacy self-management and the consent dilemma, Harvard Law Review, vol. 126, no. 7, pp. 1880–1903, May 2013.
- N. Sadeh, M. Degeling, A. Das, A. S. Zhang, A. Acquisti, L. Bauer, L. Cranor, A. Datta, and D. Smullen, A Privacy Assistant for the Internet of Things: https://www.usenix.org/sites/default/files/soups17_poster_sadeh.pdf
- H. Lee, R. Chow, M. R. Haghighat, H. M. Patterson and A. Kobsa, "IoT Service Store: A Web-based System for Privacy-aware IoT Service Discovery and Interaction," *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Athens, pp. 107-112, 2018.
- L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," W3C Recommendation, [Online]. Available: www.w3.org/TR/P3P/, Apr. 2002.
- L. F. Cranor, "Personal Privacy Assistants in the Age of the Internet of Things," in World Economic Forum Annual Meeting, 2016.

Section 2—Curate

To retain agency in the algorithmic era, we must provide every individual with a personal data or algorithmic agent they curate to represent their terms and conditions in any real, digital, or virtual environment. This "agent" would be empowered to act as an individual's legal proxy in the digital and virtual arena. Oftentimes, the functionality of this agent will be automated, operating along the lines of current ad blockers which do not permit prespecified algorithms to access a user's data. For other situations that might be unique or new to this agent, a user could specify that notices or updates be sent on a case-by-case basis to determine where there could be a concern.

Issue: What would it mean for a person to have an algorithmic agent helping them actively represent and curate their terms and conditions at all times?

Background

While it's essential to create your own terms and conditions to broadcast your preferences, it's also important to recognize that humans do not operate at an algorithmic speed or level. A significant part of retaining your agency in this

Personal Data and Individual Agency

way involves identifying trusted services that can essentially act on your behalf when making decisions about your data.

Part of this logic entails putting you “at the center of your data”. One of the greatest challenges to user agency is that once you give your data away, you do not know how it is being used or by whom. But when all transactions about your data go through your A/IS agent honoring your preferences, you have better opportunities to control how your information is shared.

As an example, with medical data—while it is assumed most would share all their medical data with their spouse—most would also not wish to share that same amount of data with their local gym. This is an issue that extends beyond privacy, meaning one’s cultural or individual preferences about what personal information to share, to utility and clarity. This type of sharing also benefits users or organizations on the receiving end of data from these exchanges. For instance, the local gym in the previous example may only need basic heart or general health information and would actually not wish to handle or store sensitive cancer or other personal health data for reasons of liability.

A precedent for this type of patient- or user-centric model comes from Glimpse, a service described by Jordan Crook from *TechCrunch* in his article, [“Apple acquired Glimpse, a personal health data startup”](#): “Glimpse works by letting users pull their own medical info into a single virtual space, with the ability to add documents and pictures to fill out the profile. From there, users can share that data (as a comprehensive picture) to whomever they wish.” The fact that

Apple acquired the startup points to the potential for the successful business model of user-centric data exchange and putting individuals at the center of their data.

A person’s A/IS agent is a proactive algorithmic tool honoring their terms and conditions in the digital, virtual, and physical worlds. Any public space where a user may not be aware they are under surveillance by facial recognition, biometric, or other tools that could track, store, and utilize their data can now provide overt opportunity for consent via an A/IS agent platform. Even where an individual is not sure they are being tracked, by broadcasting their terms and conditions via digital means, they can demonstrate their preferences in the public arena. Via Bluetooth or similar technologies, individuals could offer their terms and conditions in a ubiquitous and always-on manner. This means even when an individual’s terms and conditions are not honored, people would have the ability to demonstrate their desire not to be tracked which could provide a methodology for the democratic right to protest in a peaceful manner. And where those terms and conditions are recognized meaning technically recognized even if they are not honored one’s opinions could be formally logged via GPS and timestamp data.

The A/IS agent could serve as an educator and negotiator on behalf of its user by suggesting how requested data could be combined with other data that has already been provided, inform the user if data are being used in a way that was not authorized, or make recommendations to the user based on a personal profile. As a negotiator, the agent could broker conditions for sharing data and could include payment to the user as a

Personal Data and Individual Agency

term, or even retract consent for the use of data previously authorized, for instance, if a breach of conditions was detected.

Recommendations

Algorithmic agents should be developed for individuals to curate and share their personal data. Specifically:

- For purposes of privacy, a person must be able to set up complex permissions that reflect a variety of wishes.
- The agent should help a person foresee and mitigate potential ethical implications of specific machine learning data exchanges.
- A user should be able to override his/her personal agents should he/she decide that the service offered is worth the conditions imposed.
- An agent should enable machine-to-machine processing of information to compare, recommend, and assess offers and services.
- Institutional systems should ensure support for and respect the ability of individuals to bring their own agent to the relationship

without constraints that would make some guardians inherently incompatible or subject to censorship.

- Vulnerable parts of the population will need protection in the process of granting access.

Further Resources

- [IEEE P7006™ - IEEE Standards Project on Personal Data AI Agent Working Group](#). Designed as a tool to allow any individual to create their own personal “terms and conditions” for their data, the AI Agent will also provide a technological tool for individuals to manage and control their identity in the digital and virtual world.
- Tools allowing an individual to create a form of an algorithmic guardian are often labeled as PIMS, or Personal Information Management Services. [Nesta in the United Kingdom was one of the funders of early research about PIMS](#) conducted by [CtrlShift](#).

Personal Data and Individual Agency

Section 3—Control

To retain agency in the algorithmic era, we must provide every individual access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data.

Issue: How can we increase agency by providing individuals access to services allowing them to create a trusted identity to control the safe, specific, and finite exchange of their data?

Background

Pervasive behavior-tracking adversely affects human agency by recognizing our identity in every action we take on and offline. This is why identity as it relates to individual data is emerging at the forefront of the risks and opportunities related to use of personal information for A/IS. Across the identity landscape there is increasing tension between the requirement for federated identities versus a range of identities. In federated identities, all data are linked to a natural and identified person. When one has a range of identities, or personas, these can be context specific and determined by the use case. New movements, such as “Self-Sovereign Identity”—defined as the right of a person to determine his or her own identity—are emerging alongside legal identities, e.g., those issued by governments, banks, and regulatory authorities, to help put individuals at the center of their data in the algorithmic age.

Personas, identities that act as proxies, and pseudonymity are also critical requirements for privacy management and agency. These help individuals select an identity that is appropriate for the context they are in or wish to join. In these settings, trust transactions can still be enabled without giving up the “root” identity of the user. For example, it is possible to validate that a user is over eighteen or is eligible for a service.

Attribute verification will play a significant role in enabling individuals to select the identity that provides access without compromising agency. This type of access is especially important in dealing with the myriad of algorithms interacting with narrow segments of our identity data. In these situations, individuals typically are not aware of the context for how their data will be used.

Recommendation

Individuals should have access to trusted identity verification services to validate, prove, and support the context-specific use of their identity.

Further Resources

- Sovrin Foundation, [The Inevitable Rise of Self-Sovereign Identity](#), Sept. 29, 2016.
- T. Ruff, “[Three Models of Digital Identity Relationships](#),” [Evernym](#), Apr. 24, 2018.
- C. Pettey, [The Beginner’s Guide to Decentralized Identity](#). Gartner, 2018.
- C. Allen, [The Path to Self-Sovereign Identity](#). GitHub, 2017.

Personal Data and Individual Agency

Section 4—Children’s Data Issues

While the focus of this chapter is to provide all individuals with agency regarding their personal data, some sectors of society have little or no control. For some elderly individuals or the mentally ill, it is because they have been found to not have “mental capacity”, and for prisoners in the criminal justice system, society has taken control away as punishment. In the case of children, this is because they are considered human beings in development with evolving capacities.

We examine the issues of children as an example case and recommend either regulation or a technical architecture that provides a veil and buffer from harm until a child is at an age where they can claim personal responsibility for their decisions.

In many parts of the world, children are viewed by the law as being primarily charges of their parents who make choices on their behalf. In Europe, however, the state has a role in ensuring the “best interests of the child”^{1, 2}. In schools, the two interests operate side-by-side, with parents being given some control over their child’s education but with many decisions being made by the schools.

Many of the issues described above concern choices around personal data and the future impacts of how the data are gathered and shared. Children are at the forefront of technological developments with future educational and recreational technology gathering data from them all day at school and intelligent toys throughout their time at home.

As children post, click, search, and share information, their data are linked to various profiles, grouped into segmented audiences, and fed into machine learning algorithms. Some of these may be designed to target campaigns that increase sales, influence sentiment, encourage online games, impact social networks, or influence religious and political views. Data fed into algorithmic advertising is not only gathered from children’s online actions but also from their devices. An example of device data is browser fingerprinting.³ It includes a set of data about a child’s browser or operating system. Fingerprinting vastly increases privacy risks because it is used to link to an individual.

Increasingly, children’s beliefs and social norms are established by what they see and experience online. Their actions reflect what they believe is possible and expected. The report, “Digital Deceit: Technologies Behind Precision Propaganda on the Internet”⁴, explains how companies collect, process, and then monetize personal preferences, socioeconomic status, fears, political and religious beliefs, location, and patterns of internet use.

Companies, governments, political parties, and philosophical and religious organizations use data available about students and children to influence how they spend their time, money, and the people or institutions they trust and with whom they spend time and build relationships.

Many aspects of a child’s life can be digitized. Their behavioral, device, and network data are combined and used by machine learning

Personal Data and Individual Agency

algorithms to determine the information and content that best achieve the educational goals of the schools and the economic goals of the advertisers and platform companies.

Issue: Mass personalization of instruction

Background

The mass personalization of education offers better education for all at very low cost through A/IS-enabled computer-based instruction that promises to free up teachers to work with kids individually to pursue their passions. These applications will rely on the continuous gathering of personal data regarding mood, thought processes, private stories, physiological data, and more. The data will be used to construct a computational model of each child's interests, understanding, strengths, and weaknesses. The model provides an intimate understanding of how they think, what they understand, how they process information, or react to new information; all of which can be used to drive instructional content and feedback.

Sharing of this data between classes, enabling it to follow students through their schooling, will make the models more effective and beneficial to children, but it also exposes children and their families to social control. If performance data are correlated with social data on a family, it could be used by social authorities in decision-making about the family. For example, since 2015-2018, well-being digital tests were performed in schools in Denmark. Children were asked

about everything from bullying, loneliness, and stomachaches. Recently it was disclosed that although the collected data was presented as anonymous, they were not. Data were stored with social security numbers, correlated with other test data, and even used in case management by some Danish municipalities.⁵ Commercial profiling and correlation of different sets of personal data may further affect these children in future job or educational situations.

Recommendation

Educational data offer a unique opportunity to model individuals' thought processes and could be used to predict or change individuals' behavior in many situations. Governments and organizations should classify educational data as being sensitive and implement special protective standards.

Children's data should be held in "escrow" and not used for any commercial purposes until a child reaches the age of majority and is able to authorize use as they choose.

Further Resources

- The journal of the International Artificial Intelligence in Education Society: <http://iaied.org/journal/>
- Deeper discussion and bibliography of future trends of AI-based education with utopian and dystopian case scenarios: N. Pinkwart, "Another 25 Years of AIED? Challenges and Opportunities for Intelligent Educational Technologies of the Future," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 2, pp. 771–783, 2016. [Online].

Personal Data and Individual Agency

Available: <https://doi.org/10.1007/s40593-016-0099-7> [Accessed Dec. 2018].

- Information Commissioners Office (ico.), "What if we want to profile children or make automated decisions about them?" <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/children-and-the-gdpr/what-if-we-want-to-profile-children-or-make-automated-decisions-about-them/>
- K. Firth-Butterfield, "What happens when your child's friend is an AI toy that talks back?" in World Economic Forum: Generation AI, <https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/>, May 22, 2018.

Issue: Technology choice-making in schools

Background

Children, as minors, have no standing to give or deny consent, or to control the use of their personal data. Parents only have limited choices in what are often school-wide implementations of educational technology. Examples include the use of Google applications, face recognition in security systems, and computer driven instruction as described above. In many cases, parents' only choice would be to send their children to a different school, but that choice is seldom available.

How should schools make these choices? How much input should parents have? Should parents be able to demand technology-free teaching?

There are many gaps in current student data regulation. In June 2018, CLIP, The Center on Law and Information Policy at Fordham Law School published, "Transparency and the Marketplace for Student Data".⁶ This study concluded that "student lists are commercially available for purchase on the basis of ethnicity, affluence, religion, lifestyle, awkwardness, and even a perceived or predicted need for family planning services". Fordham found that the data market is becoming one of the largest and most profitable marketplaces in the United States. Data brokers have databases that store billions of data elements on nearly every United States consumer. However, information from students in the pursuit of an education should not be exploited and commercialized without restraint.

Fordham researchers found at least 14 data brokers who advertise the sale of student information. One sold lists of students as young as two years old. Another sold lists of student profiles on the basis of ethnicity, religion, economic factors, and even gawkiness.

Recommendation

Local and national educational authorities must work to develop policies surrounding students' personal data with all stakeholders: administrators, teachers, technology providers, students, and parents in order to balance the best educational interests of each child with the best practices to ensure safety of their personal data. Such efforts will raise awareness among all stakeholders of the promise and the compromises inherent in new educational technologies.

Personal Data and Individual Agency

Further Resources

- Common Sense Media privacy evaluation project: <https://www.commonsense.org/education/privacy>
- D. T. Ritvo, L. Plunkett, and P. Haduong, "Privacy and Student Data: Companion Learning Tools." Berkman Klein Center for Internet and Society at Harvard University, 2017. [Online]. Available: http://blogs.harvard.edu/youthandmediaalpha/files/2017/03/PrivacyStudentData_Companion_Learning_Tools.pdf [Accessed Dec. 2018].
- F. Alim, N. Cardozo, G. Gebhart, K. Gullo, and A. Kalia, "Spying on Students: School-Issued Devices and Student Privacy," Electronic Frontier Foundation, <https://www.eff.org/wp/school-issued-devices-and-student-privacy>, April 13, 2017.
- N. C. Russell, J. R. Reidenberg, E. Martin, and T. Norton, "Transparency and the Marketplace for Student Data," *Virginia Journal of Law and Technology*, Forthcoming. Available at SSRN: <https://ssrn.com/abstract=3191436>, June 6, 2018.

Issue: Intelligent toys

Background

Children will not only be exposed to A/IS at school but also at home, while they play and while they sleep. Toys are already being sold that offer interactive, intelligent opportunities for play. Many of them collect video and audio data which is stored on company servers and either is or could be mined for profiling or marketing data.

There is currently little regulatory oversight. In the United States COPPA⁷ offers some protection for the data of children under 13. Germany has outlawed such toys using legislation banning spying equipment enacted in 1981. Corporate A/IS are being embodied in toys and given to children to play with, to talk to, tell stories to, and to explore all the personal development issues that we learn about in private play as children.

Recommendations

Child data should be held in "escrow" and not used for any commercial purposes until a child reaches the age of majority and is able to authorize use as they choose.

Governments and organizations need to educate and inform parents of the mechanisms of A/IS and data collection in toys and the possible impact on children in the future.

Further Resources

- K. Firth-Butterfield, "What happens when your child's friend is an AI toy that talks back?" in World Economic Forum: Generation AI, <https://www.weforum.org/agenda/2018/05/generation-ai-what-happens-when-your-childs-invisible-friend-is-an-ai-toy-that-talks-back/>, May 22, 2018.
- D. Basulto, "How artificial intelligence is moving from the lab to your kid's playroom," Washington Post, Oct. 15, 2015. [Online]. Available: https://www.washingtonpost.com/news/innovations/wp/2015/10/15/how-artificial-intelligence-is-moving-from-the-lab-to-your-kids-playroom/?utm_term=.89a1431a05a7 [Accessed Dec. 1, 2018].

Personal Data and Individual Agency

- S. Chaudron, R. Di Gioia, M. Gemo, D. Holloway, J. Marsh, G. Mascheroni J. Peter, and D. Yamada-Rice , <http://publications.jrc.ec.europa.eu/repository/handle/JRC105061>, 2016.
- S. Chaudron, R. Di Gioia, M. Gemo, D. Holloway, J. Marsh, G. Mascheroni, J. Peter, D. Yamada-Rice [Kaleidoscope on the Internet of Toys - Safety, security, privacy and societal insights](#), EUR 28397 EN, doi:10.2788/05383, Luxembourg: Publications Office of the European Union, 2017.
- Z. Kleinman, "Alexa, are you friends with our kids?" *BBC News*, July 16, 2018. [Online]. Available: <https://www.bbc.com/news/technology-44847184.5b>. [Accessed Dec. 1, 2018].
- J. Wakefield, "Germany bans children's smartwatches." *BBC News*, Nov. 17 2017. [Online]. Available: <https://www.bbc.co.uk/news/technology-42030109>. [Accessed Dec. 2018].

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Personal Data and Individual Agency Committee

- **Katryna Dow** (Co-Chair) – CEO & Founder at Meeco
- **John C. Havens** (Co-Chair) – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Mads Schaarup Andersen** – Senior Usable Security Expert in the Alexandra Institute's Security Lab
- **Ajay Bawa** – Technology Innovation Lead, Avanade Inc.
- **Ariel H. Brio** – Privacy and Data Counsel at Sony Interactive Entertainment
- **Walter Burrough** – Co-Founder, Augmented Choice; PhD Candidate (Computer Science) – Serious Games Institute
- **Danny W. Devriendt** – Managing director of Mediabrands Dynamic (IPG) in Brussels, and the CEO of the Eye of Horus, a global think-tank for communication-technology related topics
- **Dr. D. Michael Franklin** – Assistant Professor, Kennesaw State University, Marietta Campus, Marietta, GA
- **Jean-Gabriel Ganascia** – Professor, University Pierre et Marie Curie; LIP6 Laboratory ACASA Group Leader

Personal Data and Individual Agency

- **Bryant Joseph Gilot, MD CM DPhil MSc** – Center for Personalised Medicine, University of Tuebingen Medical Center, Germany & Chief Medical Officer, Blockchain Health Co., San Francisco
- **David Goldstein** – Seton Hall University
- **Adrian Gropper, M.D.** – CTO, Patient Privacy Rights Foundation; HIE of One Project
- **Marsali S. Hancock** – Chair, IEEE Standards for Child and Student Data governance, CEO and Co-Foundation EP3 Foundation.F
- **Gry Hasselbalch** – Founder DataEthics, Author, *Data Ethics - The New Competitive Advantage*
- **Yanqing Hong** – Graduate, University of Utrecht Researcher at Tsinghua University
- **Professor Meg Leta Jones** – Assistant Professor in the Communication, Culture & Technology program at Georgetown University
- **Mahsa Kiani** – Chair of Student Activities, IEEE Canada; Vice Editor, IEEE Canada Newsletter (ICN); PhD Candidate, Faculty of Computer Science, University of New Brunswick
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Emma Lindley** – Founder, Innovate Identity
- **Ewa Luger** – Chancellor's Fellow at the University of Edinburgh, within the Design Informatics Group
- **Sean Martin McDonald** – CEO of FrontlineSMS, Fellow at Stanford's Digital Civil Society Lab, Principal at Digital Public
- **Hiroshi Nakagawa** – Professor, The University of Tokyo, and AI in Society Research Group Director at RIKEN Center for Advanced Intelligence Project (AIP)
- **Sofia C. Olhede** – Professor of Statistics and an Honorary Professor of Computer Science at University College London, London, U.K; Member of the Programme Committee of the International Centre for Mathematical Sciences.
- **Ugo Pagallo** – University of Turin Law School; Center for Transnational Legal Studies, London; NEXA Center for Internet & Society, Politecnico of Turin
- **Dr. Juuso Parkkinen** – Senior Data Scientist, Nightingale Health; Programme Team Member, MyData 2017 conference
- **Eleonore Pauwels** – Research Fellow on AI and Emerging Cybertechnologies, United Nations University (NY) and Director of the AI Lab, Woodrow Wilson International Center for Scholars (DC)
- **Dr. Deborah C. Peel** – Founder, Patient Privacy Rights & Creator, the International Summits on the Future of Health Privacy
- **Walter Pienciak** – Principal Architect, Advanced Cognitive Architectures, Ltd.
- **Professor Serena Quattrocchio** – University of Turin Law School
- **Carolyn Robson** – Group Data Privacy Manager at Etihad Aviation Group
- **Gilad Rosner** – Internet of Things Privacy Forum; Horizon Digital Economy Research Institute, UK; UC Berkeley Information School

Personal Data and Individual Agency

- **Prof. Dr.-Ing. Ahmad-Reza Sadeghi** – Director System Security Lab, Technische Universität Darmstadt / Director Intel Collaborative Research Institute for Secure Computing
- **Rose Shuman** – Partner at BrightFront Group & Founder, Question Box
- **Dr. Zoltán Szlávik** – Lead/Researcher, IBM Center for Advanced Studies Benelux
- **Udbhav Tiwari** – Centre for Internet and Society, India

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Endnotes

¹ Europäische Union, Europäischer Gerichtshof für Menschenrechte, & Europarat (Eds.). (2015). Handbook on European law relating to the rights of the child. Luxembourg: Publications Office of the European Union. https://www.echr.coe.int/Documents/Handbook_rights_child_ENG.PDF

² Children Act (1989). Retrieved from <https://www.legislation.gov.uk/ukpga/1989/41/section/1>

³ “Browser fingerprints, and why they are so hard to erase | Network World.” 17 Feb. 2015, <https://www.networkworld.com/article/2884026/security0/browser-fingerprints-and-why-they-are-so-hard-to-erase.html>. Accessed 25 July. 2018.

⁴ D. Gosh and B. Scott, “Digital Deceit: The Technologies behind Precision Propaganda on the Internet” 23 Jan. 2018, <https://www.newamerica.org/public-interest-technology/policy-papers/digital-deceit/>. Accessed 10 Nov 2018.

⁵ Case described in Danish here <https://dataethics.eu/trivsel-enhver-pris/>

⁶ Russell, N. Cameron, Reidenberg, Joel R., Martin, Elizabeth, and Norton, Thomas, “Transparency and the Marketplace for Student Data” (June 6, 2018). Virginia Journal of Law and Technology, Forthcoming. Available at SSRN: <https://ssrn.com/abstract=3191436>

⁷ Children’s Online Privacy Protection Act (COPPA) - <https://www.ftc.gov/tips-advice/business-center/privacy-and-security/children%27s-privacy>

Methods to Guide Ethical Research and Design

Autonomous and intelligent systems (A/IS) research and design must be developed against the backdrop that technology is not neutral. A/IS embody values and biases that can influence important social processes like voting, policing, and banking. To ensure that A/IS benefit humanity, A/IS research and design must be underpinned by ethical and legal norms. These should be instantiated through values-based research and design methods. Such methods put human well-being at the core of A/IS development.

To help achieve these goals, researchers, product developers, and technologists across all sectors need to embrace research and development methods that evaluate their processes, products, values, and design practices in light of the concerns and considerations raised in this chapter. This chapter is split up into three sections:

Section 1—Interdisciplinary Education and Research

Section 2—Corporate Practices on A/IS

Section 3—Responsibility and Assessment

Each of the sections highlights various areas of concern (issues) as well as recommendations and further resources.

Overall, we address both structural and individual approaches. We discuss how to improve the ethical research and business practices surrounding the development of A/IS and attend to the responsibility of the technology sector vis-à-vis the public interest. We also look at that what can be done at the level of educational institutions, among others, informing engineering students about ethics, social justice, and human rights. The values-based research and design method will require a change of current system development approaches for organizations. This includes a commitment of research institutions to strong ethical guidelines for research and of businesses to values that transcend narrow economic incentives.

Methods to Guide Ethical Research and Design

Section 1—Interdisciplinary Education and Research

Integrating applied ethics into education and research to address the issues of A/IS requires an interdisciplinary approach, bringing together humanities, social sciences, physical sciences, engineering, and other disciplines.

Issue: Integration of ethics in A/IS-related degree programs

Background

A/IS engineers and design teams do not always thoroughly explore the ethical considerations implicit in their technical work and design choices. Moreover, the overall science, technology, engineering, and mathematics (STEM) field struggles with the complexity of ethical considerations, which cannot be readily articulated and translated into the formal languages of mathematics and computer programming associated with algorithms and machine learning.

Ethical issues can easily be rendered invisible or inappropriately reduced and simplified in the context of technical practice. For the dangers of this approach see for instance, Lipton and Steinhardt (2018), listed under “Further Resources”. This problem is further compounded by the fact that many STEM programs do not

sufficiently integrate applied ethics throughout their curricula. When they do, often ethics is relegated to a stand-alone course or module that gives students little or no direct experience in ethical decision-making. Ethics education should be meaningful, applicable, and incorporate best practices from the broader field.

The aim of these recommendations is to prepare students for the technical training and engineering development methods that incorporate ethics as essential so that ethics, and relevant principles, like human rights, become naturally a part of the design process.

Recommendations

- Ethics training needs to be a core subject for all those in the STEM field, beginning at the earliest appropriate level and for all advanced degrees.
- Effective STEM ethics curricula should be informed by experts outside the STEM community from a variety of cultural and educational backgrounds to ensure that students acquire sensitivity to a diversity of robust perspectives on ethics and design.
- Such curricula should teach aspiring engineers, computer scientists, and statisticians about the relevance and impact of their decisions in designing A/IS technologies. Effective

Methods to Guide Ethical Research and Design

ethics education in STEM contexts and beyond should span primary, secondary, and postsecondary education, and include both universities and vocational training schools.

- Relevant accreditation bodies should reinforce this integrated approach as outlined above.

Further Resources

- [IEEE P7000™ Standards Project for a Model Process for Addressing Ethical Concerns During System Design](#). IEEE P7000 aims to enhance corporate IT innovation practices by providing processes for embedding a values- and virtue-based thinking, culture, and practice into them.
- Z. Lipton and J. Steinhardt, [Troubling Trends in Machine Learning Scholarship](#). ICML conference paper, July 2018.
- J. Holdren, and M. Smith. "[Preparing for the Future of Artificial Intelligence](#)." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016.
- Comparing the UK, EU, and US approaches to AI and ethics: C. Cath, S. Wachter, B. Mittelstadt, et al., "[Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach](#)." *Science and Engineering Ethics*, vol. 24, pp. 505-528, 2017.

Issue: Interdisciplinary collaborations

Background

More institutional resources and incentive structures are necessary to bring A/IS engineers and designers into sustained and constructive contact with ethicists, legal scholars, and social scientists, both in academia and industry. This contact is necessary as it can enable meaningful interdisciplinary collaboration and shape the future of technological innovation. More could be done to develop methods, shared knowledge, and lexicons that would facilitate such collaboration.

This issue relates, among other things, to funding models as well as the lack of diversity of backgrounds and perspectives in A/IS-related institutions and companies, which limit cross-pollination between disciplines. To help bridge this gap, additional translation work and resource sharing, including websites and Massive Open Online Courses (MOOCs), need to happen among technologists and other relevant experts, e.g., in medicine, architecture, law, philosophy, psychology, and cognitive science. Furthermore, there is a need for more cross-disciplinary conversation and multi-disciplinary research, as is being done, for instance, at the annual ACM Fairness, Accountability, and Transparency (FAT*) conference or the work done by the Canadian Institute For Advanced Research (CIFAR), which is developing Canada's AI strategy.

Methods to Guide Ethical Research and Design

Recommendations

Funding models and institutional incentive structures should be reviewed and revised to prioritize projects with interdisciplinary ethics components to encourage integration of ethics into projects at all levels.

Further Resources

- S. Barocas, Course Material for Ethics and Policy in Data Science, Cornell University, 2017.
- L. Floridi, and M. Taddeo. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society*, vol. 374, no. 2083, 1–4. DOI [10.1098/rsta.2016.0360](https://doi.org/10.1098/rsta.2016.0360), 2016.
- S. Spiekermann, Ethical IT Innovation: A Value-Based System Design Approach. Boca Raton, FL: Auerbach Publications, 2015.
- K. Crawford, "[Artificial Intelligence's White Guy Problem](https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=1)", *New York Times*, July 25, 2016. [Online]. Available: http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=1. [Accessed October 28, 2018].

Issue: A/IS culture and context

Background

A responsible approach to embedding values into A/IS requires that algorithms and systems are created in a way that is sensitive to the variation of ethical practices and beliefs across cultures. The designers of A/IS need to be mindful of cross-cultural ethical variations while also respecting widely held international legal norms.

Recommendation

Establish a leading role for [intercultural information ethics](#) (IIE) practitioners in ethics committees informing technologists, policy makers, and engineers. Clearly demonstrate through examples how cultural variation informs not only information flows and information systems, but also algorithmic decision-making and value by design.

Further Resources

- D. J. Pauleen, et al. "[Cultural Bias in Information Systems Research and Practice: Are You Coming From the Same Place I Am?](#)" *Communications of the Association for Information Systems*, vol. 17, no. 17, 2006.
- J. Bielby, "[Comparative Philosophies in Intercultural Information Ethics](#)," *Confluence: Online Journal of World Philosophies* 2, no. 1, pp. 233–253, 2016.

Methods to Guide Ethical Research and Design

Issue: Institutional ethics committees in the A/IS fields

Background

It is unclear how research on the interface of humans and A/IS, animals and A/IS, and biological hazards will impact research ethical review boards. Norms, institutional controls, and risk metrics appropriate to the technology are not well established in the relevant literature and research governance infrastructure. Additionally, national and international regulations governing review of human-subjects research may explicitly or implicitly exclude A/IS research from their purview on the basis of legal technicalities or medical ethical concerns, regardless of the potential harms posed by the research.

Research on A/IS human-machine interaction, when it involves intervention or interaction with identifiable human participants or their data, typically falls to the governance of research ethics boards, e.g., institutional review boards. The national level and institutional resources, e.g., hospitals and universities, necessary to govern ethical conduct of Human-Computer Interaction (HCI), particularly within the disciplines pertinent to A/IS research, are underdeveloped.

First, there is limited international or national guidance to govern this form of research. Sections of IEEE standards governing research on A/IS in medical devices address some of the issues related to the security of A/IS-enabled devices. However, the ethics of testing those devices for the purpose of bringing them

to market are not developed into policies or guidance documents from recognized national and international bodies, e.g., U.S. Food and Drug Administration (FDA) and EU European Medicines Agency (EMA). Second, the bodies that typically train individuals to be gatekeepers for the research ethics bodies are under-resourced in terms of expertise for A/IS development, e.g., Public Responsibility in Medicine and Research (PRIM&R) and the Society of Clinical Research Associates (SoCRA). Third, it is not clear whether there is sufficient attention paid to A/IS ethics by research ethics board members or by researchers whose projects involve the use of human participants or their identifiable data.

For example, research pertinent to the ethics-governing research at the interface of animals and A/IS research is underdeveloped with respect to systematization for implementation by the Institutional Animal Care and Use Committee (IACUC) or other relevant committees. In institutions without a veterinary school, it is unclear that the organization would have the relevant resources necessary to conduct an ethical review of such research.

Similarly, research pertinent to the intersection of radiological, biological, and toxicological research—ordinarily governed under institutional biosafety committees—and A/IS research is not often found in the literature pertinent to research ethics or research governance.

Methods to Guide Ethical Research and Design

Recommendation

The IEEE and other standards-setting bodies should draw upon existing standards, empirical research, and expertise to identify priorities and develop standards for the governance of A/IS research and partner with relevant national agencies, and international organizations, when possible.

Further Resources

- S. R. Jordan, "The Innovation Imperative." *Public Management Review* 16, no. 1, pp. 67–89, 2014.
- B. Schneiderman, "[The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight](#)." *Proceedings of the National Academy of Sciences of the United States of America* 113, no. 48, 13538–13540, 2016.
- J. Metcalf and K. Crawford, "[Where are Human Subjects in Big Data Research? The Emerging Ethics Divide](#)." *Big Data & Society*, May 14, 2016. [Online]. Available: SSRN: <https://ssrn.com/abstract=2779647>. [Accessed Nov. 1, 2018].
- R. Calo, "[Consumer Subject Review Boards: A Thought Experiment](#)." *Stanford Law Review Online* 66 97, Sept. 2013.

Methods to Guide Ethical Research and Design

Section 2—Corporate Practices on A/IS

Corporations are eager to develop, deploy, and monetize A/IS, but there are insufficient structures in place for creating and supporting ethical systems and practices around A/IS funding, development, and use.

Issue: Values-based ethical culture and practices for industry

Background

Corporations are built to create profit while competing for market share. This can lead corporations to focus on growth at the expense of avoiding negative ethical consequences. Given the deep ethical implications of widespread deployment of A/IS, in addition to laws and regulations, there is a need to create values-based ethical culture and practices for the development and deployment of those systems. To do so, we need to further identify and refine corporate processes that facilitate values-based design.

Recommendations

The building blocks of such practices include top-down leadership, bottom-up empowerment, ownership, and responsibility, along with the need to consider system deployment contexts and/or ecosystems. Corporations should identify stages in their processes in which ethical considerations, “ethics filters”, are in place before products are further developed and deployed.

For instance, if an ethics review board comes in at the right time during the A/IS creation process, it would help mitigate the likelihood of creating ethically problematic designs. The institution of an ethical A/IS corporate culture would accelerate the adoption of the other recommendations within this section focused on business practices.

Further Resources

- [ACM Code of Ethics and Professional Ethics](#), which includes various references to human well-being and human rights, 2018.
- Report of UN Special Rapporteur on [Freedom of Expression. AI and Freedom of Expression](#), 2018.
- The [website of the Benefit corporations](#) (B-corporations) provides a good overview of a range of companies that personify this type of culture.
- R. Sisodia, J. N. Sheth and D. Wolfe, [Firms of Endearment](#), 2nd edition. Upper Saddle River, NJ: FT Press, 2014. This book showcases how companies embracing values and a stakeholder approach outperform their competitors in the long run.

Methods to Guide Ethical Research and Design

Issue: Values-based leadership

Background

Technology leadership should give innovation teams and engineers direction regarding which human values and legal norms should be promoted in the design of A/IS. Cultivating an ethical corporate culture is an essential component of successful leadership in the A/IS domain.

Recommendations

Companies should create roles for senior-level marketers, engineers, and lawyers who can collectively and pragmatically implement ethically aligned design. There is also a need for more in-house ethicists, or positions that fulfill similar roles. One potential way to ensure values are on the agenda in A/IS development is to have a Chief Values Officer (CVO), a role first suggested by Kay Firth-Butterfield, see “Further Resources”. However, ethical responsibility should not be delegated solely to CVOs. They can support the creation of ethical knowledge in companies, but in the end, all members of an organization will need to act responsibly throughout the design process.

Companies need to ensure that their understanding of values-based system innovation is based on *de jure* and *de facto* international human rights standards.

Further Resources

- K. Firth-Butterfield, “[How IEEE Aims to Instill Ethics in Artificial Intelligence Design](http://theinstitute.ieee.org/ieee-roundup/blogs/blog/how-ieee-aims-to-instill-ethics-in-artificial-intelligence-design),” The Institute. Jan. 19, 2017. [Online]. Available: <http://theinstitute.ieee.org/ieee-roundup/blogs/blog/how-ieee-aims-to-instill-ethics-in-artificial-intelligence-design>. [Accessed October 28, 2018].
- United Nations, [Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework](#), New York and Geneva: UN, 2011.
- Institute for Human Rights and Business (IHRB), and Shift, ICT Sector Guide on Implementing the UN Guiding Principles on Business and Human Rights, 2013.
- C. Cath, and L. Floridi, “[The Design of the Internet’s Architecture by the Internet Engineering Task Force \(IETF\) and Human Rights](#),” *Science and Engineering Ethics*, vol. 23, no. 2, pp. 449–468, Apr. 2017.

Issue: Empowerment to raise ethical concerns

Background

Engineers and design teams may encounter obstacles to raising ethical concerns regarding their designs or design specifications within their organizations. Corporate culture should incentivize technical staff to voice the full range of ethical questions to relevant corporate actors throughout the full product lifecycle, including the design, development, and deployment

Methods to Guide Ethical Research and Design

phases. Because raising ethical concerns can be perceived as slowing or halting a design project, organizations need to consider how they can recognize and incentivize values-based design as an integral component of product development.

Recommendations

Employees should be empowered and encouraged to raise ethical concerns in day-to-day professional practice.

To be effective in ensuring adoption of ethical considerations during product development or internal implementation of A/IS, organizations should create a company culture and set of norms that encourage incorporating ethical considerations in the design and implementation processes.

New categories of considerations around these issues need to be accommodated, along with updated Codes of Conduct, company value-statements, and other management principles so individuals are empowered to share their insights and concerns in an atmosphere of trust. Additionally, bottom-up approaches like company “town hall meetings” should be explored that reward, rather than punish, those who bring up ethical concerns.

Further Resources

- [The British Computer Society \(BCS\)](#), Code of Conduct, 2019.
- C. Cath, and L. Floridi, “[The Design of the Internet’s Architecture by the Internet Engineering Task Force \(IETF\) and Human Rights](#),” *Science and Engineering Ethics*, vol. 23, no. 2, pp. 449–468, Apr. 2017.

Issue: Ownership and responsibility

Background

There is variance within the technology community on how it sees its responsibility regarding A/IS. The difference in values and behaviors are not necessarily aligned with the broader set of social concerns raised by public, legal, and professional communities. The current makeup of most organizations has clear delineations among engineering, legal, and marketing functions. Thus, technologists will often be incentivized in terms of meeting functional requirements, deadline, and financial constraints, but for larger social issues may say, “Legal will handle that.” In addition, in employment and management technology or work contexts, “ethics” typically refers to a code of conduct regarding professional behavior versus a values-driven design process mentality.

As such, ethics regarding professional conduct often implies moral issues such as integrity or the lack thereof, in the case of whistleblowing, for instance. However, ethics in A/IS design include broader considerations about the consequences of technologies.

Recommendations

Organizations should clarify the relationship between professional ethics and applied A/IS ethics by helping or enabling designers, engineers, and other company representatives to discern the differences between these kinds of ethics and where they complement each other.

Methods to Guide Ethical Research and Design

Corporate ethical review boards, or comparable mechanisms, should be formed to address ethical and behavioral concerns in relation to A/IS design, development and deployment. Such boards should seek an appropriately diverse composition and use relevant criteria, including both research ethics and product ethics, at the appropriate levels of advancement of research and development. These boards should examine justifications of research or industrial projects.

Further Resources

- HH van der Kloot Meijberg and RHJ ter Meulen, "[Developing Standards for Institutional Ethics Committees: Lessons from the Netherlands](#)," *Journal of Medical Ethics* 27 i36-i40, 2001.

Issue: Stakeholder inclusion

Background

The interface between A/IS and practitioners, as well as other stakeholders, is gaining broader attention in domains such as healthcare diagnostics, and there are many other contexts where there may be different levels of involvement with the technology. We should recognize that, for example, occupational therapists and their assistants may have on-the-ground expertise in working with a patient, who might be the "end user" of a robot or social A/IS technology. In order to develop a product that is ethically aligned, stakeholders' feedback is crucial to design a system that takes ethical and social issues into account. There are successful user experience (UX) design concepts, such

as accessibility, that consider human physical disabilities, which should be incorporated into A/IS as they are more widely deployed. It is important to continuously consider the impact of A/IS through unanticipated use and on unforeseen interests.

Recommendations

To ensure representation of stakeholders, organizations should enact a planned and controlled set of activities to account for the interests of the full range of stakeholders or practitioners who will be working alongside A/IS and incorporating their insights to build upon, rather than circumvent or ignore, the social and practical wisdom of involved practitioners and other stakeholders.

Further Resources

- C. Schroeter, et al., "[Realization and User Evaluation of a Companion Robot for People with Mild Cognitive Impairments](#)," *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2013)*, Karlsruhe, Germany 2013. pp. 1145–1151.
- T. L. Chen, et al. "[Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities](#)," *IEEE Robotics and Automation Magazine*, vol. 20, no. 1, pp. 30–39, 2013.
- R. Hartson, and P. S. Pyla. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Waltham, MA: Elsevier, 2012.

Methods to Guide Ethical Research and Design

Issue: Values-based design

Background

Ethics are often treated as an impediment to innovation, even among those who ostensibly support ethical design practices. In industries that reward rapid innovation in particular, it is necessary to develop ethical design practices that integrate effectively with existing engineering workflows. Those who advocate for ethical design within a company should be seen as innovators seeking the best outcomes for the company, end users, and society. Leaders can facilitate that mindset by promoting an organizational structure that supports the integration of dialogue about ethics throughout product life cycles.

AI/IS design processes often present moments where ethical consequences can be highlighted. There are no universally prescribed models for this because organizations vary significantly in structure and culture. In some organizations, design team meetings may be brief and informal. In others, the meetings may be lengthy and structured. The transition points between discovery, prototyping, release, and revisions are natural contexts for conducting such reviews. Iterative review processes are also advisable, in part because changes to risk profiles over time can illustrate needs or opportunities for improving the final product.

Recommendations

Companies should study design processes to identify situations where engineers and researchers can be encouraged to raise and resolve questions of ethics and foster a proactive environment to realize ethically aligned design. Achieving a distributed responsibility for ethics requires that all people involved in product design are encouraged to notice and respond to ethical concerns. Organizations should consider how they can best encourage and facilitate deliberations among peers.

Organizations should identify points for formal review during product development. These reviews can focus on “red flags” that have been identified in advance as indicators of risk. For example, if the datasets involve minors or focus on users from protected classes, then it may require additional justification or alterations to the research or development protocols.

Further Resources

- A. Sinclair, “[Approaches to Organizational Culture and Ethics](#),” *Journal of Business Ethics*, vol. 12, no. 1, pp. 63–73, 1993.
- Al Y. S. Chen, R. B. Sawyers, and P. F. Williams. “[Reinforcing Ethical Decision Making Through Corporate Culture](#),” *Journal of Business Ethics* 16, no. 8, pp. 855–865, 1997.
- K. Crawford and R. Calo, “[There Is a Blind Spot in AI Research](#),” *Nature* 538, pp. 311–313, 2016.

Methods to Guide Ethical Research and Design

Section 3—Responsibility and Assessment

Lack of accountability of the A/IS design and development process presents a challenge to ethical implementation and oversight. This section presents four issues, moving from macro oversight to micro documentation practices.

Issue: Oversight for algorithms

The algorithms behind A/IS are not subject to consistent oversight. This lack of assessment causes concern because end users have no account of how a certain algorithm or system came to its conclusions. These recommendations are similar to those made in the “General Principles” and “Embedding Values into Autonomous and Intelligent Systems” chapters of *Ethically Aligned Design*, but here the recommendations are used as they apply to the narrow scope of this chapter .

Recommendations

Accountability: As touched on in the General Principles chapter of *Ethically Aligned Design*, algorithmic transparency is an issue of concern. It is understood that specifics relating to algorithms or systems contain intellectual property that cannot, or will not, be released to the general public. Nonetheless, standards providing oversight of the manufacturing process of A/IS technologies need to be created to avoid harm and negative consequences. We can look to other technical domains, such as biomedical, civil, and aerospace engineering, where commercial

protections for proprietary technology are routinely and effectively balanced with the need for appropriate oversight standards and mechanisms to safeguard the public.

Human rights and algorithmic impact assessments should be explored as a meaningful way to improve the accountability of A/IS. These need to be paired with public consultations, and the final impact assessments must be made public.

Further Resources

- F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press, 2016.
- R. Calo, “Artificial Intelligence Policy: A Primer and Roadmap,” *UC Davis Law Review*, 52: pp. 399–435, 2017.
- ARTICLE 19. “Privacy and Freedom of Expression in the Age of Artificial Intelligence,” Privacy International, April 2018. [Online]. Available: <https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf>. [Accessed October 28, 2018].

Methods to Guide Ethical Research and Design

Issue: Independent review organization

Background

We need independent, expert opinions that provide guidance to the general public regarding A/IS. Currently, there is a gap between how A/IS are marketed and their actual performance or application. We need to ensure that A/IS technology is accompanied by best-use recommendations and associated warnings. Additionally, we need to develop a certification scheme for A/IS which ensures that the technologies have been independently assessed as being safe and ethically sound.

For example, today it is possible for systems to download new self-parking functionality to cars, and no independent reviewer establishes or characterizes boundaries or use. Or, when a companion robot promises to watch your children, there is no organization that can issue an independent seal of approval or limitation on these devices. We need a ratings and approval system ready to serve social/automation technologies that will come online as soon as possible. We also need further government funding for research into how A/IS technologies can best be subjected to review, and how review organizations can consider both traditional health and safety issues, as well as ethical considerations.

Recommendations

An independent, internationally coordinated body—akin to ISO—should be formed to oversee whether A/IS products actually meet ethical criteria, both when designed, developed, deployed, and when considering their evolution after deployment and during interaction with other products. It should also include a certification process.

Further Resources

- A. Tutt, “An FDA for Algorithms,” *Administrative Law Review* 69, 83–123, 2016.
- M. U. Scherer, “[Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies](#),” *Harvard Journal of Law and Technology* vol. 29, no. 2, 354–400, 2016.
- D. R. Desai and J. A. Kroll, “[Trust But Verify: A Guide to Algorithms and the Law](#),” *Harvard Journal of Law and Technology*, Forthcoming; Georgia Tech Scheller College of Business Research Paper No. 17-19, 2017.

Issue: Use of black-box components

Background

Software developers regularly use “black box” components in their software, the functioning of which they often do not fully understand. “Deep” machine learning processes, which are driving many advancements in autonomous and intelligent systems, are a growing source of black box software. At least for the foreseeable future, A/IS developers will likely be unable to build systems that are guaranteed to operate as intended.

Methods to Guide Ethical Research and Design

Recommendations

When systems are built that could impact the safety or well-being of humans, it is not enough to just presume that a system works. Engineers must acknowledge and assess the ethical risks involved with black box software and implement mitigation strategies.

Technologists should be able to characterize what their algorithms or systems are going to do via documentation, audits, and transparent and traceable standards. To the degree possible, these characterizations should be predictive, but given the nature of A/IS, they might need to be more retrospective and mitigation-oriented. As such, it is also important to ensure access to remedy adverse impacts.

Technologists and corporations must do their ethical due diligence before deploying A/IS technology. Standards for what constitutes ethical due diligence would ideally be generated by an international body such as IEEE or ISO, and barring that, each corporation should work to generate a set of ethical standards by which their processes are evaluated and modified. Similar to a flight data recorder in the field of aviation, algorithmic traceability can provide insights on what computations led to questionable or dangerous behaviors. Even where such processes remain somewhat opaque, technologists should seek indirect means of validating results and detecting harms.

Further Resources

- M. Ananny and K. Crawford, "[Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability](#)," *New Media & Society*, vol. 20, no. 3, pp. 973-989, Dec. 13, 2016.
- D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," AI NOW 2018. [Online]. Available: <https://ainowinstitute.org/aiareport2018.pdf>. [Accessed October 28, 2018].
- J. A. Kroll "[The Fallacy of Inscrutability](#)," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, C. Cath, S. Wachter, B. Mittelstadt and L. Floridi, Eds., October 15, 2018 DOI: 10.1098/rsta.2018.0084.

Issue: Need for better technical documentation

Background

A/IS are often construed as fundamentally opaque and inscrutable. However, lack of transparency is often the result of human decision. The problem can be traced to a variety of sources, including poor documentation that excludes vital information about the limitations and assumptions of a system. Better documentation combined with internal and external auditing are crucial to understanding a system's ethical impact.

Methods to Guide Ethical Research and Design

Recommendation

Engineers should be required to thoroughly document the end product and related data flows, performance, limitations, and risks of A/IS. Behaviors and practices that have been prominent in the engineering processes should also be explicitly presented, as well as empirical evidence of compliance and methodology used, such as training data used in predictive systems, algorithms and components used, and results of behavior monitoring. Criteria for such documentation could be: auditability, accessibility, meaningfulness, and readability.

Companies should make their systems auditable and should explore novel methods for external and internal auditing.

Further Resources

- S. Wachter, B. Mittelstadt, and L. Floridi. "[Transparent, Explainable, and Accountable AI for Robotics.](#)" *Science Robotics*, vol. 2, no. 6, May 31, 2017. [Online]. Available: DOI: 10.1126/scirobotics.aan6080. [Accessed Nov. 2017].
- S. Barocas, and A. D. Selbst, "[Big Data's Disparate Impact.](#)" *California Law Review* 104, 671-732, 2016.
- J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. "[Accountable Algorithms.](#)" *University of Pennsylvania Law Review* 165, no. 1, 633–705, 2017.
- J. M. Balkin, "[Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation.](#)" *UC Davis Law Review*, 2017.

Methods to Guide Ethical Research and Design

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Methods to Guide Ethical Research and Design Committee

- **Corinne Cath-Speth** (Co-Chair) – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Raja Chatila** (Co-Chair) – CNRS-Sorbonne Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society
- **Thomas Arnold** – Research Associate at Tufts University Human-Robot Interaction Laboratory
- **Jared Bielby** – President, Netizen Consulting Ltd; Chair, International Center for Information Ethics; editor, *Information Cultures in the Digital Age*
- **Reid Blackman, PhD** – Founder & CEO Virtue Consultants, Assistant Professor of Philosophy Colgate University
- **Tom Guarriello, PhD** – Founding Faculty member in the Master's in Branding program at New York City's School of Visual Arts, Host of RoboPsych Podcast and author of RoboPsych Newsletter
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Jason Millar** – Professor, robot ethics at Carleton University
- **Sarah Spiekermann** – Chair of the Institute for Information Systems & Society at Vienna University of Economics and Business; Author of the textbook "Ethical IT-Innovation", the popular book "Digitale Ethik—Ein Wertesystem für das 21. Jahrhundert" and Blogger on "The Ethical Machine"
- **Shannon Vallor** – William J. Rewak Professor in the Department of Philosophy at Santa Clara University in Silicon Valley and Executive Board member of the Foundation for Responsible Robotics
- **Klein, Wilhelm E. J., PhD** – Senior Research Associate & Lecturer in Technology Ethics, City University of Hong Kong

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

A/IS for Sustainable Development

Autonomous and intelligent systems (A/IS) offer unique and impactful opportunities as well as risks both to people living in high-income countries (HIC) and in low-and middle-income countries (LMIC). The scaling and use of A/IS represent a genuine opportunity across the globe to provide individuals and communities—be they rural, semi-urban, or urban—with the means to satisfy their needs and develop their full potential, with greater autonomy and choice. A/IS will potentially disrupt economic, social, and political relationships and interactions at many levels. Those disruptions could provide an historical opportunity to reset those relationships in order to distribute power and wealth more equitably and thus promote social justice.¹ They could also leverage quality and better standards of life and protect people's dignity, while maintaining cultural diversity and protecting the environment.

One possible vehicle that can be used to agree on priorities and prioritize resources and actions is the United Nations Agenda for Sustainable Development, which was adopted by the UN General Assembly in 2015; 193 nations voted in favor of the Agenda, which also includes 17 Sustainable Development Goals (SDGs) for the world to achieve by 2030. The Agenda challenges all member states to make concerted efforts toward the above mentioned goals, and thus toward a sustainable, prosperous, and resilient future for people and the planet. These universally applicable goals should be reached by 2030.²

The value of A/IS is significantly associated with the generation of various types of superior and unique insights, many of which could help achieve positive socioeconomic outcomes for both HIC and LMIC societies, in keeping with the SDGs. The ethical imperative driving this chapter is that A/IS must be harnessed to benefit humanity, promote equality, and realize the world community's vision of a sustainable future and the SDGs:

.....of universal respect for human rights and human dignity, the rule of law, justice, equality and nondiscrimination; of respect for race, ethnicity and cultural diversity; and of equal opportunity permitting the full realization of human potential and contributing to shared prosperity. A world which invests in its children and in which every child grows up free from violence and exploitation. A world in which every woman and girl enjoys full gender equality and all legal, social and economic barriers to their empowerment have been removed. A just, equitable, tolerant, open and socially inclusive world in which the needs of the most vulnerable are met.³

A/IS for Sustainable Development

We recognize that how A/IS are deployed globally will be a determining factor in whether, in fact, “no-one gets left behind”, whether human rights and dignity of all people are respected, whether children are protected, and whether the gap between rich and poor, within and between nations, narrows or widens. A/IS can advance the Sustainable Development Agenda’s transformative vision, but at the same time, A/IS can undermine it if risks reviewed in this chapter are not managed properly.

For example, A/IS create the risk of accelerating inequality within and among nations, if their development and marketing are controlled by a few select companies, primarily in HIC. The benefits would largely accrue to the highly educated and wealthier segment of the population, while displacing the less educated workforce, both by automation and by the absence of educational or retraining systems capable of imparting skills and knowledge needed to work productively alongside A/IS. These risks, although differentiated by IT infrastructure, educational attainment, economic, and cultural contexts, exist in HIC and LMIC alike. The inequality in accessing and using the internet, both within and among countries, raises questions on how to spread A/IS benefits across humanity. Ensuring A/IS “for the common good” is an ethical imperative and at the core of *Ethically Aligned Design, First Edition*; the key elements of this “common good” are that it is human-centered, accountable, and ensure outcomes that are fair and inclusive.

This chapter explores the imperative for A/IS to serve humanity by improving the quality and standard of life for all people everywhere. It makes recommendations for advancing equal access to this transformative technology, so that it drives the well-being of all people, rather than further concentrating wealth, resources, and decision-making power in the hands of a few countries, companies, or citizens. The recommendations further reflect policies and collaborative public, private, and people programs which, if implemented, will respect the ethical imperative embedded in the Sustainable Development Agenda’s transformative vision. The respect of human rights and dignity, and the advancement of “common good” with equal benefit to both HIC and LMIC, are central to every recommendation within this chapter.

A/IS for Sustainable Development

Section 1—A/IS in Service to Sustainable Development for All

A/IS have the potential to contribute to the resolution of some of the world's most pressing problems, including: violation of fundamental rights, poverty, exploitation, climate change, lack of high-quality services to excluded populations, increased violence, and the achievement of the SDGs.

Issue: Current roadmaps for development and deployment of A/IS are not aligned with or guided by their impact in the most important challenges of humanity, defined in the seventeen United Nations Sustainable Development Goals (SDGs), which collectively aspire to create a more equal world of prosperity, peace, planet protection, and human dignity for all people.⁴

Background

SDGs promoting prosperity, peace, planet protection, human dignity, and respect for human rights of all, apply to HIC and LMIC alike. Yet ensuring that the benefits of A/IS will accrue to humanity as a whole, leaving “no one behind”, requires an ethical commitment to global

citizenship and well-being, and a conscious effort to counter the nature of the tech economy, with its tendency to concentrate wealth within high income populations. Implementation of the SDGs should benefit excluded sectors of society in every country, regardless of A/IS infrastructure.

“The Road to Dignity by 2030” document of the UN Secretary General reports on resources and methods for implementing the 2030 Agenda for Sustainable Development and emphasizes the importance of science, technology, and innovation for a sustainable future.⁵ The UN Secretary General posits that:

“A sustainable future will require that we act now to phase out unsustainable technologies and to invest in innovation and in the development of clean and sound technologies for sustainable development. We must ensure that they are fairly priced, broadly disseminated and fairly absorbed, including to and by developing countries.” (para. 120)

A/IS are among the technologies that can play an important role in the solution of the deep social problems plaguing our global civilization, contributing to the transformation of society away from an unsustainable, unequal socioeconomic system, towards one that realizes the vision of universal human dignity, peace, and prosperity.

However, with all the potential benefits of A/IS, there are also risks. For example, given A/IS technology's immense power needs, without

A/IS for Sustainable Development

new sources of sustainable energy harnessed to power A/IS in the future, there is a risk that it will increase fossil fuel use and have a negative impact on the environment and the climate.

While 45% of the world's population is not connected to the internet, they are not necessarily excluded from A/IS' potential benefits: in LMIC mobile networks can provide data for A/IS applications. However, only those connected are likely to benefit from the income-producing potential of internet technologies. In 2017, internet penetration in HIC left behind certain portions of the population often in rural or remote areas; 12% of U.S. residents and 20% of residents across Europe were unable to access the internet. In Asia with its concentration of LMIC, 52% of the population, on average, had no access, a statistic skewed by the large population of China, where internet penetration reached 45% of the population. In numerous other countries in the region, 99% of residents had no access. This nearly total exclusion also exists in several countries in Africa, where the overall internet penetration is only 35%: 2 of every 3 residents in Africa have no access.⁶ Those with no internet access also do not generate data needed to "train" A/IS, and are thereby excluded from benefits of the technology, the development of which risks systematic discriminatory bias, particularly against people from minority populations, and those living in rural areas, or in low-income countries. As a comparison, one study estimated that "in the US, just one home automation product can generate a data point every six seconds."⁷ In Mozambique, where about 90% of the population lack internet access, "the average household generates zero digital data

points."⁸ With mobile phones generating much of the data needed for developing A/IS applications in LMIC, unequal phone ownership may build in bias. For example, there is a risk of discrimination against women, who across LMIC are 14% less likely than men to own a mobile phone, and in South Asia where 38% are less likely to own a mobile phone.⁹

Recommendations

The current range of A/IS applications in sectors crucial to the SDGs, and to excluded populations everywhere, should be studied, with the strengths, weaknesses, and potential of the most significant recent applications analyzed, and the best ones developed at scale. Specific objectives to consider include:

- Identifying and experimenting with A/IS technologies relevant to the SDGs, such as: big data for development relevant to, for example, agriculture and medical tele-diagnosis; geographic information systems needed in public service planning, disaster prevention, emergency planning, and disease monitoring; control systems used in, for example, naturalizing intelligent cities through energy and traffic control and management of urban agriculture; applications that promote human empathy focused on diminishing violence and exclusion and increasing well-being.
- Promoting the potential role of A/IS in sustainable development by collaboration between national and international government agencies and nongovernmental organizations (NGOs) in technology sectors.

A/IS for Sustainable Development

- Analyzing the cost of and proposing strategies for publicly providing internet access for all, as a means of diminishing the gap in A/IS' potential benefit to humanity, particularly between urban and rural populations in HIC and LMIC alike.
- Investing in the documentation and dissemination of innovative applications of A/IS that advance the resolution of identified societal issues and the SDGs.
- Researching sustainable energy to power A/IS computational capacity.
- Investing in the development of transparent monitoring frameworks to track the concrete results of donations by international organizations, corporations, independent agencies, and the State, to ensure efficiency and accountability in applied A/IS.
- Developing national legal, policy, and fiscal measures to encourage competition in the A/IS domestic markets and the flourishing of scalable A/IS applications.
- Integrating the SDGs into the core of private sector business strategies and adding SDG indicators to companies' key performance indicators, going beyond corporate social responsibility (CSR).
- Applying the well-being indicators¹⁰ to evaluate A/IS' impact from multiple perspectives in HIC and LMIC alike.

Further Resources

- R. Van Est and J.B.A. Gerritsen, with assistance of L. Kool, Human Rights in the Robot Age: Challenges arising from the use of Robots, Artificial Intelligence and Augmented Reality Expert Report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE), The Hague: Rathenau Instituut 2017.
- World Economic Forum Global Future Council on Human Rights 2016-18, "White Paper: How to Prevent Discriminatory Outcomes in Machine Learning," World Economic Forum, March 2018.
- United Nations General Assembly, *Transforming Our World: The 2030 Agenda for Sustainable Development* (A/RES/70/1: 21 October 2015) Preamble. http://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf.
- United Nations Global Pulse, Big Data for Development: Challenges and Opportunities, 2012.

A/IS for Sustainable Development

Issue: A/IS are often viewed only as having impact in market contexts, yet these technologies also have an impact on social relations and culture.

Background

A/IS are expected to have an impact beyond market domains and business models, diffusing throughout the global society. For instance, A/IS have and will impact social relationships in a way similar to how mobile phones changed our daily lives, reflecting directly on our culture, customs, and language. The extent and direction of this impact is not yet clear, but documented experience in HIC and high internet-penetration environments of trolls, “fake news,” and cyberbullying on social media offer a cautionary tale.¹¹ Depression, social isolation, aggression, and the dissemination of violent behavior with damage to human relations, so extreme that, in some cases, it has resulted in suicide, are all correlated with the internet.¹² As an example, the technology for “smart homes” has been used for inflicting domestic violence by remotely locking doors, turning off heat/AC, and otherwise harassing a partner. This problem could be easily extended to include elder and child abuse.¹³ Measures need to be developed to prevent A/IS from contributing to the emergence or amplification of social disorders.

Recommendations

To understand the impact of A/IS on society, it is necessary to consider product and process innovation, as well as wider sociocultural and ethical implications, from a global perspective, including the following:

- Exploring the development of algorithms capable of detecting and reporting discrimination, cyberbullying, deceptive content and identities, etc., and of notifying competent authorities; recognizing that the use of such algorithms must address ethical concerns related to algorithm explainability as well as take into account the risk to certain aspects of human rights, notably to privacy and freedom from oppression.
- Developing a globally recognized professional Code of Ethics with and for technology companies.
- Identifying social disorders, such as depression, anxiety, psychological violence, political manipulation, etc., correlated with the use of A/IS-based technologies as a world health problem; monitoring and measuring their impact.
- Elaborating metrics measuring how, where and on whom there is a cultural impact of new A/IS-based technologies.

A/IS for Sustainable Development

Further Resources

- T. Luong, "Thermostats, Locks and Lights: Digital Tools of Domestic Abuse," *The New York Times*, June 23, 2018, <https://www.nytimes.com/2018/06/23/technology/smart-home-devices-domestic-abuse.html>.
- J. Naughton, "The internet of things has opened up a new frontier of domestic abuse." *The Guardian*, July 2018.
- M. Pianta, *Innovation and Employment, Handbook of Innovation*. Oxford, U.K.: Oxford University Press, 2003.
- M.J. Salganik, *Bit by Bit*. Princeton, NJ: Princeton University Press 2018.
- J. Torresen, "A Review of Future and Ethical Perspectives of Robotics and AI" *Frontiers in Robotics and AI*, Jan. 15, 2018. [Online]. Available: <https://doi.org/10.3389/frobt.2017.00075>. [Accessed Nov. 1, 2018].

Issue: The right to truthful information is key to a democratic society and to achieving sustainable development and a more equal world, but A/IS poses risks to this right that must be managed.

Background

Social media have become the dominant technological infrastructure for the dissemination of information such as news, opinion, advertising,

etc., and are currently in the vanguard of the movement toward customized/targeted information based on user profiling that involves significant use of A/IS techniques. Analysis of opinion polls and trends in social networks, blogs, etc., and of the emotional response to news items can be used for the purposes of manipulation, facilitating both the selection of news that guides public opinion in the desired direction and the practice of sensationalism.

The "personalization of the consumer experience", that is, the adaptation of articles to the interests, political vision, cultural level, education, and geographic location of the reader, is a new challenge for the journalism profession that expands the possibilities of manipulation.

The information infrastructure is currently lacking in transparency, such that it is difficult or impossible to know (except perhaps for the infrastructure operator):

- what private information is being collected for user profiling and by whom,
- which groups are targeted and by whom,
- what information has been received by any given targeted group,
- who financed the creation and dissemination of this information,
- the percentage of the information being disseminated by bots, and
- who is financing these bots.

Many actors have found this opaque infrastructure ideal for spreading politically motivated disinformation, which has a negative

A/IS for Sustainable Development

effect on the creation of a more equal world, democracy, and the respect for fundamental rights. This disinformation can have tragic consequences. For instance, human rights groups have unearthed evidence that the military authorities of Myanmar used Facebook for inciting hatred against the Rohingya Muslim minority, hatred which facilitated an ethnic cleansing campaign and the murder of up to 50,000 people.¹⁴ The UN determined that these actions constituted genocide, crimes against humanity, and war crimes.¹⁵

Recommendations

To protect democracy, respect fundamental rights, and promote sustainable development, governments should implement a legislative agenda which prevents the spread of misinformation and hate speech, by:

- Ensuring more control and transparency in the use of A/IS techniques for user profiling in order to protect privacy and prevent user manipulation.
- Using A/IS techniques to detect untruthful information circulating in the infrastructures, overseen by a democratic body to prevent potential censorship.
- Obliging companies owning A/IS infrastructures to provide more transparency regarding their algorithms, sources of funding, services, and clients.
- Defining a new legal status somewhere between "platforms" and "content providers" for A/IS infrastructures.
- Reformulating the deontological codes of the journalistic profession to take into account the intensive use of A/IS techniques foreseen in the future.
- Promoting the right to information in official documents, and developing A/IS techniques to automate journalistic tasks such as verification of sources and checking the accuracy of the information in official documents, or in the selection, hierarchy, assessment, and development of news, thereby contributing to objectivity and reliability.

Further Resources

- M. Broussard, "Artificial intelligence for Investigative Reporting: Using an expert system to enhance journalists' ability to discover original public affairs stories." *Digital Journalism*, vol. 3, no. 6, pp. 814-831, 2015.
- M. Carlson, "The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority." *Digital Journalism*, vol. 3, no. 3, pp. 416-431, 2015.
- A. López Barriuso, F. de la Prieta Pintado, Á. Lozano Murciego, , D. Hernández de la Iglesia and J. Revuelta Herrero, *JOUR-MAS: A Multi-agent System Approach to Help Journalism Management*, vol. 4, no. 4, 2015.
- P. Mozur, "A Genocide Incited on Facebook with Posts from Myanmar's Military," *The New York Times*, Oct. 15 2018. <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-k-genocide.html>
- UK Parliament, House of Commons, Digital, Culture, Media and Sport Committee Disinformation and 'fake news': Interim Report, Fifth Report of Session 2017–19UK Parliament, Published on July 29, 2018.

A/IS for Sustainable Development

Section 2—Equal Availability

Issue: Vastly different power structures among and within countries create risk that A/IS deployment accelerates, rather than reduces, inequality in the pursuit of a sustainable future. It is unclear how LMIC can best implement A/IS via existing resources and take full advantage of the technology's potential to achieve a sustainable future.

Background

The potential use of A/IS to create sustainable economic growth for LMIC is uniquely powerful. Yet, many of the debates surrounding A/IS take place within HIC, among highly educated and financially secure individuals. It is imperative that all humans, in any condition around the world, are considered in the general development and application of these systems to avoid the risk of bias, excessive inequality, classism, and general rejection of these technologies. With much of the financial and technical resources for A/IS development and deployment residing in HIC, not only are A/IS benefits more difficult to access for LMIC populations, but those A/IS applications that are deployed outside of HIC realities may not be appropriate. This is for reasons of cultural/ethnic bias, language difficulties, or simply an inability to adapt to local internet infrastructure constraints.

Furthermore, technological innovation in LMIC comes up against many potential obstacles, which could be considered when undertaking initiatives aimed at enhancing LMIC access:

- Reluctance to provide open source licensing of technological development innovations,
- Lack of the human capital and knowledge required to adapt HIC-developed technologies to resolving problems in the LMIC context, or to develop local technological solutions to these problems,
- Retention of A/IS capacity in LMIC due to globally uncompetitive salaries,
- Lack of infrastructure for deployment, and difficulties in taking technological solutions to where they are needed,
- Lack of organizational and business models for adapting technologies to the specific needs of different regions,
- Lack of active participation of the target population,
- Lack of political will to allow people to have access to technological resources,
- Existence of oligopolies that hinder new technological development,
- Lack of inclusive and high-quality education at all levels, and
- Bureaucratic policies ill-adapted to highly dynamic scenarios.

A/IS for Sustainable Development

For A/IS capacities and benefits to become equally available worldwide, training, education, and opportunities should be provided particularly for LMIC. Currently, access to products that facilitate A/IS research of timely topics is quite limited for researchers in LMIC, due to cost considerations.

If A/IS capacity and governance problems, such as relevant laws, policies, regulations, and anti-corruption safeguards, are addressed, LMIC could have the ability to use A/IS to transform their economies and leapfrog into a new era of inclusive growth. Indeed, A/IS itself can contribute to good governance when applied to the detection of corruption in state and banking institutions, one of the most serious recognized constraints to investment in LMIC. Particular attention, however, must be paid to ensure that the use of A/IS is for the common good—especially in the context of LMIC—and does not reinforce existing socioeconomic inequities through systematic discriminatory bias in both design and application, or undermine fundamental rights through, among other issues, lax data privacy laws and practice.

Recommendations

A/IS benefits should be equally available to populations in HIC and LMIC, in the interest of universal human dignity, peace, prosperity, and planet protection. Specific measures for LMIC should include:

- Deploying A/IS to detect fraud and corruption, to increase the transparency of power structures, to contribute to a favorable investment, governance, and innovation environment.
- Supporting LMIC in the development of their own A/IS strategies, and in the retention or return of their A/IS talent to prevent “brain drain”.
- Encouraging global standardization/harmonization and open source A/IS software.
- Promoting distribution of knowledge and wealth generated by the latest A/IS, including through formal public policy and financial mechanisms to advance equity worldwide.
- Developing public datasets to facilitate the access of people from LMIC to data resources to facilitate their applied research, while ensuring the protection of personal data.
- Creating A/IS international research centers in every continent, that promote culturally appropriate research, and allow the remote access of LMIC's communities to high-end technology.¹⁶
- Facilitating A/IS access in LMIC through online courses in local languages.
- Ensuring that, along with the use of A/IS, discussions related to identity, platforms, and blockchain are conducted, such that core enabling technologies are designed to meet the economic, social, and cultural needs of LMIC.
- Diminishing the barriers and increase LMIC access to technological products, including the formation of collaborative networks between developers in HIC and LMIC, supporting the latter in attending global A/IS conferences.¹⁷
- Promoting research into A/IS-based technologies, for example, mobile lightweight A/IS applications, that are readily available in LMIC.
- Facilitating A/IS research and development in LMIC through investment incentives, public-

A/IS for Sustainable Development

private partnerships, and/or joint grants, and collaboration between international organizations, government bodies, universities, and research institutes.

- Prioritizing A/IS infrastructure in international development assistance, as necessary to improve the quality and standard of living and advance progress towards the SDGs in LMIC.
- Recognizing data issues that may be particular to LMIC contexts, i.e., insufficient sample size for machine learning which sometimes results in *de facto* discrimination, and inadequate laws for, and the practice of, data protection.
- Supporting research on the adaptation of A/IS methods to scarce data environments and other remedies that facilitate an optimal A/IS enabling environment in LMIC.

Further Resources

- A. Akubue, "Appropriate Technology for Socioeconomic Development in Third World Countries." *The Journal of Technology Studies* 26, no. 1, pp. 33–43, 2000.
- O. Ajakaiye and M. S. Kimenyi. "Higher Education and Economic Development in Africa: Introduction and Overview." *Journal of African Economies* 20, no. 3, iii3–iii13, 2011.
- D. Allison-Hope and M. Hodge, "Artificial Intelligence: A Rights-Based Blueprint for Business," San Francisco: BSF, Aug. 28, 2018
- D. E. Bloom, D. Canning, and K. Chan. *Higher Education and Economic Development in Africa* (Vol. 102). Washington, DC: World Bank, 2006.
- N. Bloom, "Corporations in the Age of Inequality." *Harvard Business Review*, April 21, 2017.
- C. Dahlman, *Technology, Globalization, and Competitiveness: Challenges for Developing Countries. Industrialization in the 21st Century*. New York: United Nations, 2006.
- M. Fong, *Technology Leapfrogging for Developing Countries. Encyclopedia of Information Science and Technology*, 2nd ed. Hershey, PA: IGI Global, 2009 (pp. 3707–3713).
- C. B. Frey and M. A. Osborne. "The Future of Employment: How Susceptible Are Jobs to Computerisation?" (working paper). Oxford, U.K.: Oxford University, 2013.
- B. Hazeltine and C. Bull. *Appropriate Technology: Tools, Choices, and Implications*. New York: Academic Press, 1999.
- McKinsey Global Institute. "Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy" (report), May 2013.
- D. Rotman, "How Technology Is Destroying Jobs." *MIT Technology Review*, June 12, 2013.
- R. Sauter and J. Watson. "Technology Leapfrogging: A Review of the Evidence, A Report for DFID." Brighton, England: University of Sussex. October 3, 2008.
- "The Rich and the Rest." *The Economist*. October 13, 2012.
- "Wealth without Workers, Workers without Wealth." *The Economist*. October 4, 2014.
- World Bank. "Global Economic Prospects 2008: Technology Diffusion in the Developing World." Washington, DC: World Bank, 2008.
- World Development Report 2016: Digital Dividends. Washington, DC: World Bank. doi:10.1596/978-1-4648-0671-1.
- World Wide Web Foundation "Artificial Intelligence: The Road ahead in Low and Middle-income Countries," webfoundation.org, June 2017.

A/IS for Sustainable Development

Section 3—A/IS and Employment

Issue: A/IS are changing the nature of work, disrupting employment, while technological change is happening too fast for existing methods of (re)training the workforce.

Background

The current pace of technological development will heavily influence changes in employment structure. In order to properly prepare the workforce for such evolution, actions should be proactive and not only reactive. The wave of automation caused by the A/IS revolution will displace a very large share of jobs across domains and value chains. The U.S. “automated vehicle” case study analyzed in the White House 2016 report *Artificial Intelligence, Automation, and the Economy* is emblematic of what is at stake: “2.2 to 3.1 million existing part- and full-time U.S. jobs are exposed over the next two decades, although the timeline remains uncertain.”¹⁸

The risk of unemployment for LMIC is more serious than for developed countries. The industry of most LMIC is labor intensive. While labor may be cheap(er) in LMIC economies, the ripple effects of A/IS and automation will be felt much more than in the HIC economies. The 2016 World Bank Development Report stated that the share of occupations susceptible to automation and A/IS is higher in LMIC than in

HIC, where such jobs have already disappeared. In addition, the qualities which made certain jobs easy to outsource to LMIC where wages are lower are those that may make them easy to automate.¹⁹ An offsetting factor is the reality that many LMIC lack the communication, energy, and IT infrastructure required to support highly automated industries.²⁰ Notwithstanding this reality, the World Bank estimated the automatable share of employment, unadjusted for adoption time lag, for LMIC ranges from 85% in Ethiopia to 62% in Argentina, compared to the OECD average of 57%.²¹

In the coming decades, the automation wave calls for higher investment and the transformation of labor market capacity development programs. Innovative and fair ways of funding such an investment are required; the solutions should be designed in cooperation with the companies benefiting from the increase of profitability, thanks to automation. This should be done in a responsible way so that the innovation cycle is not broken, and yet workforce capacity does not fall behind the needs of 21st century employment. At the same time, A/IS and other digital technologies offer real potential to innovate new approaches to job-search assistance, placement, and hiring processes in the age of personalized services. The efficiency of matching labor supply and demand can be tremendously enhanced by the rise of multisided platforms and predictive analytics, provided they do not entrench discrimination.²² The case of platforms, such as LinkedIn, for instance, with its 470 million

A/IS for Sustainable Development

registered users, and online job consolidators such as indeed.com and Simply Hired, are interesting as an evolution in hiring practices, at least for those able to access the internet.

Tailored counseling and integrated retraining programs also represent promising grounds for innovation. In addition, much will have to be done to create fair and effective lifelong skill development/training, infrastructures, and mechanisms capable of empowering millions of people to viably transition jobs, sectors, and potentially locations, and to address differential geographic impacts that exacerbate income and wealth disparities. Effectively enabling the workforce to be more mobile—physically, legally, and virtually—will be crucial. This implies systemic policy approaches which encompass housing, transportation, licensing, tax incentives, and crucially in the age of A/IS, universal broadband access, especially in rural areas of both HIC and LMIC.

Recommendations

To thrive in the A/IS age, workers must be provided training in skills that improve their adaptability to rapid technological changes; programs should be available to any worker, with special attention to the low-skilled workforce. Those programs can be private, that is, sponsored by the employer, or publicly and freely offered through specific public channels and government policies, and should be available regardless of whether the worker is in between jobs or still employed. Specific measures include:

- Offering new technical programs, possibly earlier than high school, to increase the workforce capacity to close the skills gap and thrive in employment alongside A/IS.
- Creating opportunities for apprenticeships, pilot programs, and scaling up data-driven evidence-based solutions that increase employment and earnings.
- Supporting new forms of public-private partnerships involving civil society, as well as new outcome-oriented financial mechanisms, e.g., social impact bonds, that help scale up successful innovations.
- Supporting partnerships between universities, innovation labs in corporations, and governments to research and incubate startups for A/IS graduates.²³
- Developing regulations to hold corporations responsible for employee retraining necessary due to increased automation and other technological applications having impact on the workforce.
- Facilitating private sector initiatives by public policy for co-investment in training and retraining programs through tax incentives.
- Establishing and resourcing public policies that assure the survival and well-being of workers, displaced by A/IS and automation, who cannot be retrained.
- Researching complementary areas, to lay solid foundations for the transformation outlined above.
 - Requiring more policy research on the dynamics of professional transitions in different labor market conditions.

A/IS for Sustainable Development

- Researching the fairest and most efficient public-private options for financing labor force transformation due to A/IS.
- Developing national and regional future of work strategies based on sound research and strategic foresight.

Further Resources

- V. Cerf and D. Norfors, *The People-centered Economy: The New Ecosystem for Work*. California: IIIJ Foundation, 2018.
- Executive Office of the President. *Artificial Intelligence, Automation, and the Economy*. December 20, 2016.
- S. Kilcarr, "Defining the American Dream for Trucking ... and the Nation, Too," *FleetOwner*, April 26, 2016.
- M. Mason, "Millions of Californians' Jobs could be Affected by Automation—a Scenario the next Governor has to Address," *Los Angeles Times*, October 14, 2018.
- OECD, "Labor Market Programs: Expenditure and Participants," *OECD Employment and Labor Market Statistics* (database), 2016.
- M. Vivarelli, "Innovation and Employment: A Survey," Institute for the Study of Labor (IZA) Discussion Paper No. 2621, February 2007.

Issue: Analysis of the A/IS impact on employment is too focused on the number and category of jobs affected, whereas more attention should be addressed to the complexities of changing the task content of jobs.

Background

Current attention on automation and employment tends to focus on the sheer number of jobs lost or gained. It is important to focus the analysis on how employment structures will be changed by A/IS, rather than solely dwelling on the number of jobs that might be impacted. For example, rather than carrying out a task themselves, workers will need to shift to supervision of robots performing that task. Other concerns include changes in traditional employment structures, with an increase in flexible, contract-based temporary jobs, without employee protection, and a shift in task composition away from routine/repetitive and toward complex decision-making. This is in addition to the enormous need for the aforementioned retraining. Given the extent of disruption, workforce trends will need to measure time spent unemployed or underemployed, labor force participation rates, and other factors beyond simple unemployment numbers.

A/IS for Sustainable Development

The *Future of Jobs 2018* report of the World Economic Forum highlights:

"...the potential of new technologies to create as well as disrupt jobs and to improve the quality and productivity of the existing work of human employees. Our findings indicate that, by 2022, *augmentation* of existing jobs through technology may free up workers from the majority of data processing and information search tasks—and may also increasingly support them in high-value tasks such as reasoning and decision-making as

augmentation becomes increasingly common over the coming years as a way to supplement and complement human labour."²⁴

The report predicts the shift in skill demand between today and 2022 will be significant and that "proactive, strategic and targeted efforts will be needed to map and incentivize workforce redeployment... [and therefore]... investment decisions [on] whether to prioritize automation or augmentation and the question of whether or not to invest in workforce reskilling."²⁵

Comparing Skills Demand, 2018 Versus 2022, Top Ten

TODAY, 2018	TRENDING, 2022	DECLINING, 2022
<ol style="list-style-type: none"> 1. Analytical thinking and innovation 2. Complex problem-solving 3. Critical thinking and analysis 4. Active learning and learning strategies 5. Creativity, originality, and initiative 6. Attention to detail, trustworthiness 7. Emotional Intelligence 8. Reasoning, problem-solving, and ideation 9. Leadership and social influence 10. Coordination and time management 	<ol style="list-style-type: none"> 1. Analytical thinking and innovation 2. Active learning and learning strategies 3. Creativity, originality, and initiative 4. Technology design and programming 5. Critical thinking and analysis 6. Complex problem-solving 7. Leadership and social influence 8. Emotional intelligence 9. Reasoning, problem-solving, and ideation 10. Systems analysis and evaluation 	<ol style="list-style-type: none"> 1. Manual dexterity, endurance, and precision 2. Memory, verbal, auditory, and spatial abilities 3. Management of financial and material resources 4. Technology installation and maintenance 5. Reading, writing, math, and active listening 6. Management of personnel 7. Quality control and safety awareness 8. Coordination and time-management 9. Visual, auditory, and speech abilities 10. Technology use, monitoring, and control

Source: Future of Jobs Survey 2018, World Economic Forum, Table 4

A/IS for Sustainable Development

Recommendations

While there is evidence that robots and automation are taking jobs away in various sectors, a more balanced, granular, analytical, and objective treatment of A/IS impact on the workforce is needed to effectively inform policy making and essential workforce reskilling. Specifics to accomplish this include:

- Creating an international and independent agency able to properly disseminate objective statistics and inform the media, as well as the general public, about the impact of robotics and A/IS on jobs, tax revenue, growth,²⁶ and well-being.
- Analyzing and disseminating data on how current task content of jobs have changed, based on a clear assessment of the automatability of the occupational description of such jobs.
- Promoting automation with augmentation, as recommended in the *Future of Jobs Report 2018* ([see chart on page 154](#)), to maximize the benefit of A/IS to employment and meaningful work.
- Integrating more granulated dynamic mapping of the future jobs, tasks, activities, workplace-structures, associated work-habits, and skills base spurred by the A/IS revolution, in order to innovate, align, and synchronize skill development and training programs with future requirements. This workforce mapping is needed at the macro, but also crucially at the micro, levels where labor market programs are deployed.
- Considering both product and process innovation, and looking at them from a global perspective in order to understand properly the global impact of A/IS on employment.
- Proposing mechanisms for redistribution of productivity increases and developing an adaptation plan for the evolving labor market.

Further Resources

- E. Brynjolfsson and A. McAfee. *The Second Age of Machine Intelligence: Work Progress and Prosperity in a Time of Brilliant Technologies*. New York, NY: W. W. Norton & Company, 2014.
- P.R. Daugherty, and H.J. Wilson, *Human + Machine: Reimagining Work in the Age of AI*. Watertown, MA: Harvard Business Review Press, 2018.
- International Federation of Robotics. "The Impact of Robots on Productivity, Employment and Jobs," A positioning paper by the International Federation of Robotics, April 2017.
- RockEU. "Robotics Coordination Action for Europe Report on Robotics and Employment," Deliverable D3.4.1, June 30, 2016.
- World Economic Forum, Centre for the New Economy and Society, *The Future of Jobs 2018*, Geneva: WEF 2018.

A/IS for Sustainable Development

Section 4—Education for the A/IS Age

Issue: Education to prepare the future workforce, in both HIC and LMIC, to design ethical A/IS applications or to have a comparative advantage in working alongside A/IS, is either lacking or unevenly available, risking inequality perpetuated across generations, within and between countries, constraining equitable growth, supporting a sustainable future, and achievement of the SDGs.

Background

Multiple international institutions, in particular educational engineering organizations,²⁷ have called on universities to play an active role, both locally and globally, in the resolution of the enormous problems that the world faces in securing peace, prosperity, planet protection, and universal human dignity: armed conflict, social injustice, rapid climate change, abuse of human rights, etc. Addressing global social problems is one of the central objectives of many universities, transversal to their other functions, including research in A/IS. UNESCO points out that universities' preparation of future scientists and engineers for social responsibility is presently

very limited, in view of the enormous ethical and social problems associated with technology.²⁸ Enhancing the global dimension of engineering in undergraduate and postgraduate A/IS education is necessary, so that students can be prepared as technical professionals, aware of the opportunities and risks that A/IS present, and ready for work anywhere in the world in any sector.

Engineering studies at the university and postgraduate levels is just one dimension of the A/IS education challenge. For instance, business, law, public policy, and medical students will also need to be prepared for professions where A/IS are a partner, and to have internalized ethical principles to guide the deployment of such technologies. LMIC need financial and academic support to incorporate global A/IS professional curricula in their own universities, and all countries need to develop the pipeline by preparing elementary and secondary school students to access such professional programs. While the need for curriculum reform is recognized, the impact of A/IS on various professions and socioeconomic contexts is, at this time, both evolving and largely undocumented. Thus, the overhaul of education systems at all levels should be preceded by A/IS research.

Much of LMIC education is not globally competitive today, so there is a risk that the global advent of A/IS could negatively affect the chances of young people in LMIC finding

A/IS for Sustainable Development

productive employment, further fueling global inequality. Education systems worldwide have to be reformed and transformed to fit the new demands of the information age, in view of the changing mix of skills demanded from the workforce.²⁹ In 21st century education, it has been observed that children need less rote knowledge, given so much is instantly accessible on the web and more tools to network and innovate are available; less memory and more imagination should be developed; and fewer physical books and more internet access is required. Young people everywhere need to develop their capacities for creativity, human empathy, ethics, and systems thinking in order to work productively alongside robots and A/IS technologies. Science, Technology, Engineering, Art/design, and Math (STEAM) subjects need to be more extensive and more creatively taught.³⁰ In addition, research is needed to establish ways that a new subject, empathy, can be added to these crucial 21st century subjects in order to educate the future A/IS workforce in social skills. Instead, in rich and poor countries alike, children are continuing to be educated for an industrial age which has disappeared or never even arrived. LMIC education systems, being less entrenched in many countries, may have the potential to be more flexible than those in HIC. Perhaps A/IS can be harnessed to help educational systems to leapfrog into the 21st century, just as mobile phone technology enabled LMIC leapfrog over the phase of wired communication infrastructure.

Recommendations

Education with respect to A/IS must be targeted to three sets of students: the general public, present and future professionals in A/IS, and present and future policy makers. To prepare the future workforce to develop culturally appropriate A/IS, to work productively and ethically alongside such technologies, and to advance the UN SDGs, the curricula in HIC and LMIC universities and professional schools require innovation. Equally importantly, preuniversity education systems, starting with early childhood education, need to be reformed to prepare society for the risks and opportunities of the A/IS age, rather than the current system which prepares society for work in an industrial age that ended with the 20th century. Specific recommendations include:

- Preparing future managers, lawyers, engineers, civil servants, and entrepreneurs to work productively and ethically as global citizens alongside A/IS, through reform of undergraduate and graduate curricula as well as of preschool, primary, and secondary school curricula. This will require:
- Fomenting interaction between universities and other actors such as companies, governments, NGOs, etc., with respect to A/IS research through definition of research priorities and joint projects, subcontracts to universities, participation in observatories, and co-creation of curricula, cooperative teaching, internships/service learning, and conferences/seminars/courses.
- Establishing and supporting more multidisciplinary degrees that include

A/IS for Sustainable Development

A/IS, and adapting university curricula to provide a broad, integrated perspective which allows students to understand the impact of A/IS in the global, economic, environmental, and sociocultural domains and trains them as future policy makers in A/IS fields.

- Integrating the teaching of ethics and A/IS across the education spectrum, from preschool to postgraduate curricula, instead of relegating ethics to a standalone module with little direct practical application.
- Promoting service learning opportunities that allow A/IS undergraduate and graduate students to apply their knowledge to meet the needs of a community.
- Creating international exchange programs, through both private and public institutions, which expose students to different cultural contexts for A/IS applications in both HIC and LMIC.
- Creating experimental curricula to prepare people for information-based work in the 21st century, from preschool through postgraduate education.
- Taking into account transversal competencies students need to acquire to become ethical global citizens, i.e., critical thinking, empathy, sociocultural awareness, flexibility, and deontological reasoning in the planning and assessment of A/IS curricula.
- Training teachers in teaching methodologies suited to addressing challenges imposed in the age of A/IS.
- Stimulating STEAM courses in preuniversity education.
- Encouraging high-quality HIC-LMIC collaborative A/IS research in both private and public universities.
- Conducting research to support innovation in education and business for the A/IS world, which could include:
 - Researching the impact of A/IS on the governance and macro/micro strategies of companies and organizations, together with those companies, in an interdisciplinary manner which harnesses expertise of both social scientists and technology experts.
 - Researching the impact of A/IS on the business model for the development of new products and services through the collaborative efforts of management, operations, and the technical research and development function.
 - Researching how empathy can be taught and integrated into curricula, starting at the preschool level.
 - Researching how schools and education systems in low-income settings of both HIC and LMIC can leverage their less-entrenched interests to leapfrog into a 21st century-ready education system.

A/IS for Sustainable Development

- Establishing ethics observatories in universities with the purpose of fostering an informed public opinion capable of participating in policy decisions regarding the ethics and social impact of A/IS applications.
- Creating professional continuing education and employment opportunities in A/IS for current professionals, including through online and executive education courses.
- Creating educative mass media campaigns to elevate society's ongoing baseline level of understanding of A/IS systems, including what it is, if and how it can be trusted in various contexts, and what are its limitations.

Further Resources

- ABET Computing and Engineering Accreditation Criteria 2018. Available at: <http://www.abet.org/accreditation/accreditation-criteria/>
- ABET, 2017 ABET Impact Report, Working Together for a Sustainable Future, 2017.
- emlyon business school, Artificial Intelligence in Management (AIM) Institute <http://aim.em-lyon.com>
- UNESCO, *The UN Decade of Education for Sustainable Development, Shaping the Education of Tomorrow*. UNESCO 2012.

A/IS for Sustainable Development

Section 5—A/IS and Humanitarian Action

Issue: A/IS are contributing to humanitarian action to save lives, alleviate suffering, and maintain human dignity both during and in the aftermath of man-made crises and natural disasters, as well as to prevent and strengthen preparedness for the occurrence of such situations. However, there are ethical concerns with both the collection and use of data during humanitarian emergencies.

Background

There have been a number of promising A/IS applications that relieve suffering in humanitarian crises, such as extending the reach of the health system by using drones to deliver blood to remote parts of Rwanda,³¹ locating and removing landmines,³² efforts to use A/IS to track movements and population survival needs following a natural disaster, and to meet the multiple management requirements of refugee camps.³³ There are also promising developments using A/IS and robotics to assist people with disabilities to recover mobility, and robots to rescue people trapped in collapsed buildings.³⁴ A/IS are also being used to monitor

conflict zones and to enable early warning systems.³⁵ For example, Microsoft has partnered with the UN Human Rights Office of the High Commissioner (OHCHR) to use big data in order to track and analyze human rights violations in conflict zones.³⁶ Machine learning is being used for improved decision-making regarding asylum adjudication and refugee resettlement, with a view to increasing successful integration between refugees and host communities.³⁷ In addition, there is evidence that a recent growth in human empathy has increased well-being while diminishing psychological and physical violence,³⁸ inspiring some researchers to look for ways of harnessing the power of A/IS to introduce more empathy and less violence into society.

The design and ethical deployment of these technologies in crisis settings are both essential and challenging. Large volumes of both personally identifiable and demographically identifiable data are collected in fragile environments, where tracking of individuals or groups may compromise their security if data privacy cannot be assured. Consent to data use is also impractical in such environments, yet crucial for the respect of human rights.

A/IS for Sustainable Development

Recommendations

The potential for A/IS to contribute to humanitarian action to save and improve lives should be prioritized for research and development, including by organizing global research challenges, while also building in safeguards to protect the creation, collection, processing, sharing, use, and disposal of information, including data from and about individuals and populations. Specific recommendations include:

- Promoting awareness of the vulnerable condition of certain communities around the globe and the need to develop and use A/IS applications for humanitarian purposes.
- Elaborating competitions and challenges in high impact conferences and university hackathons to engage both technical and nontechnical communities in the development of A/IS for humanitarian purposes and to address social issues.
- Support civil society groups who organize themselves for the purpose of A/IS research and advocacy to develop applications to benefit humanitarian causes.³⁹
- Developing and applying ethical standards for the collection, use, sharing, and disposal of data in fragile settings.
- Following privacy protection frameworks for pressing humanitarian situations that ensure the most vulnerable are protected.⁴⁰
- Setting up clear ethical frameworks for exceptional use of A/IS technologies in life-saving humanitarian situations, compared to "normal" situations.⁴¹
- Stimulating the development of low-cost and open source solutions based on A/IS to address specific humanitarian problems.
- Training A/IS experts in humanitarian action and norms, and humanitarian practitioners to catalyze collaboration in designing, piloting, developing, and implementing A/IS technologies for humanitarian purposes. Forging public-private A/IS participant alliances that develop crisis scenarios in advance.
- Working on cultural and contextual acceptance of any A/IS introduced during emergencies.
- Documenting and developing quantifiable metrics for evaluating the outcomes of humanitarian digital projects, and educating the humanitarian ecosystem on the same.

A/IS for Sustainable Development

Further Resources

- E. Prestes et al., "The 2016 Humanitarian Robotics and Automation Technology Challenge [Competitions]," in *IEEE Robotics & Automation Magazine*, vol. 23, no. 3, pp. 23-24, Sept. 2016. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7565695&isnumber=7565655>
- L. Marques et al., "Automation of humanitarian demining: The 2016 Humanitarian Robotics and Automation Technology Challenge," *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, Kollam, 2016, pp. 1-7. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7931893&isnumber=7931858>
- CYBATHLON 2020 Preliminary Race Task Descriptions <http://www.cyathlon.ethz.ch/cyathlon-2020/preliminary-race-task-descriptions.html>
- CYBATHLON Scientific Publications <http://www.cyathlon.ethz.ch/>
- Immigration Policy Lab (IPL), "Harnessing Big Data to Improve Refugee Resettlement" <https://immigrationlab.org/project/harnessing-big-data-to-improve-refugee-resettlement/>
- Harvard Humanitarian Initiative, *The Signal Code*, <https://signalcode.org>
- J.A. Quinn, et al., "Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping" *Philosophical Transactions of the Royal Society A*, 376 20170363; DOI: 10.1098/rsta.2017.0363. Aug. 6, 2018.
- Humanitarian Innovation Guide: <https://higuide.elrha.org/>, 2019.
- P. Meier, *Digital Humanitarians: How Big Data is Changing the Face of Humanitarian Response*. Florida: CRC Press, 2015.
- "Technology for human rights: UN Human Rights Office announces landmark partnership with Microsoft" <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21620&LangID=E>
- M. Luengo-Oroz, "10 big data science challenges facing humanitarian organizations," UNHCR, Nov. 22, 2016. <http://www.unhcr.org/innovation/10-big-data-science-challenges-facing-humanitarian-organizations/>
- Optic Technologies, Press Release, Vatican Hack 2018—Results, 18 March 2018, which announced winning AI applications to benefit migrants and refugees as well as social inclusion and interfaith dialogue, <http://opticttechnology.org/index.php/en/news-en/151-vhack-2018winners-en>

A/IS for Sustainable Development

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The A/IS for Sustainable Development Committee

- **Elizabeth D. Gibbons** (Chair) – Senior Fellow and Director of the Child Protection Certificate Program, FXB Center for Health and Human Rights, Harvard T.H. Chan School of Public Health
- **Kay Firth-Butterfield** (Founding Co-Chair) – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Raj Madhavan** (Founding Co-Chair) – Founder & CEO of Humanitarian Robotics Technologies, LLC, Maryland, U.S.A.
- **Ronald C. Arkin** – Regents' Professor & Director of the Mobile Robot Laboratory; Associate Dean for Research & Space Planning, College of Computing, Georgia Institute of Technology
- **Joanna J. Bryson** – Reader (Associate Professor), University of Bath, Intelligent Systems Research Group, Department of Computer Science
- **Renaud Champion** – Director of Emerging Intelligences, emlyon business school; Founder of Robolution Capital & CEO of PRIMNEXT
- **Chandramauli Chaudhuri** – Senior Data Scientist; Fractal Analytics
- **Rozita Dara** – Assistant Professor, Principal Investigator of Data Management and Data Governance program, School of Computer Science, University of Guelph, Canada
- **Scott L. David** – Director of Policy at University of Washington—Center for Data Management and Privacy Governance LabInformation Assurance and Cybersecurity
- **Jia He** – Executive Director of Toutiao Research (Think Tank), Bytedance Inc.
- **William Hoffman** – Associate director and head of Data-Driven Development, The World Economic Forum
- **Michael Lennon** – Senior Fellow, Center for Excellence in Public Leadership, George Washington University; Co-Founder, Govpreneur.org; Principal, CAIPP.org (Consortium for Action Intelligence and Positive Performance); Member, Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems Committee
- **Miguel Luengo-Oroz** – Chief Data Scientist, United Nations Global Pulse.

A/IS for Sustainable Development

- **Angeles Manjarrés** – Professor of the Department of Artificial Intelligence of the Spanish National Distance-Learning University
- **Nicolas Mialhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po). Visiting Professor, IE School of Global and Public Affairs
- **Roya Pakzad** – Research Associate and Project Leader in Technology and Human Rights, Global Digital Policy Incubator (GDPI), Stanford University
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow
- **Simon Pickin** – Professor, Dpto. de Sistemas Informáticos y Computación, Facultad de Informática, Universidad Complutense de Madrid, Spain
- **Rose Shuman** – Partner at BrightFront Group & Founder, Question Box
- **Hruy Tsegaye** – One of the founders of iCog Labs; a pioneer company in East Africa to work on Research and Development of Artificial General Intelligence, Ethiopia

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

A/IS for Sustainable Development

Endnotes

¹ See, for example, the writing of T. Piketty, *Capital in the Twenty-First Century* (Cambridge: Belknap Press 2014).

² See preamble of the United Nations General Assembly, *Transforming our world: the 2030 Agenda for Sustainable Development* (A/RES/70/1: 21 October 2015): "This Agenda is a plan of action for people, planet and prosperity. It also seeks to strengthen universal peace in larger freedom. We recognize that eradicating poverty in all its forms and dimensions, including extreme poverty, is the greatest global challenge and an indispensable requirement for sustainable development. All countries and all stakeholders, acting in collaborative partnership, will implement this plan. We are resolved to free the human race from the tyranny of poverty and want and to heal and secure our planet. We are determined to take the bold and transformative steps which are urgently needed to shift the world on to a sustainable and resilient path. As we embark on this collective journey, we pledge that no one will be left behind. The 17 Sustainable Development Goals and 169 targets which we are announcing today demonstrate the scale and ambition of this new universal Agenda."

³ Ibid, paragraph 8.

⁴ A/IS has the potential to advance positive change toward all seventeen 2030 Sustainable Development Goals, which are:

Goal 1. End poverty in all its forms everywhere

Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture

Goal 3. Ensure healthy lives and promote well-being for all at all ages

Goal 4. Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all

Goal 5. Achieve gender equality and empower all women and girls

Goal 6. Ensure availability and sustainable management of water and sanitation for all

Goal 7. Ensure access to affordable, reliable, sustainable and modern energy for all

Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

Goal 10. Reduce inequality within and among countries

Goal 11. Make cities and human settlements inclusive, safe, resilient and sustainable

Goal 12. Ensure sustainable consumption and production patterns

Goal 13. Take urgent action to combat climate change and its impacts

A/IS for Sustainable Development

Goal 14. Conserve and sustainably use the oceans, seas and marine resources for sustainable development

Goal 15. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss

Goal 16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

Goal 17. Strengthen the means of implementation and revitalize the global partnership for sustainable development

Source: United Nations General Assembly, Transforming our world: the 2030 Agenda for Sustainable Development (A/RES/70/1: 21 October 2015) p. 14

⁵ United Nations Secretary General “The road to dignity by 2030: ending poverty, transforming all lives and protecting the planet” United Nations, A/69/700, 4 December 2014, pp. 25-27 http://www.un.org/ga/search/view_doc.asp?symbol=A/69/700&Lang=E

⁶ Internet World Stats <https://www.internetworldstats.com/stats.htm>, accessed 17 May 2018.

⁷ (“Internet of Things, Privacy and Security in a Connected World,” FTC, <https://www.ftc.gov/system/les/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf>)

⁸ World Economic Forum Global Future Council on Human Rights 2016-18 “White Paper: How to Prevent Discriminatory Outcomes in Machine Learning” (WEF: March 2018).

⁹ World Wide Web Foundation *Artificial Intelligence: the Road ahead in Low and Middle-income Countries* (June 2017: webfoundation.org) p.13

¹⁰ See the Well-being chapter of *Ethically Aligned Design*, First Edition

¹¹ See, for example, S. Vosougi, D. Roy, and S. Aral, “The spread of true and false news online” *Science* 09 Mar 2018: Vol. 359, Issue 6380, pp. 1146-1151 and M. Fox, “Fake News: Lies spread faster on social media than Truth does” *NBC Health News*, 8 March 2018 <https://www.nbcnews.com/health/health-news/fake-news-lies-spread-faster-social-media-truth-does-n854896>; Cyberbullying Research Center: Summary of Cyberbullying Research 2004-2016 <https://cyberbullying.org/summary-of-our-cyberbullying-research> and TeenSafe “Cyberbullying Facts and Statistics” TeenSafe October 4, 2016, <https://www.teensafe.com/blog/cyber-bullying-facts-and-statistics/>

A. Hutchison, “Social Media Still Has a Fake News Problem and Digital Literacy is Largely to Blame” *Social Media Today*, October 5, 2018 <https://www.socialmediatoday.com/news/social-media-still-has-a-fake-news-problem-and-digital-literacy-is-largel/538930/>; D.D. Luxton, J.D. June, and J. M. Fairall, “Social Media and Suicide: A Public Health Perspective”, *Am J Public Health*. 2012 May; 102(Suppl 2): S195–S200. J. Twege, T. E. Joiner, M.L. Rogers, “Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links

A/IS for Sustainable Development

to Increased New Media Screen Time" Clinical Psychological Science, November 14, 2017 <https://doi.org/10.1177/2167702617723376>

¹² D.D. Luxton, J.D. June, and J. M. Fairall, "Social Media and Suicide: A Public Health Perspective", Am J Public Health. 2012 May; 102(Suppl 2): S195–S200. J. Twege, T. E. Joiner, M.L. Rogers, "Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time" Clinical Psychological Science, November 14, 2017 <https://doi.org/10.1177/2167702617723376>

¹³ T. Luong, "Thermostats, Locks and Lights: Digital Tools of Domestic Abuse." *The New York Times*, June 23, 2018, <https://www.nytimes.com/2018/06/23/technology/smart-home-devices-domestic-abuse.html>

¹⁴ P. Mozur, "A Genocide incited on Facebook with posts from Myanmar's Military", *The New York Times*, October 15, 2018. <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>

¹⁵ United Nations Human Rights Council "Human rights situations that require the Council's attention Report of the independent international fact-finding mission on Myanmar*" (A/HRC/39/64, 12 September 2018)

¹⁶ See for example Google AI in Ghana <https://www.blog.google/around-the-globe/google-africa/google-ai-ghana/>

¹⁷ See *Artificial Intelligence: the Road ahead in Low and Middle-income Countries*

¹⁸ Executive Office of the President of the United States. *Artificial Intelligence, Automation, and the Economy*. December 20, 2016. page 21.

¹⁹ From World Wide Web Foundation *Artificial Intelligence: The Road ahead in Low and Middle-income Countries* (June 2017: webfoundation.org) page 8.

²⁰ Ibid.

²¹ World Bank, 2016. *World Development Report 2016: Digital Dividends*. Washington, DC: World Bank. doi:10.1596/978-1-4648-0671-1 page 129.

²² See for example: J. Dasten, "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters Business News October 9, 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08C>

²³ For example, The Vector Institute, CIFAR and the Legal Innovation Group at Ryerson University. See <https://vectorinstitute.ai> and <http://www.legalinnovationzone.ca>.

²⁴ World Economic Forum, Centre for the New Economy and Society *the Future of Jobs 2018* (Geneva: WEF 2018) p. 3.

²⁵ Ibid, page 9

²⁶ It must be noted that the OECD is already engaged in this work as well as are some government bodies. See <http://www.oecd.org/employment/future-of-work/>

A/IS for Sustainable Development

²⁷ UNESCO, WHO, ABET, Bologna Follow-Up Group Secretariat for the European Higher Education Area

²⁸ UNESCO, The UN Decade of Education for Sustainable Development, Shaping the Education of Tomorrow. (UNESCO: Paris 2012).

²⁹ See *Future of Jobs Report 2018 Survey* table, p. 154.

³⁰ National Math and Science Initiative, STEM Education and Workforce, 2014 <https://www.nms.org/Portals/0/Docs/STEM%20Crisis%20Page%20Stats%20and%20References.pdf>

³¹ <https://www.bloomberg.com/news/articles/2018-08-16/this-27-year-old-launches-drones-that-deliver-blood-to-rwanda-s-hospitals>

³² <https://www.theguardian.com/sustainable-business/2015/may/25/robots-rescue-lethal-rehabilitation-landmines-drones>

³³ See for example, C. Fey, "Tech can improve lives in refugee camps" Cambridge Network, 10 May 2018 <https://www.cambridgenetwork.co.uk/news/tech-can-improve-lives-in-refugee-camps/>; <https://github.com/qcri-social/AIDR/wiki/AIDR-Overview>

³⁴ <https://www.sciencemag.org/news/2017/10/searching-survivors-mexico-earthquake-snake-robots>

<https://www.livescience.com/48473-search-and-rescue-robot-algorithm.html>

³⁵ <http://focus.barcelonagse.eu/can-machine-learning-help-policy-makers-detect-conflict/>

<https://www.worldbank.org/en/news/press-release/2018/09/23/united-nations-world-bank-humanitarian-organizations-launch-innovative-partnership-to-end-famine>

³⁶ "United Nations Human Rights Office of the High Commissioner, press release, "Technology for human rights: UN Human Rights Office announces landmark partnership with Microsoft" 16 May 2017." <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21620&LangID=E>

³⁷ For example, researchers at Stanford University are running a pilot project to develop machine learning algorithms for a better resettlement program. To train their algorithm, the Immigration Policy Lab (IPL) at Stanford University and ETH Zurich gathered data from refugee resettlement agencies in the US and Switzerland. The model is optimized based on refugees' background and skill sets to match them to a host city in which the individual has a higher chance of finding employment.

³⁸ See for example S. Pinker, *The Better Angels of Our Nature: Why Violence has Declined* (Penguin 2012) and R. Krznaric, *Empathy: How it matters and how to get it.* (Perigee 2015).

³⁹ See for example TechToronto: <https://www.techtoronto.org> and #AI and Big Data

⁴⁰ See for example Harvard Humanitarian Initiative Signal Code <https://signalcode.org>

⁴¹ See Humanitarian Innovation Guide: <https://higuide.elrha.org/>

Embedding Values into Autonomous and Intelligent Systems

Society has not established universal standards or guiding principles for embedding human values and norms into autonomous and intelligent systems (A/IS) today. But as these systems are instilled with increasing autonomy in making decisions and manipulating their environment, it is essential that they are designed to adopt, learn, and follow the norms and values of the community they serve. Moreover, their actions should be transparent in signaling their norm compliance and, if needed, they must be able to explain their actions. This is essential if humans are to develop appropriate levels of trust in A/IS in the specific contexts and roles in which A/IS function.

At the present time, the conceptual complexities surrounding what “values” are (Hitlin and Piliavin 2004¹; Malle and Dickert 2007²; Rohan 2000³; Sommer 2016⁴) make it difficult to envision A/IS that have computational structures directly corresponding to social or cultural values such as “security,” “autonomy,” or “fairness”. It may be a more realistic goal to embed explicit norms into such systems. Since norms are observable in human behavior, they can therefore be represented as instructions to act in defined ways in defined contexts, for a specific community—from family to town to country and beyond. A community’s network of social and moral norms is likely to reflect the community’s values, and A/IS equipped with such a network would, therefore, also reflect the community’s values. For discussion of specific values that are critical for ethical considerations of A/IS, see the chapters of *Ethically Aligned Design*, “Personal Data and Individual Agency” and “Well-being”.

Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally (Malle, Scheutz, and Austerweil 2017⁵; Vázquez-Salceda, Aldewereld and Dignum 2004⁶). They are typically qualitative in nature, e.g., do not stand too close to people. However, the implementation of norms also has a quantitative component—the measurement of the physical distance we mean by “too close”, and the possible instantiations of the quantitative component technically enable the qualitative norm.

Embedding Values into Autonomous and Intelligent Systems

To address the broad objective of embedding norms and, by implication, values into A/IS, this chapter addresses three more concrete goals:

1. Identifying the norms of the specific community in which the A/IS operate,
2. Computationally implementing the norms of that community within the A/IS, and
3. Evaluating whether the implementation of the identified norms in the A/IS are indeed conforming to the norms reflective of that community.

Pursuing these three goals represents an iterative process that is sensitive to the purpose of the A/IS and to its users within a specific community. It is understood that there may be conflicts of values and norms when identifying, implementing, and evaluating these systems. Such conflicts are a natural part of the dynamically changing and renegotiated norm systems of any community. As a result, we advocate for an approach in which systems are designed to provide transparent signals describing the specific nature of their behavior to the individuals in the community they serve. Such signals may include explanations or offers for inspection and must be in a language or form that is meaningful to the community.

Further Resources

- S. Hitlin and J. A. Piliavin, "Values: Reviving a Dormant Concept." *Annual Review of Sociology* 30, pp.359–393, 2004.
- B. F. Malle, and S. Dickert. "Values," in *Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs. Thousand Oaks, CA: Sage, 2007.
- B. F. Malle, M. Scheutz, and J. L. Austerweil. "Networks of Social and Moral Norms in Human and Robot Agents," in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, edited by M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, 3–17. Cham, Switzerland: Springer International Publishing, 2017.
- M. J. Rohan, "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review* 4, pp. 255–277, 2000.
- U. Sommer, *Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt*. [Values. Why We Need Them Even Though They Don't Exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. "Implementing Norms in Multiagent Systems," in *Multiagent System Technologies. MATES 2004*, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. (Lecture Notes in Computer Science, vol. 3187.) Berlin: Springer, 2004.

Embedding Values into Autonomous and Intelligent Systems

Section 1—Identifying Norms for Autonomous and Intelligent Systems

We identify three issues that must be addressed in the attempt to identify norms and corresponding values for A/IS. The first issue asks which norms should be identified and with which properties. Here we highlight context specificity as a fundamental property of norms. Second, we emphasize another important property of norms: their dynamically changing nature (Mack 2018⁷), which requires A/IS to have the capacity to update their norms and learn new ones. Third, we address the challenge of norm conflicts that naturally arise in a complex social world. Resolving such conflicts requires priority structures among norms, which help determine whether, in a given context, adhering to one norm is more important than adhering to another norm, often in light of overarching standards, e.g., laws and international humanitarian principles.

Issue 1: Which norms should be identified?

Background

If machines engage in human communities, then those agents will be expected to follow the community's social and moral norms. A necessary step in enabling machines to do so is to identify these norms. But which norms should be identified? Laws are publicly

documented and therefore easy to identify, so they can be incorporated into A/IS as long as they do not violate humanitarian or community moral principles. Social and moral norms are more difficult to ascertain, as they are expressed through behavior, language, customs, cultural symbols, and artifacts. Most important, communities ranging from families to whole nations differ to various degrees in the norms they follow. Therefore, generating a universal set of norms that applies to all A/IS in all contexts is not realistic, but neither is it advisable to completely tailor the A/IS to individual preferences. We suggest that it is feasible to identify broadly observed norms of communities in which a technology is deployed.

Furthermore, the difficulty of generating a universal set of norms is not inconsistent with the goal of seeking agreement over Universal Human Rights (see the “General Principles” chapter of *Ethically Aligned Design*). However, these universal rights are not sufficient for devising A/IS that conform to the specific norms of its community. Universal Human Rights must, however, constrain the kinds of norms that are implemented in the A/IS (cf. van de Poel 2016⁸).

Embedding norms in A/IS requires a careful understanding of the communities in which the A/IS are to be deployed. Further, even within a particular community, different types of A/IS will demand different sets of norms. The relevant

Embedding Values into Autonomous and Intelligent Systems

norms for self-driving vehicles, for example, may differ greatly from those for robots used in healthcare. Thus, we recommend that to develop A/IS capable of following legal, social, and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks and roles for which the A/IS are designed. Even when designating a narrowly defined community, e.g., a nursing home, an apartment complex, or a company, there will be variations in the norms that apply, or in their relative weighting. The norm identification process must heed such variation and ensure that the identified norms are representative, not only of the dominant subgroup in the community but also of vulnerable and underrepresented groups.

The most narrowly defined “community” is a single person, and A/IS may well have to adapt to the unique expectations and needs of a given individual, such as the arrangement of a disabled person’s living accommodations. However, unique individual expectations must not violate norms in the larger community. Whereas the arrangement of someone’s kitchen or the frequency with which a care robot checks in with a patient can be personalized without violating any community norms, encouraging the robot to use derogatory language to talk about certain social groups does violate such norms. In the next section, we discuss how A/IS might handle such norm conflicts.

Innovation projects and development efforts for A/IS should always rely on empirical research, involving multiple disciplines and multiple methods; to investigate and document both context- and task-specific norms, spoken and

unspoken, that typically apply in a particular community. Such a set of empirically identified norms should then guide system design. This process of norm identification and implementation must be iterative and revisable. A/IS with an initial set of implemented norms may betray biases of original assessments (Misra, Zitnick, Mitchell, and Girshick 2016⁹) that can be revealed by interactions with, and feedback from, the relevant community. This leads to a process of norm updating, which is described next in Issue 2.

Recommendation

To develop A/IS capable of following social and moral norms, the first step is to identify the norms of the specific community in which the A/IS are to be deployed and, in particular, norms relevant to the kinds of tasks and roles that the A/IS are designed for. This norm identification process must use appropriate scientific methods and continue through the system’s life cycle.

Further Resources

- Mack, Ed., “Changing social norms.” *Social Research: An International Quarterly*, 85, no.1, 1–271, 2018.
- I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2939. doi:[10.1109/CVPR.2016.320](https://doi.org/10.1109/CVPR.2016.320)
- I. van de Poel, “[An Ethical Framework for Evaluating Experimental Technology](#),” *Science and Engineering Ethics*, 22, no. 3, pp. 667–686, 2016.

Embedding Values into Autonomous and Intelligent Systems

Issue 2: The need for norm updating

Background

Norms are not static. They change over time, in response to social progress, political change, new legal measures, or novel opportunities (Mack 2018¹⁰). Norms can fade away when, for whatever reasons, fewer and fewer people adhere to them. And new norms emerge when technological innovation invites novel behaviors and novel standards, e.g., cell phone use in public.

A/IS should be equipped with a starting set of social and legal norms before they are deployed in their intended community (see Issue 1), but this will not suffice for A/IS to behave appropriately over time. A/IS or the designers of A/IS, must be adept at identifying and adding new norms to its starting set, because the initial norm identification process in the community will undoubtedly have missed some norms and because the community's norms change.

Humans rely on numerous capacities to update their knowledge of norms and learn new ones. They observe other community members' behavior and are sensitive to collective norm change; they explicitly ask about new norms when joining new communities, e.g., entering college or a job in a new town; and they respond to feedback from others when they exhibit uncertainty about norms or have violated a norm.

Likewise, A/IS need multiple capacities to improve their own norm knowledge and to adapt to a community's dynamically changing norms. These capacities include:

- Processing behavioral trends by members of the target community and comparing them to trends predicted by the baseline norm system,
- Asking for guidance from the community when uncertainty about applicable norms exceeds a critical threshold,
- Responding to instruction from the community members who introduce a robot to a previously unknown context or who notice the A/IS' uncertainty in a familiar context, and
- Responding to formal or informal feedback from the community when the A/IS violate a norm.

The modification of a normative system can occur at any level of the system: it could involve altering the priority weightings between individual norms, changing the qualitative expression of a norm, or altering the quantitative parameters that enable the norm.

We recommend that the system's norm changes be transparent. That is, the system or its designer should consult with users, designers, and community representatives when adding new norms to its norm system or adjusting the priority or content of existing norms. Allowing a system to learn new norms without public or expert review has detrimental consequences (Green and Hu 2018¹¹). The form of consultation

Embedding Values into Autonomous and Intelligent Systems

and the specific review process will vary by machine sophistication e.g., linguistic capacity and function/role, or a flexible social companion versus a task-defined medical robot and best practices will have to be established. In some cases, the system may document its dynamic change, and the user can consult this documentation as desired. In other cases, explicit announcements and requests for discussion with the designer may be appropriate. In yet other cases, the A/IS may propose changes, and the relevant human community, e.g., drawn from a representative crowdsourced panel, will decide whether such changes should be implemented in the system.

Recommendation

To respond to the dynamic change of norms in society A/IS or their designers must be able to amend their norms or add new ones, while being transparent about these changes to users, designers, broader community representatives, and other stakeholders.

Further Resources

- B. Green and L. Hu. "The Myth in the Methodology: Towards a Recontextualization of Fairness in ML." Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.
- Mack, Ed., "Changing social norms," *Social Research: An International Quarterly*, 85 (1, Special Issue), 1-271, 2018.

Issue 3: A/IS will face norm conflicts and need methods to resolve them.

Background

Often, even within a well-specified context, no action is available that fulfills all obligations and prohibitions. Such situations—often described as moral dilemmas or moral overload (Van den Hoven 2012¹²)—must be computationally tractable by A/IS; they cannot simply stop in their tracks and end on a logical contradiction. Humans resolve such situations by accepting trade-offs between conflicting norms, which constitute priorities of one norm or value over another in a given context. Such priorities may be represented in the norm system as hierarchical relations.

Along with identifying the norms within a specific community and task domain, empirical research must identify the ways in which people prioritize competing norms and resolve norm conflicts, and the ways in which people expect A/IS to resolve similar norm conflicts. These more local conflict resolutions will be further constrained by some general principles, such as the "Common Good Principle" (Andre and Velasquez 1992¹³) or local and national laws. For example, a self-driving vehicle's prioritization of one factor over another in its decision-making will need to reflect the laws and norms of the population in which the A/IS are deployed, e.g., the traffic laws of a U.S. state and the United States as a whole.

Embedding Values into Autonomous and Intelligent Systems

Some priority orders can be built into a given norm network as hierarchical relations, e.g., more general prohibitions against harm to humans typically override more specific norms against lying. Other priority orders can stem from the override that norms in the larger community exert on norms and preferences of an individual user. In the earlier example discussing personalization (see Issue 1), the A/IS of a racist user who demands the A/IS use derogatory language for certain social groups will have to resist such demands because community norms hierarchically override an individual user's preferences. In many cases, priority orders are not built in as fixed hierarchies because the priorities are themselves context-specific or may arise from net moral costs and benefits of the particular case at hand. A/IS must have learning capacities to track such variations and incorporate user and community input, e.g., about the subtle differences between contexts, so as to refine the system's norm network (see Issue 2).

Tension may sometimes arise between a community's social and legal norms and the normative considerations of designers or manufacturers. Democratic processes may need to be developed that resolve this tension—processes that cannot be presented in detail in this chapter. Often such resolution will favor the local laws and norms, but in some cases the community may have to be persuaded to accept A/IS favoring international law or broader humanitarian principles over, say, racist or sexist local practices.

In general, we recommend that the system's resolution of norm conflicts be transparent—that is, documented by the system and ready to be made available to users, the relevant community of deployment, and third-party evaluators. Just like people explain to each other why they made decisions, they will expect any A/IS to be able to explain their decisions and be sensitive to user feedback about the appropriateness of the decisions. To do so, design and development of A/IS should specifically identify the relevant groups of humans who may request explanations and evaluate the systems' behaviors. In the case of a system detecting a norm conflict, the system should consult and offer explanations to representatives from the community, e.g., randomly sampled crowdsourced members or elected officials, as well as to third-party evaluators, with the goal of discussing and resolving the norm conflict.

Recommendation

A/IS developers should identify the ways in which people resolve norm conflicts and the ways in which they expect A/IS to resolve similar norm conflicts. A system's resolution of norm conflicts must be transparent—that is, documented by the system and ready to be made available to users, the relevant community of deployment, and third-party evaluators.

Embedding Values into Autonomous and Intelligent Systems

Further Resources

- M. Velasquez, C. Andre, T. Shanks, S.J., and M. J. Meyer, "The Common Good." *Issues in Ethics*, vol. 5, no. 1, 1992.
- J. Van den Hoven, "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics*, vol. 18, no. 1, pp. 143–155, 2012.
- D. Abel, J. MacGlashan, and M. L. Littman. "Reinforcement Learning as a Framework for Ethical Decision Making." *AAAI Workshop AI, Ethics, and Society, Volume WS-16-02 of 13th AAAI Workshops*. Palo Alto, CA: AAAI Press, 2016.
- O. Bendel, *Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik*. Hannover, Germany: Heise Medien, 2016.
 - Accessible popular-science contributions to philosophical issues and technical implementations of machine ethics
- S. V. Burks, and E. L. Krupka. "A Multimethod Approach to Identifying Norms and Normative Expectations within a Corporate Hierarchy: Evidence from the Financial Services Industry." *Management Science*, vol. 58, pp. 203–217, 2012.
 - Illustrates surveys and incentivized coordination games as methods to elicit norms in a large financial services firm
- F. Cushman, V. Kumar, and P. Railton, "Moral Learning," *Cognition*, vol. 167, pp. 1–282, 2017.
- M. Flanagan, D. C. Howe, and H. Nissenbaum, "Embodying Values in Technology: Theory and Practice." *Information Technology and Moral Philosophy*, J. van den Hoven and J. Weckert, Eds., Cambridge University Press, 2008, pp. 322–53. Cambridge Core, *Cambridge University Press*. Preprint available at <http://www.nyu.edu/projects/nissenbaum/papers/Nissenbaum-VID.4-25.pdf>
- B. Friedman, P. H. Kahn, A. Borning, and A. Hultgren. "Value Sensitive Design and Information Systems," in *Early Engagement and New Technologies: Opening up the Laboratory*, N. Doorn, Schuurbijs, I. van de Poel, and M. Gorman, Eds., vol. 16, pp. 55–95. Dordrecht: Springer, 2013.
 - A comprehensive introduction into Value Sensitive Design and three sample applications
- G. Mackie, F. Moneti, E. Denny, and H. Shakya. "What Are Social Norms? How Are They Measured?" UNICEF Working Paper. University of California at San Diego: UNICEF, Sept. 2014. <https://dmeforpeace.org/sites/default/files/4%2009%2030%20Whole%20What%20are%20Social%20Norms.pdf>
 - A broad survey of conceptual and measurement questions regarding social norms.
- J. A. Leydens and J. C. Lucena. *Engineering Justice: Transforming Engineering Education and Practice*. Hoboken, NJ: John Wiley & Sons, 2018.
 - Identifies principles of engineering for social justice.

Embedding Values into Autonomous and Intelligent Systems

- B. F. Malle, "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology*, vol. 18, no. 4, pp. 243–256, 2016.
 - Discusses how a robot's norm capacity fits in the larger vision of a robot with moral competence.
- K. W. Miller, M. J. Wolf, and F. Grodzinsky, "This 'Ethical Trap' Is for Roboticians, Not Robots: On the Issue of Artificial Agent Ethical Decision-Making." *Science and Engineering Ethics*, vol. 23, pp. 389–401, 2017.
 - This article raises doubts about the possibility of imbuing artificial agents with morality, or of claiming to have done so.
- Open Roboethics Initiative: www.openroboethics.org. A series of poll results on differences in human moral decision-making and changes in priority order of values for autonomous systems (e.g., on [care robots](#)), 2019.
- A. Rizzo and L. L. Swisher, "Comparing the Stewart–Sprinthall Management Survey and the Defining Issues Test-2 as Measures of Moral Reasoning in Public Administration." *Journal of Public Administration Research and Theory*, vol. 14, pp. 335–348, 2004.
 - Describes two assessment instruments of moral reasoning (including norm maintenance) based on Kohlberg's theory of moral development.
- S. H. Schwartz, "An Overview of the Schwartz Theory of Basic Values." *Online Readings in Psychology and Culture* 2, 2012.
 - Comprehensive overview of a specific theory of values, understood as motivational orientations toward abstract outcomes (e.g., self-direction, power, security).
- S. H. Schwartz and K. Boehnke. "Evaluating the Structure of Human Values with Confirmatory Factor Analysis." *Journal of Research in Personality*, vol. 38, pp. 230–255, 2004.
 - Describes an older method of subjective judgments of relations among valued outcomes and a newer, formal method of analyzing these relations.
- W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
 - This book describes some of the challenges of having a one-size-fits-all approach to embedding human values in autonomous systems.

Embedding Values into Autonomous and Intelligent Systems

Section 2—Implementing Norms in Autonomous and Intelligent Systems

Once the norms relevant to A/IS' role in a specific community have been identified, including their properties and priority structure, we must link these norms to the functionalities of the underlying computational system. We discuss three issues that arise in this process of norm implementation. First, computational approaches to enable a system to represent, learn, and execute norms are only slowly emerging. However, the diversity of approaches may soon lead to substantial advances. Second, for A/IS that operate in human communities, there is a particular need for transparency—ranging from the technical process of implementation to the ethical decisions that A/IS will make in human-machine interactions, which will require a high level of explainability. Third, failures of normative reasoning can be considered inevitable and mitigation strategies should therefore be put in place to handle such failures when they occur.

As a general guideline, we recommend that, through the entire process of implementation of norms, designers should consider various forms and metrics of evaluation, and they should define and incorporate central criteria for assessing the A/IS' norm conformity, e.g., human-machine agreement on moral decisions, verifiability of A/IS decisions, or justified trust. In this way, implementation already prepares for the critical third phase of evaluation (discussed in Section 3).

Issue 1: Many approaches to norm implementation are currently available, and it is not yet settled which ones are most suitable.

Background

The prospect of developing A/IS that are sensitive to human norms and factor them into morally or legally significant decisions has intrigued science fiction writers, philosophers, and computer scientists alike. Modest efforts to realize this worthy goal in limited or bounded contexts are already underway. This emerging field of research appears under many names, including: machine morality, machine ethics, moral machines, value alignment, computational ethics, artificial morality, safe AI, and friendly AI.

There are a number of different implementation routes for implementing ethics into autonomous and intelligent systems. Following Wallach and Allen (2008)¹⁴, we might begin to categorize these as either:

- A. Top-down approaches, where the system, e.g., a software agent, has some symbolic representation of its activity, and so can identify specific states, plans, or actions as ethical or unethical with respect to particular ethical requirements (Dennis,

Embedding Values into Autonomous and Intelligent Systems

Fisher, Slavkovik, Webster 2016¹⁵; Pereira and Saptawijaya 2016¹⁶; Rötzer, 2016¹⁷; Scheutz, Malle, and Briggs 2015¹⁸); or

- B. Bottom-up approaches, where the system, e.g., a learning component, builds up, through experience of what is to be considered ethical and unethical in certain situations, an implicit notion of ethical behavior (Anderson and Anderson 2014¹⁹; Riedl and Harrison 2016²⁰).

Relevant examples of these two are: (A) symbolic agents that have explicit representations of plans, actions, goals, etc.; and (B) machine learning systems that train subsymbolic mechanisms with acceptable ethical behavior. For more detailed discussion, see Charisi et al. 2017²¹.

Many of the existing experimental approaches to building moral machines are top-down, in the sense that norms, rules, principles, or procedures are used by the system to evaluate the acceptability of differing courses of action, or as moral standards or goals to be realized. Increasingly, however, A/IS will encounter situations that initially programmed norms do not clearly address, requiring algorithmic procedures to select the better of two or more novel courses of action. Recent breakthroughs in machine learning and perception enable researchers to explore bottom-up approaches in which the A/IS learn about their context and about human norms, similar to the manner in which a child slowly learns which forms of behavior are safe and acceptable. Of course, unlike current A/IS, children can feel pain and pleasure, and empathize with others. Still, A/IS can learn to detect and take into account others' pain and pleasure, thus at least achieving some of the positive effects of empathy. As research on A/IS

progresses, engineers will explore new ways to improve these capabilities.

Each of the first two options has obvious limitations, such as option A's inability to learn and adapt and option B's unconstrained learning behavior. A third option tries to address these limitations:

- C. Hybrid approaches, combining (A) and (B).

For example, the selection of action might be carried out by a subsymbolic system, but this action must be checked by a symbolic "gateway" agent before being invoked. This is a typical approach for "Ethical Governors" (Arkin, 2008²²; Winfield, Blum, and Liu 2014²³) or "Guardians" (Etzioni 2016²⁴) that monitor, restrict, and even adapt certain unacceptable behaviors proposed by the system (see Issue 3). Alternatively, action selection in light of norms could be done in a verifiable logical format, while many of the norms constraining those actions can be learned through bottom-up learning mechanisms (Arnold, Kasenberg, and Scheutz 2017²⁵).

These three architectures do not cover all possible techniques for implementing norms into A/IS. For example, some contributors to the multi-agent systems literature have integrated norms into their agent specifications (Andrighetto et al. 2013²⁶), and even though these agents live in societal simulations and are too underspecified to be translated into individual A/IS such as robots, the emerging work can inform cognitive architectures of such A/IS that fully integrate norms. Of course, none of these experimental systems should be deployed outside of the laboratory before testing or before certain criteria are met, which we outline in the remainder of this section and in Section 3.

Embedding Values into Autonomous and Intelligent Systems

Recommendation

In light of the multiple possible approaches to computationally implement norms, diverse research efforts should be pursued, especially collaborative research between scientists from different schools of thought and different disciplines.

Further Resources

- M. Anderson, and S. L. Anderson, "GenEth: A General Ethical Dilemma Analyzer," *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, July 27 –31, 2014, pp. 253–261, Palo Alto, CA, The AAAI Press, 2014.
- G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Amsterdam, Netherlands, March 12 -15, 2008, IEEE, pp. 121–128, 2008.
- T. Arnold, D. Kasenberg, and M. Scheutz. "Value Alignment or Misalignment—What Will Keep Systems Accountable?" *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports*, WS-17-02: AI, Ethics, and Society, pp. 81–88. Palo Alto, CA: The AAAI Press, 2017.
- V. Charisi, L. Dennis, M. Fisher, et al. "Towards Moral Autonomous Systems," 2017.
- A. Conn, "[How Do We Align Artificial Intelligence with Human Values?](#)" *Future of Life Institute*, Feb. 3, 2017.
- L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016.
- A. Etzioni and O. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM*, vol. 59, no. 9, pp. 29–31, Sept. 2016.
- L. M. Pereira and A. Saptawijaya, *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.
- M. O. Riedl and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *AAAI Workshops 2016*. Phoenix, Arizona, February 12–13, 2016.
- F. Rötzer, ed. *Programmierte Ethik: Brauchen Roboter Regeln oder Moral?* Hannover, Germany: Heise Medien, 2016.
- M. Scheutz, B. F. Malle, and G. Briggs. "Towards Morally Sensitive Action Selection for Autonomous Social Robots." *Proceedings of the 24th International Symposium on Robot and Human Interactive Communication, RO-MAN 2015* (2015): 492–497.
- U. Sommer, *Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt*. [Values. Why we need them even though they don't exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- I. Sommerville, *Software Engineering*. Harlow, U.K.: Pearson Studium, 2001.
- W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
- F. T. Winfield, C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection" in *Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume*, M. Mistry, A. Leonardis, Witkowski, and C. Melhuish, eds. pp. 85–96. Springer, 2014.

Embedding Values into Autonomous and Intelligent Systems

Issue 2: The need for transparency from implementation to deployment

Background

When A/IS become part of social communities and behave according to the norms of their communities, people will want to understand the A/IS decisions and actions, just as they want to understand each other's decisions and actions. This is particularly true for morally significant actions or omissions: an ethical reasoning system should be able to explain its own reasoning to a user on request. Thus, transparency, or “explainability”, of A/IS is paramount (Chaudhuri 2017²⁷; Wachter, Mittelstadt, and Floridi 2017²⁸), and it will allow a community to understand, predict, and modify the A/IS (see Section 1, Issue 2; for a nuanced discussion see Selbst and Barocas²⁹). Moreover, as the norms embedded in A/IS are continuously updated and refined (see Section 1, Issue 2), transparency allows for appropriate trust to be developed (Grodzinsky, Miller, and Wolf 2011³⁰), and, where necessary, allows the community to modify a system's norms, reasoning, and behavior.

Transparency can occur at multiple levels, e.g., ordinary language or coder verification, and for multiple stakeholders, e.g., user, engineer, and attorney. (See [IEEE P7001™](#), IEEE Standards Project for Transparency of Autonomous Systems). It should be noted that transparency to all parties may not always be advisable, such as in the case of security programs that prevent a system from being hacked (Kroll et al. 2016³¹). Here we briefly illustrate the broad

range of transparency by reference to four ways in which systems can be transparent—traceability, verifiability, honest design, and intelligibility—and apply these considerations to the implementation of norms in A/IS.

Transparency as traceability—Most relevant for the topic of implementation is the transparency of the software engineering process during implementation (Cleland-Huang, Gotel, and Zisman 2012³²). It allows for the originally identified norms (Section 1, Issue 1) to be traced through to the final system. This allows technical inspection of which norms have been implemented, for which contexts, and how norm conflicts are resolved, e.g., priority weights given to different norms. Transparency in the implementation process may also reveal biases that were inadvertently built into systems, such as racism and sexism, in search engine algorithms (Noble 2013³³). (See Section 3, Issue 2.) Such traceability in turn calibrates a community's trust about whether A/IS are conforming to the norms and values relevant in their use contexts (Fleischmann and Wallace 2005³⁴).

Transparency as verifiability—Transparency concerning how normative reasoning is approached in the implementation is important as we wish to verify that the normative decisions the system makes match the required norms and values. Explicit and exact representations of these normative decisions can then provide the basis for a range of strong mathematical techniques, such as formal verification (Fisher, Dennis, and Webster 2013³⁵). Even if a system cannot explain every single reasoning step in understandable human terms, a log of ethical reasoning should be available for inspection of later evaluation purposes (Hind et al. 2018³⁶).

Embedding Values into Autonomous and Intelligent Systems

Transparency as honest design—German designer Dieter Rams coined the term “honest design” to refer to design that “does not make a product more innovative, powerful or valuable than it really is” (Vitsoe 2018³⁷; see also Donelli 2015³⁸; Jong 2017³⁹). Honest design of A/IS is one aspect of their transparency, because it allows the user to “see through” the outward appearance and accurately infer the A/IS’ actual capacities. At times, however, the physical appearance of a system does not accurately represent what the system is capable of doing—e.g., the agent displays signs of a certain human-like emotion but its internal state does not represent that human emotion. Humans are quick to make strong inferences from outward appearances of human-likeness to the mental and social capacities the A/IS might have. Demands for transparency in design therefore put a responsibility on the designer to “not attempt to manipulate the consumer with promises that cannot be kept” (Vitsoe 2018⁴⁰).

Transparency as intelligibility—As mentioned above, humans will want to understand the A/IS’ decisions and actions, especially the morally significant ones. A clear requirement for an ethical A/IS is that the system be able to explain its own reasoning to a user, when asked—or, ideally, also when suspecting the user’s confusion, and the system should do so at a level of ordinary human reasoning, not with incomprehensible technical detail (Tintarev and Kutlak 2014⁴¹). Furthermore, when the system cannot explain some of its actions, technicians or designers should be available to make those actions intelligible. Along these lines, the European Union’s General Data Protection Regulation (GDPR), in effect since May 2018, states that, for automated decisions based on personal data, individuals have a right

to “an explanation of the [algorithmic] decision reached after such assessment and to challenge the decision”. (See boyd [sic] 2016⁴², for a critical discussion of this regulation.)

Recommendation

A/IS, especially those with embedded norms, must have a high level of transparency, shown as traceability in the implementation process, mathematical verifiability of their reasoning, honesty in appearance-based signals, and intelligibility of the systems’ operation and decisions.

Further Resources

- d. boyd, “Transparency ≠ Accountability.” *Data & Society: Points*, November 29, 2016.
- A. Chaudhuri, “Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology,” *The EDP Audit, Control, and Security Newsletter*, vol. 56, no. 4, pp. 7-18, DOI: 10.1080/07366981.2017.1380474, 2017.
- J. Cleland-Huang, O. Gotel, and A. Zisman, eds. *Software and Systems Traceability*. London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- G. Donelli, “Good design is honest.” (blog). March 13, 2015. Accessed Oct 22, 2018. <https://blog.astropad.com/good-design-is-honest/>
- M. Fisher, L. A. Dennis, and M. P. Webster. “Verifying Autonomous Systems.” *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.

Embedding Values into Autonomous and Intelligent Systems

- K. R. Fleischmann and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM*, vol. 48, no. 5, pp. 93–97, 2005.
- F. S. Grodzinsky, K. W. Miller, and M. J. Wolf. "Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from This Artificial Agent?" *Ethics and Information Technology*, vol. 13, pp. 17–27, 2011.
- M. Hind, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *ArXiv E-Prints*, Aug. 2018. [Online] Available: <https://arxiv.org/abs/1808.07261>. [Accessed October 28, 2018].
- C. W. De Jong, ed., *Dieter Rams: Ten Principles for Good Design*. New York, NY: Prestel Publishing, 2017.
- J. A. Kroll, J. Huey, S. Barocas et al. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 2017.
- S. U. Noble, "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture* 19, 2013.
- D. Selbst and S. Barocas, "The Intuitive Appeal of Explainable Machines," *87 Fordham Law Review* 1085, Available at SSRN: <https://ssrn.com/abstract=3126971> or <http://dx.doi.org/10.2139/ssrn.3126971>, Feb. 19, 2018.
- N. Tintarev and R. Kutlak. "Demo: Making Plans Scrutable with Argumentation and Natural Language Generation." *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, pp. 29–32, 2014.
- Vitsoe. "The Power of Good Design." *Vitsoe*, 2018. Retrieved Oct 22, 2018 from <https://www.vitsoe.com/us/about/good-design>.
- S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics*, vol. 2, no. 6, eaan6080. doi:10.1126/scirobotics.aan6080, 2017.

Embedding Values into Autonomous and Intelligent Systems

Issue 3: Failures will occur.

Background

Operational failures and, in particular, violations of a system's embedded community norms, are unavoidable, both during system testing and during deployment. Not only are implementations never perfect, but A/IS with embedded norms will update or expand their norms over time (see Section 1, Issue 2) and interactions in the social world are particularly complex and uncertain. Thus, prevention and mitigation strategies must be adopted, and we sample four possible ones.

First, anticipating the process of evaluation during the implementation phase requires defining criteria and metrics for such evaluation, which in turn better allows the detection and mitigation of failures. Metrics will include:

- Technical variables, such as traceability and verifiability,
- User-level variables such as reliability, understandable explanations, and responsiveness to feedback, and
- Community-level variables such as justified trust (see Issue 2) and the collective belief that A/IS are generally creating social benefits rather than, for example, technological unemployment.

Second, a systematic risk analysis and management approach can be useful (Oetzel and Spiekermann 2014⁴³) for an application to privacy

norms. This approach tries to anticipate potential points of failure, e.g., norm violations, and, where possible, develops some ways to reduce or remove the effects of failures. Successful behavior, and occasional failures, can then iteratively improve predictions and mitigation attempts.

Third, because not all risks and failures are predictable (Brundage et al 2018⁴⁴; Vanderelst and Winfield 2018⁴⁵), especially in complex human-machine interactions in social contexts, additional mitigation mechanisms must be made available. Designers are strongly encouraged to augment the architectures of their systems with components that handle unanticipated norm violations with a fail-safe, such as the symbolic "gateway" agents discussed in Section 2, Issue 1. Designers should identify a number of strict laws, that is, task- and community-specific norms that should never be violated, and the fail-safe components should continuously monitor operations against possible violations of these laws. In case of violations, the higher-order gateway agent should take appropriate actions, such as safely disabling the system's operation, or greatly limiting its scope of operation, until the source of failure is identified. The fail-safe components need to be understandable, extremely reliable, and protected against security breaches, which can be achieved, for example, by validating them carefully and not letting them adapt their parameters during execution.

Fourth, once failures have occurred, responsible entities, e.g., corporate, government, science, and engineering, shall create a publicly accessible

Embedding Values into Autonomous and Intelligent Systems

database with undesired outcomes caused by specific A/IS systems. The database would include descriptions of the problem, background information on how the problem was detected, which context it occurred in, and how it was addressed.

In summary, we offer the following recommendation.

Recommendation

Because designers and developers cannot anticipate all possible operating conditions and potential failures of A/IS, multiple strategies to mitigate the chance and magnitude of harm must be in place.

Further Resources

- M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," CoRR abs/1802.07228 [cs.AI]. 2018. <https://arxiv.org/abs/1802.07228>
- M. C. Oetzel and S. Spiekermann, "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems*, vol. 23, pp. 126–150, 2014. <https://link.springer.com/article/10.1057/ejis.2013.18>
- D. Vanderelst and A.F. Winfield, 2018 "The Dark Side of Ethical Robots," In Proc. The First AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society, New Orleans, LA, Feb. 1 -3, 2018.

Embedding Values into Autonomous and Intelligent Systems

Section 3—Evaluating the Implementation of A/IS

The success of implementing appropriate norms in A/IS must be rigorously evaluated. This evaluation process must be anticipated during design and incorporated into the implementation process and continue throughout the life cycle of the system's deployment. Assessment before full-scale deployment would best take place in systematic test beds that allow human users—from the defined community and representing all demographic groups—to engage safely with the A/IS in intended tasks. Multiple disciplines and methods should contribute to developing and conducting such evaluations.

Evaluation criteria must capture, among others, the quality of human-machine interactions, human approval and appreciation of the A/IS, appropriate trust in the A/IS, adaptability of the A/IS to human users, and benefits to human well-being in the presence or under the influence of the A/IS. A range of normative aspects to be considered can be found in British Standard BS 8611:2016 on Robot Ethics (British Standards Institution 2016⁴⁶). These are important general evaluation criteria, but they do not yet fully capture evaluation of a system that has “norm capacities”.

To evaluate a system's norm-conforming behavior, one must describe—and ideally, formally specify—criterion behaviors that reflect the previously identified norms, describe what

the user expects the system to do, verify that the system really does this, and validate that the specification actually matches the criteria. Many different evaluation techniques are available in the field of software engineering (Sommerville 2015⁴⁷), ranging from formal mathematical proof, through rigorous empirical testing against criteria of normatively correct behavior, to informal analysis of user interactions and responses to the machine's norm awareness and compliance. All these approaches can, in principle, be applied to the full range of A/IS including robots (Fisher, Dennis, and Webster 2013⁴⁸). More general principles from system quality management may also be integrated into the evaluation process, such as the Plan-Do-Check-Act (PDCA) cycle that underlies standards like ISO 9001 (International Organization for Standardization 2015⁴⁹).

Evaluation may be done by first parties, e.g., designers, manufacturers, and users, as well as third parties, e.g., regulators, independent testing agencies, and certification bodies. In either case, the results of evaluations should be made available to all parties, with strong encouragement to resolve discovered system limitations and resolve potential discrepancies among multiple evaluations.

As a general guideline, we recommend that evaluation of A/IS implementations must be anticipated during a system's design, incorporated

Embedding Values into Autonomous and Intelligent Systems

into the implementation process, and continue throughout the system's deployment (cf. ITIL principles, BMC 2016⁵⁰). Evaluation must include multiple methods, be made available to all parties—from designers and users to regulators, and should include procedures to resolve conflicting evaluation results. Specific issues that need to be addressed in this process are discussed next.

Further Resources

- British Standards Institution. BS8611:2016, "Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems," 2016.
- BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. <http://www.bmc.com/guides/itil-introduction.html>, Dec. 6, 2016.
- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, no. 9, pp. 84–93, 2013.
- International Organization for Standardization (2015). ISO 9001:2015, Quality management systems —Requirements. Retrieved July 12, 2018 from <https://www.iso.org/standard/62085.html>.
- I. Sommerville, *Software Engineering*. 10th ed. Harlow, U.K.: Pearson Studium, 2015.

Issue 1: Not all norms of a target community apply equally to human and artificial agents

Background

An intuitive criterion for evaluations of norms embedded in A/IS would be that the A/IS norms should mirror the community's norms—that is, the A/IS should be disposed to behave the same way that people expect each other to behave. However, for a given community and a given A/IS use context, A/IS and humans are unlikely to have identical sets of norms. People will have some unique expectations for humans than they do not for machines, e.g., norms governing the regulation of negative emotions, assuming that machines do not have such emotions. People may in some cases have unique expectations of A/IS that they do not have for humans, e.g., a robot worker, but not a human worker, is expected to work without regular breaks.

Recommendation

The norm identification process must document the similarities and differences between the norms that humans apply to other humans and the norms they apply to A/IS. Norm implementations should be evaluated specifically against the norms that the community expects the A/IS to follow.

Embedding Values into Autonomous and Intelligent Systems

Issue 2: A/IS can have biases that disadvantage specific groups

Background

Even when reflecting the full system of community norms that was identified, A/IS may show operation biases that disadvantage specific groups in the community or instill biases in users by reinforcing group stereotypes. A system's bias can emerge in perception. For example, a passport application AI rejected an Asian man's photo because it insisted his eyes were closed (Griffiths 2016⁵¹). Bias can emerge in information processing. For instance, speech recognition systems are notoriously less accurate for female speakers than for male speakers (Tatman 2016⁵²). System bias can affect decisions, such as a criminal risk assessment device which overpredicts recidivism by African Americans (Angwin et al. 2016⁵³). The system's bias can present itself even in its own appearance and presentation: the vast majority of humanoid robots have white "skin" color and use female voices (Riek and Howard 2014⁵⁴).

The norm identification process detailed in Section 1 is intended to minimize individual designers' biases because the community norms are assessed empirically. The identification process also seeks to incorporate norms against prejudice and discrimination. However, biases may still emerge from imperfections in the norm identification process itself, from unrepresentative training sets for machine learning systems, and from programmers' and designers' unconscious

assumptions. Therefore, unanticipated or undetected biases should be further reduced by including members of diverse social groups in both the planning and evaluation of A/IS and integrating community outreach into the evaluation process, e.g., [DO-IT](#) program and [RRI](#) framework. Behavioral scientists and members of the target populations will be particularly valuable when devising criterion tasks for system evaluation and assessing the success of evaluating the A/IS performance on those tasks. Such tasks would assess, for example, whether the A/IS apply norms in discriminatory ways to different races, ethnicities, genders, ages, body shapes, or to people who use wheelchairs or prosthetics, and so on.

Recommendation

Evaluation of A/IS must carefully assess potential biases in the systems' performance that disadvantage specific social and demographic groups. The evaluation process should integrate members of potentially disadvantaged groups in efforts to diagnose and correct such biases.

Further Resources

- J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica, May 23, 2016.
- J. Griffiths, "New Zealand Passport Robot Thinks This Asian Man's Eyes Are Closed." CNN.com, December 9, 2016.

Embedding Values into Autonomous and Intelligent Systems

- L. D. Riek and D. Howard, "A Code of Ethics for the Human-Robot Interaction Profession." *Proceedings of We Robot*, April 4, 2014.
- R. Tatman, "Google's Speech Recognition Has a Gender Bias." *Making Noise and Hearing Things*, July 12, 2016.

Issue 3: Challenges to evaluation by third parties

Background

A/IS should have sufficient transparency to allow evaluation by third parties, including regulators, consumer advocates, ethicists, post-accident investigators, or society at large. However, transparency can be severely limited in some systems, especially in those that rely on machine learning algorithms trained on large data sets. The data sets may not be accessible to evaluators; the algorithms may be proprietary information or mathematically so complex that they defy common-sense explanation; and even fellow software experts may be unable to verify reliability and efficacy of the final system because the system's specifications are opaque.

For less inscrutable systems, numerous techniques are available to evaluate the implementation of the A/IS' norm conformity. On one side there is formal verification, which provides a mathematical proof that the A/IS will always match specific normative and ethical requirements, typically devised in a top-down

approach (see Section 2, Issue 1). This approach requires access to the decision-making process and the reasons for each decision (Fisher, Dennis, and Webster 2013⁵⁵). A simpler alternative, sometimes suitable even for machine learning systems, is to test the A/IS against a set of scenarios and assess how well they matches their normative requirements, e.g., acting in accordance with relevant norms and recognizing other agents' norm violations. A "red team" may also devise scenarios that try to get the A/IS to break norms so that its vulnerabilities can be revealed.

These different evaluation techniques can be assigned different levels of "strength": strong ones demonstrate the exhaustive set of the A/IS' allowable behaviors for a range of criterion scenarios; weaker ones sample from criterion scenarios and illustrate the systems' behavior for that subsample. In the latter case, confidence in the A/IS' ability to meet normative requirements is more limited. An evaluation's concluding judgment must therefore acknowledge the strength of the verification technique used, and the expressed confidence in the evaluation—and in the A/IS themselves—must be qualified by this level of strength.

Transparency is only a necessary requirement for a more important long-term goal: having systems be accountable to their users and community members. However, this goal raises many questions such as to whom the A/IS are accountable, who has the right to correct the systems, and which kind of A/IS should be subject to accountability requirements.

Embedding Values into Autonomous and Intelligent Systems

Recommendation

To maximize effective evaluation by third parties, e.g., regulators and accident investigators, A/IS should be designed, specified, and documented so as to permit the use of strong verification and validation techniques for assessing the system's safety and norm compliance, in order to achieve accountability to the relevant communities.

Further Resources

- M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM*, vol. 56, pp. 84–93, 2013.
- K. Abney, G. A. Bekey, and P. Lin. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: The MIT Press, 2011.
- M. Anderson and S. L. Anderson, eds. *Machine Ethics*. New York: Cambridge University Press, 2011.
- M. Boden, J. Bryson, et al. "Principles of Robotics: Regulating Robots in the Real World." *Connection Science* 29, no. 2, pp. 124–129, 2017.
- M. Coeckelbergh, "[Can We Trust Robots?](#)" *Ethics and Information Technology*, vol.14, pp. 53–60, 2012.
- L. A. Dennis, M. Fisher, N. Lincoln, A. Lisitsa, and S. M. Veres, "Practical Verification of Decision-Making in Agent-Based Autonomous Systems." *Automated Software Engineering*, vol. 23, no. 3, pp. 305–359, 2016.
- M. Fisher, C. List, M. Slavkovik, and A. F. T. Winfield. "Engineering Moral Agents—From Human Morality to Artificial Morality" (Dagstuhl Seminar 16222). *Dagstuhl Reports* 6, no. 5, pp. 114–137, 2016.
- K. R. Fleischmann, *Information and Human Values*. San Rafael, CA: Morgan and Claypool, 2014.
- G. Governatori and A. Rotolo. "How Do Agents Comply with Norms?" in *Normative Multi-Agent Systems*, G. Boella, P. Noriega, G. Pigozzi, and H. Verhagen, eds., *Dagstuhl Seminar Proceedings*. Dagstuhl, Germany: Schloss Dagstuhl—Leibniz-Zentrum für Informatik, 2009.
- B. Higgins, "New York City Task Force to Consider Algorithmic Harm." *Artificial Intelligence Technology and the Law Blog*, Feb. 7, 2018. [Online]. Available: <http://aitechnologylaw.com/2018/02/new-york-city-task-force-algorithmic-harm/>. [Accessed Nov. 1, 2018].
- S. L. Jarvenpaa, N. Tractinsky, and L. Saarinen. "[Consumer Trust in an Internet Store: A Cross-Cultural Validation](#)" *Journal of Computer-Mediated Communication*, vol. 5, no. 2, pp. 1–37, 1999.
- E. H. Leet and W. A. Wallace. "Society's Role and the Ethics of Modeling," in *Ethics in Modeling*, W. A. Wallace, ed., Tarrytown, NY: Elsevier, 1994, pp. 242–245.
- M. A. Mahmoud, M. S. Ahmad, M. Z. M. Yusoff, and A. Mustapha. "[A Review of Norms and Normative Multiagent Systems](#)," *The Scientific World Journal*, vol. 2014, Article ID 684587, 2014.

Embedding Values into Autonomous and Intelligent Systems

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Embedding Values into Autonomous Intelligent Systems Committee

- **AJung Moon** (Founding Chair) – Director of Open Roboethics Institute
- **Bertram F. Malle** (Co-Chair) – Professor, Department of Cognitive, Linguistic, and Psychological Sciences, Co-Director of the Humanity-Centered Robotics Initiative, Brown University
- **Francesca Rossi** (Co-Chair) – Full Professor, computer science at the University of Padova, Italy, currently at the IBM Research Center at Yorktown Heights, NY
- **Stefano Albrecht** – Postdoctoral Fellow in the Department of Computer Science at The University of Texas at Austin
- **Bijilash Babu** – Senior Manager, Ernst and Young, EY Global Delivery Services India LLP
- **Jan Carlo Barca** – Senior Lecturer in Software Engineering and Internet of Things (IoT), School of Info Technology, Deakin University, Australia
- **Catherine Berger** – IEEE Standards Senior Program Manager, IEEE
- **Malo Bourgon** – COO, Machine Intelligence Research Institute
- **Richard S. Bowyer** – Adjunct Senior Lecturer and Research Fellow, College of Science and Engineering, Centre for Maritime Engineering, Control and Imaging (cmeci), Flinders University, South Australia
- **Stephen Cave** – Executive Director of the Leverhulme Centre for the Future of Intelligence, University of Cambridge
- **Raja Chatila** – CNRS-Sorbonne Institute of Intelligent Systems and Robotics, Paris, France; Member of the French Commission on the Ethics of Digital Sciences and Technologies CERNA; Past President of IEEE Robotics and Automation Society
- **Mark Coeckelbergh** – Professor, Philosophy of Media and Technology, the University of Vienna
- **Louise Dennis** – Lecturer, Autonomy and Verification Laboratory, University of Liverpool
- **Laurence Devillers** – Professor of Computer Sciences, University Paris Sorbonne, LIMSI-CNRS 'Affective and social dimensions in spoken interactions'; member of the French Commission on the Ethics of Research in Digital Sciences and Technologies (CERNA)
- **Virginia Dignum** – Associate Professor, Faculty of Technology Policy and Management, TU Delft

Embedding Values into Autonomous and Intelligent Systems

- **Ebru Dogan** – Research Engineer, VEDECOM
- **Takashi Egawa** – Cloud Infrastructure Laboratory, NEC Corporation, Tokyo
- **Vanessa Evers** – Professor, Human-Machine Interaction, and Science Director, DesignLab, University of Twente
- **Michael Fisher** – Professor of Computer Science, University of Liverpool, and Director of the UK Network on the Verification and Validation of Autonomous Systems, vavas.org
- **Ken Fleischmann** – Associate Professor in the School of Information at The University of Texas at Austin
- **Edith Pulido Herrera** – Bioengineering group, Antonio Nariño University, Bogotá, Colombia
- **Ryan Integlia** – assistant professor, Electrical and Computer Engineering, Florida Polytechnic University; Co-Founder of the em[POWER] Energy Group
- **Catholijn Jonker** – Full professor of Interactive Intelligence at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology. Part-time full professor at Leiden Institute of Advanced Computer Science of the Leiden University
- **Sara Jordan** – Assistant Professor of Public Administration in the Center for Public Administration & Policy at Virginia Tech
- **Jong-Wook Kim** – Professor, AI.Robotics Lab, Department of Electronic Engineering, Dong-A University, Busan, Korea
- **Sven Koenig** – Professor, Computer Science Department, University of Southern California
- **Brenda Leong** – Senior Counsel, Director of Operations, The Future of Privacy Forum
- **Alan Mackworth** – Professor of Computer Science, University of British Columbia; Former President, AAAI; Co-author of “Artificial Intelligence: Foundations of Computational Agents”.
- **Pablo Noriega** – Scientist, Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC), Barcelona.
- **Rajendran Parthiban** – Professor, School of Engineering, Monash University, Bandar Sunway, Malaysia
- **Heather M. Patterson** – Senior Research Scientist, Anticipatory Computing Lab, Intel Corp.
- **Edson Prestes** – Professor, Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Brazil; Head, Phi Robotics Research Group, UFRGS; CNPq Fellow.
- **Laurel Riek** – Associate Professor, Computer Science and Engineering, University of California San Diego
- **Leanne Seeto** – Co-Founder and Strategy and Operations Precision Autonomy
- **Sarah Spiekermann** – Chair of the Institute for Information Systems & Society at Vienna University of Economics and Business; Author of the textbook “Ethical IT-Innovation”, the popular book “Digitale Ethik—Ein Wertesystem für das 21. Jahrhundert” and Blogger on “The Ethical Machine”

Embedding Values into Autonomous and Intelligent Systems

- **John P. Sullins** – Professor of Philosophy, Chair of the Center for Ethics Law and Society (CELS), Sonoma State University
- **Jaan Tallinn** – Founding engineer of Skype and Kazaa; co-founder of the Future of Life Institute
- **Mike Van der Loos** – Associate Prof., Dept. of Mechanical Engineering, Director of Robotics for Rehabilitation, Exercise and Assessment in Collaborative Healthcare (RREACH) Lab, and Associate Director of CARIS Lab, University of British Columbia
- **Wendell Wallach** – Consultant, ethicist, and scholar, Yale University's Interdisciplinary Center for Bioethics
- **Nell Watson** – CFBCS, FICS, FIAP, FIKE, FRSA, FRSS, FLS Co-Founder and Chairman, EthicsNet, AI & Robotics Faculty Singularity University, Foresight Machine Ethics Fellow
- **Karolina Zawieska** – Postdoctoral Research Fellow in Ethics and Cultural Learning of Robotics at DeMontfort University, UK and Researcher at Industrial Research Institute for Automation and Measurements PIAP, Poland

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Embedding Values into Autonomous and Intelligent Systems

Endnotes

- ¹ S. Hitlin and J. A. Piliavin. "Values: Reviving a Dormant Concept." *Annual Review of Sociology* 30 (2004): 359–393.
- ² B. F. Malle, and S. Dickert. "Values," *The Encyclopedia of Social Psychology*, edited by R. F. Baumeister and K. D. Vohs. Thousand Oaks, CA: Sage, 2007.
- ³ M. J. Rohan, "A Rose by Any Name? The Values Construct." *Personality and Social Psychology Review* 4 (2000): 255–277.
- ⁴ A. U. Sommer, *Werte: Warum Man Sie Braucht, Obwohl es Sie Nicht Gibt*. [Values. Why We Need Them Even Though They Don't Exist.] Stuttgart, Germany: J. B. Metzler, 2016.
- ⁵ B. F. Malle, M. Scheutz, and J. L. Austerweil. "Networks of Social and Moral Norms in Human and Robot Agents," in *A World with Robots: International Conference on Robot Ethics: ICRE 2015*, edited by M. I. Aldinhas Ferreira, J. Silva Sequeira, M. O. Tokhi, E. E. Kadar, and G. S. Virk, 3–17. Cham, Switzerland: Springer International Publishing, 2017.
- ⁶ J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. "Implementing Norms in Multiagent Systems," in *Multiagent System Technologies. MATES 2004*, edited by G. Lindemann, Denzinger, I. J. Timm, and R. Unland. ([Lecture Notes in Computer Science, vol. 3187](#).) Berlin: Springer, 2004.
- ⁷ A. Mack, (Ed.). "Changing social norms." *Social Research: An International Quarterly*, 85, no.1 (2018): 1–271.
- ⁸ I. van de Poel, "[An Ethical Framework for Evaluating Experimental Technology](#)", *Science and Engineering Ethics*, 22, no. 3 (2016): 667–686.
- ⁹ I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, (2016). Seeing through the human reporting bias: Visual Classifiers from Noisy Human-Centric Labels. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2930–2939). doi:[10.1109/CVPR.2016.320](#)
- ¹⁰ A. Mack, (Ed.). (2018). Changing social norms. *Social Research: An International Quarterly*, 85(1, Special Issue), 1–271.
- ¹¹ B. Green and L. Hu. "The Myth in the Methodology: Towards a Recontextualization of Fairness in ML." Paper presented at the Debates workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden 2018.
- ¹² J. Van den Hoven, "Engineering and the Problem of Moral Overload." *Science and Engineering Ethics* 18, no. 1 (2012): 143–155.
- ¹³ C. Andre and M. Velasquez. "[The Common Good](#)." *Issues in Ethics* 5, no. 1 (1992).
- ¹⁴ W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press, 2008.
- ¹⁵ L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. "Formal Verification of Ethical Choices in Autonomous Systems." *Robotics and Autonomous Systems* 77 (2016): 1–14.

Embedding Values into Autonomous and Intelligent Systems

- ¹⁶ L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*. Cham, Switzerland: Springer International, 2016.
- ¹⁷ F. Rötzer, ed. *Programmierte Ethik: Brauchen Roboter Regeln oder Moral?* Hannover, Germany: Heise Medien, 2016.
- ¹⁸ M. Scheutz, B. F. Malle, and G. Briggs. "Towards Morally Sensitive Action Selection for Autonomous Social Robots." *Proceedings of the 24th International Symposium on Robot and Human Interactive Communication, RO-MAN 2015* (2015): 492–497.
- ¹⁹ M. Anderson and S. L. Anderson. "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.
- ²⁰ M. O. Riedl and B. Harrison. "Using Stories to Teach Human Values to Artificial Agents." *Proceedings of the 2nd International Workshop on AI, Ethics and Society*, Phoenix, Arizona, 2016.
- ²¹ V. Charisi, L. Dennis, M. Fisher et al. "[Towards Moral Autonomous Systems](#)," 2017.
- ²² R. Arkin, "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *Proceedings of the 2008 3rd ACM/IEEE International Conference on Human-Robot Interaction* (2008): 121–128.
- ²³ A. F. T. Winfield, C. Blum, and W. Liu. "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection" in *Advances in Autonomous Robotics Systems, Lecture Notes in Computer Science Volume*, edited by M. Mistry, A. Leonardis, Witkowski, and C. Melhuish, 85–96. Springer, 2014.
- ²⁴ A. Etzioni, "Designing AI Systems That Obey Our Laws and Values." *Communications of the ACM* 59, no. 9 (2016): 29–31.
- ²⁵ T. Arnold, D. Kasenberg, and M. Scheutz. "Value Alignment or Misalignment—What Will Keep Systems Accountable?" *The Workshops of the Thirty-First AAAI Conference on Artificial Intelligence: Technical Reports, WS-17-02: AI, Ethics, and Society*, 81–88. Palo Alto, CA: The AAAI Press, 2017.
- ²⁶ G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, eds. *Normative Multi-Agent Systems*. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013.
- ²⁷ A. Chaudhuri, (2017) Philosophical Dimensions of Information and Ethics in the Internet of Things (IoT) Technology. The EDP Audit, Control, and Security Newsletter, 56:4, 7-18, DOI: 10.1080/07366981.2017.1380474
- ²⁸ S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2, no. 6 (2017): ean6080. doi:10.1126/scirobotics. ean6080
- ²⁹ A. D. Selbst and S. Barocas, The Intuitive Appeal of Explainable Machines (February 19, 2018). *Fordham Law Review*. Available at SSRN: <https://ssrn.com/abstract=3126971> or <http://dx.doi.org/10.2139/ssrn.3126971>
- ³⁰ F. S. Grodzinsky, K. W. Miller, and M. J. Wolf. "Developing Artificial Agents Worthy of Trust: Would You Buy a Used Car from This Artificial Agent?" *Ethics and Information Technology* 13, (2011): 17–27.

Embedding Values into Autonomous and Intelligent Systems

- ³¹ J. A. Kroll, J. Huey, S. Barocas et al. "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (2017).
- ³² J. Cleland-Huang, O. Gotel, and A. Zisman, eds. *Software and Systems Traceability*. London: Springer, 2012. doi:10.1007/978-1-4471-2239-5
- ³³ S. U. Noble, "Google Search: Hyper-Visibility as a Means of Rendering Black Women and Girls Invisible." *InVisible Culture* 19 (2013).
- ³⁴ K. R. Fleischmann and W. A. Wallace. "A Covenant with Transparency: Opening the Black Box of Models." *Communications of the ACM* 48, no. 5 (2005): 93–97.
- ³⁵ M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.
- ³⁶ M. Hind, et al. "Increasing Trust in AI Services through Supplier's Declarations of Conformity." *ArXiv E-Prints*, Aug. 2018. Retrieved October 28, 2018 from <https://arxiv.org/abs/1808.07261>.
- ³⁷ Vitsoe. "The Power of Good Design." *Vitsoe*, 2018. Retrieved Oct 22, 2018 from <https://www.vitsoe.com/us/about/good-design>.
- ³⁸ G. Donelli, (2015, March 13). Good design is honest (Blogpost). Retrieved Oct 22, 2018 from <https://blog.astropad.com/good-design-is-honest/>
- ³⁹ C. de Jong Ed., "Ten principles for good design: Dieter Rams." New York, NY: Prestel Publishing, 2017.
- ⁴⁰ Ibid.
- ⁴¹ N. Tintarev and R. Kutlak. "Demo: Making Plans Scrutable with Argumentation and Natural Language Generation." *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces* (2014): 29–32.
- ⁴² d. boyd, "[Transparency ≠ Accountability](#)." *Data & Society: Points*, November 29, 2016.
- ⁴³ C. Oetzel and S. Spiekermann, "A Systematic Methodology for Privacy Impact Assessments: A Design Science Approach." *European Journal of Information Systems* 23, (2014): 126–150. <https://link.springer.com/article/10.1057/ejis.2013.18>
- ⁴⁴ M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfunkel, A. Dafoe, P. Scharre, T. Zeitzo, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. CoRR abs/1802.07228 (2018). <https://arxiv.org/abs/1802.07228M>.
- ⁴⁵ D. Vanderelst and A.F. Winfield, 2018 The Dark Side of Ethical Robots. In *Proc. AAAI/ACM Conf. on Artificial Intelligence, Ethics and Society*, New Orleans.
- ⁴⁶ British Standards Institution. BS8611:2016, "[Robots and Robotic Devices. Guide to the Ethical Design and Application of Robots and Robotic Systems](#)," 2016.
- ⁴⁷ I. Sommerville, *Software Engineering* (10th edition). Harlow, U.K.: Pearson Studium, 2015.
- ⁴⁸ M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.

Embedding Values into Autonomous and Intelligent Systems

⁴⁹ International Organization for Standardization (2015). ISO 9001:2015, Quality management systems—Requirements. Retrieved July 12, 2018 from <https://www.iso.org/standard/62085.html>.

⁵⁰ BMC Software. *ITIL: The Beginner's Guide to Processes & Best Practices*. 6 Dec. 2016, <http://www.bmc.com/guides/itil-introduction.html>.

⁵¹ J. Griffiths, "[New Zealand Passport Robot Thinks This Asian Man's Eyes Are -Closed.](#)" CNN.com, December 9, 2016.

⁵² R. Tatman, "[Google's Speech Recognition Has a Gender Bias.](#)" Making Noise and Hearing Things, July 12, 2016.

⁵³ J. Angwin, J. Larson, S. Mattu, L. Kirchner. "[Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.](#)" ProPublica, May 23, 2016.

⁵⁴ L. D. Riek and D. Howard. "[A Code of Ethics for the Human-Robot Interaction Profession.](#)" Proceedings of We Robot, April 4, 2014.

⁵⁵ M. Fisher, L. A. Dennis, and M. P. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56 (2013): 84–93.

Policy

Introduction

Autonomous and intelligent systems (A/IS) are a part of our society. The use of these powerful technologies promotes a range of social benefits. They may spur development across economies and society through numerous applications, including in commerce, finance, employment, health care, agriculture, education, transportation, politics, privacy, public safety, national security, civil liberties, and human rights. To encourage the development of socially beneficial applications of A/IS, and to protect the public from adverse consequences of A/IS, intended or otherwise, effective policies and government regulations are needed.

Effective A/IS policies serve the public interest in several important respects. A/IS policies and regulations, at both the national level and as developed by professional organizations and governing institutions, protect and promote safety, privacy, human rights, and cybersecurity, as well as enhance the public's understanding of the potential impacts of A/IS on society. Without policies designed with these considerations in mind, there may be critical technology failures, loss of life, and high-profile social controversies. Such events could engender policies that unnecessarily hinder innovation, or regulations that do not effectively advance public interest and protect human rights.

We believe that effective A/IS policies should embody a rights-based approach¹ that addresses five issues:

1. Ensure that A/IS support, promote, and enable internationally recognized legal norms.

Establish policies for A/IS using the internationally recognized legal framework for human rights standards that is directed at accounting for the impact of technology on individuals.

Policy

2. Develop government expertise in A/IS.

Facilitate skill development, technical and otherwise, to further boost the ability of policy makers, regulators, and elected officials to make informed proposals and decisions about the various facets of these new technologies.

3. Ensure governance and ethics are core components in A/IS research, development, acquisition, and use.

Require support for A/IS research and development (R&D) efforts with a focus on the ethical impact of A/IS. To benefit from these new technologies while also ensuring they meet societal needs and values, governments should be actively involved in supporting relevant R&D efforts.

4. Create policies for A/IS to ensure public safety and responsible A/IS design.

Governments must ensure consistent and locally adaptable policies and regulations for A/IS. Effective regulation should address transparency, explainability, predictability, bias, and accountability for A/IS algorithms, as well as risk management, privacy, data protection measures, safety, and security considerations. Certification of systems involving A/IS is a key technical, societal, and industrial issue.

5. Educate the public on the ethical and societal impacts of A/IS.

Industry, academia, the media, and governments must establish strategies for informing and engaging the public on benefits and challenges posed by A/IS. Communicating accurately both the positive potential of A/IS and the areas that require caution and further development is critical to effective decision-making environments.

As A/IS comprise a greater part of our daily lives, managing the associated risks and rewards becomes increasingly important. Technology leaders and policy makers have much to contribute to the debate on how to build trust, promote safety and reliability, and integrate ethical and legal considerations into the design of A/IS technologies. This chapter provides a principled foundation for these discussions.

Policy

Issue 1: Ensure that A/IS support, promote, and enable internationally recognized legal norms

Background

A/IS technologies have the potential to impact internationally recognized economic, social, cultural, and political rights through unintended outcomes and outright design decisions. Important examples of this issue have occurred with certain unmanned aircraft systems (Bowcott 2013), use of A/IS in predictive policing (Shapiro 2017), banking (Garcia 2017), judicial sentencing (Osoba and Welser 2017), and job hunting and hiring practices (Datta, Tschantz, and Datta 2014). Even service delivery of goods (Ingold and Soper 2016) can impact human rights by automating discrimination (Eubanks 2018) and inhibiting the right of assembly, freedom of expression, and access to information. To ensure A/IS are used as a force for social benefit, nations must develop policies that safeguard human rights.

A/IS regulation, development, and deployment should, therefore, be based on international human rights standards and standards of international humanitarian laws. When put into practice, both states and private actors will consider their responsibilities to protect and respect internationally recognized political, social, economic, and cultural rights. Similarly, business actors will consider their obligations to respect international human rights, as described in the United Nations Guiding Principles on Business and Human Rights (OHCHR 2011), also known as the Ruggie principles.

The Ruggie principles have been widely referenced and endorsed by corporations and have led to the adoption of several corporate social responsibility (CSR) policies in various companies. With broadened support, the Ruggie principles will strengthen the role of businesses in protecting and promoting human rights and ensuring that the most crucial human values and legal standards of human rights are respected by A/IS technologists.

Recommendations

National policies and business regulations for A/IS should be founded on a rights-based approach. The Ruggie principles provide the internationally recognized legal framework for human rights standards that accounts for the impact of technology on individuals while also addressing inequalities, discriminatory practices, and the unjust distribution of resources.

These six considerations for a rights-based approach to A/IS flow from the recommendation above:

- *Responsibility:* Identify the right holders and the duty bearers and ensure that duty bearers have an obligation to fulfill all human rights.
- *Accountability:* Oblige states, as duty bearers, to behave responsibly, to seek to represent the greater public interest, and to be open to public scrutiny of their A/IS policies.
- *Participation:* Encourage and support a high degree of participation of duty bearers, right holders, and other interested parties.

Policy

- *Nondiscrimination*: Underlie the practice of A/IS with principles of nondiscrimination, equality, and inclusiveness. Particular attention must be given to vulnerable groups, to be determined locally, such as minorities, indigenous peoples, or persons with disabilities.
- *Empowerment*: Empower right holders to claim and exercise their rights.
- *Corporate responsibility*: Ensure that companies' developments of A/IS comply with the rights-based approach. Companies must not willingly provide A/IS to actors that will use them in ways that lead to human rights violations.
- O. A. Osoba, and W. Welser IV, "[An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence](#)," (Research Report 1744). Santa Monica, CA: RAND Corporation, 2017.
- A. Datta, M. C. Tschantz, and A. Datta. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," arXiv:1408.6491 [Cs], 2014.
- D. Ingold, and S. Soper, "[Amazon Doesn't Consider the Race of Its Customers. Should It?](#)" *Bloomberg*, April 21, 2016.
- United Nations. Office of the High Commissioner of Human Rights. [Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework](#). United Nations Office of the High Commissioner of Human Rights. New York and Geneva: UN, 2011.

Further Resources

- Human rights-based approaches have been applied to development, education and reproductive health. See the [UN Practitioners' Portal on Human Rights Based Programming](#).
- O. Bowcott, "[Drone Strikes by US May Violate International Law, Says UN](#)," *The Guardian*, October 18, 2013.
- A. Shapiro, "[Reform Predictive Policing](#)," *Nature News*, vol. 541, no. 7638, pp. 458–460, Jan. 25, 2017.
- M. Garcia, "[How to Keep Your AI from Turning Into a Racist Monster](#)," *Wired*, April 21, 2017.
- "[Mapping Regulatory Proposals for Artificial Intelligence in Europe](#)." Access Now, November 2018.
- V. Eubanks, *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, January 2018.

Policy

Issue 2: Develop government expertise in A/IS

Background

There is a consensus among private sector and academic stakeholders that effectively governing A/IS and related technologies requires a level of technical expertise that governments currently do not possess. Effective governance requires experts who understand and can analyze the interactions between A/IS technologies, policy objectives, and overall societal values. Sufficient depth and breadth of technical expertise will help ensure policies and regulations successfully support innovation, adhere to national principles, and protect public safety.

Effective governance also requires an A/IS workforce that has adequate training in ethics and access to other resources on human rights standards and obligations, along with guidance on how to apply them in practice.

Recommendations

Policy makers should support the development of expertise required to create a public policy, legal, and regulatory environment that allows innovation to flourish while protecting the public and gaining public trust.² Example strategies include the following:

- Expertise can be furthered through technical fellowships, or rotation schemes, where technologists spend an extended time in political offices, or policy makers work with

organizations³ that operate at the intersection of technology policy, technical engineering, and advocacy. This will enhance the technical knowledge of policy makers, strengthen ties between political and technical communities, and contribute to the formulation of effective A/IS policy.

- Expertise can also be developed through cross-border sharing of best practices around A/IS legislation, consumer protection, workforce transformation, and economic displacement stemming from A/IS-based automation. This can be done through governmental cooperation, knowledge exchanges, and by building A/IS components into venues and efforts surrounding existing regulation, e.g., the General Data Protection Regulation (GDPR).
- Because A/IS involve rapidly evolving technologies, both workforce training in A/IS areas and long-term science, technology, engineering, and math (STEM) educational strategies, along with ethics courses, are needed beginning in primary school and extending into university or vocational courses. These strategies will foster A/IS expertise in the next generation of many groups, e.g., supervisors of critical systems, scientists, and policy makers.

Policy

Further Resources

- J. Holdren, and M. Smith, "[Preparing for the Future of Artificial Intelligence](#)." Washington, DC: Executive Office of the President, National Science and Technology Council, 2016.
- P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller. "[Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence](#)." (Report of the 2015-2016 Study Panel). Stanford, CA: Stanford University, 2016.
- "[Japan Industrial Policy Spotlights AI, Foreign Labor](#)." *Nikkei Asian Review*, May 20, 2016.
- Y.H. Weng, "[A European Perspective on Robot Law: Interview with Mady Delvaux-Stehres](#)." *Robohub*, July 15, 2016.

Issue 3: Ensure governance and ethics are core components in A/IS research, development, acquisition, and use.

Background

Greater national investment in ethical A/IS research and development would stimulate the economy, create high-value jobs, improve governmental services to society, and encourage international innovation and collaboration (U.S. OSTP report on the Future of AI 2016). A/IS have the potential to improve our societies through

technologies such as intelligent robots and self-driving cars that will revolutionize automobile transportation and logistics systems and reduce traffic fatalities. A/IS can improve quality of life through smart cities and decision support in health care, social services, criminal justice, and the environment. To ensure such a positive effect on individuals, societies, and businesses, nations must increase A/IS R&D investments, with particular focus on the ethical development and deployment of A/IS.

International collaboration involving governments, private industry, and non-governmental organizations (NGOs) would promote the development of standards, data sharing, and norms that guide ethically aligned A/IS R&D.

Recommendations

Develop national and international standards for A/IS to enable efficient and effective public and private sector investments. Important aspects for international standards include measures of societal benefits derived from A/IS, the use of ethical considerations in A/IS investments, and risks increased or decreased by A/IS. Nations should consider their own ethical principles and develop a framework for ethics that each country could use to reflect local systems of values and laws. This will encourage actors to think both locally and globally regarding ethics. Therefore, we recommend governments to:

- Establish priorities for funding A/IS research that identify approaches and challenges for A/IS governance. This research will identify models for national and global A/IS governance and assess their benefits and adequacy to address A/IS societal needs.

Policy

- Encourage the participation of a diverse set of stakeholders in the standards development process. Standards should address A/IS issues such as fairness, security, transparency, understandability, privacy, and societal impacts of A/IS. A global framework for identification and sharing of these and other issues should be developed. Standards should incorporate independent mechanisms to properly vet, certify, audit, and assign accountability for the A/IS applications.
- Encourage and establish national and international research groups that provide incentives for A/IS research that is publicly beneficial but may not be commercially viable.
- The Networking and Information Technology Research and Development Program, [“Supplement to the President’s Budget, FY2017.”](#) NITRD National Coordination Office, April 2016.
- S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, “The SpiNNaker Project.” *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.
- H. Markram, “The Human Brain Project,” *Scientific American*, vol. 306, no. 2, pp. 50–55, June 2012.
- L. Yuan, [“China Gears Up in Artificial-Intelligence Race.”](#) *Wall Street Journal*, August 24, 2016.

Further Resources

- E. T. Kim, [“How an Old Hacking Law Hampers the Fight Against Online Discrimination.”](#) *The New Yorker*, October 1, 2016.
- National Research Council. “Developments in Artificial Intelligence, Funding a Revolution: Government Support for Computing Research.” Washington, DC: The National Academies Press, 1999.
- N. Chen, L. Christensen, K. Gallagher, R. Mate, and G. Rafert, [“Global Economic Impacts Associated with Artificial Intelligence.”](#) Analysis Group, February 25, 2016.

Issue 4: Create policies for A/IS to ensure public safety and responsible A/IS design

Background

Effective governance encourages innovation and cooperation, helps synchronize policies globally, and reduces barriers to trade. Governments must ensure consistent and appropriate policies and regulations for A/IS that address transparency, explainability, predictability, and accountability of A/IS algorithms, risk management,⁴ data protection, safety, and certification of A/IS.

Appropriate regulatory responses are context-dependent and should be developed through an approach that is based on human rights⁵ and has human well-being as a key goal.

Policy

Recommendations

Nations should develop and harmonize their policies and regulations for A/IS using a process that is based on informed input from a range of expert stakeholders, including academia, industry, NGOs, and government officials, that addresses questions related to the governance and safe deployment of A/IS. We recommend:

- Policy makers should consider similar work from around the world. Due to the transnational nature of A/IS, globally synchronized policies can benefit public safety, technological innovation, and access to A/IS.
- Policies should foster the development of economies able to absorb A/IS. Additional focus is needed to address the effect of A/IS on employment and income and how to ameliorate certain societal conditions. New models of public-private partnerships should be studied.
- Policies for A/IS should remain founded on a rights-based approach.
- Policy makers should be prepared to address issues that will arise when innovative and new practices enabled by A/IS are not consistent with current law. In A/IS, where there is often a different system developer, integrator, user, and ultimate customer, application of traditional legal concepts of agency, strict liability, and parental liability will require legal research and deliberation. Challenges from A/IS that must be considered include increasing complexity of and interactions between systems, and the potential for reduced predictability due to the nature of machine learning systems.

Further Resources

- P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller. "[Artificial Intelligence and Life in 2030': One Hundred Year Study on Artificial Intelligence](#)." (Report of the 2015-2016 Study Panel). Stanford, CA: Stanford University, 2016.
- R. Calo, "[The Case for a Federal Robotics Commission](#)," The Brookings Institution, 2014.
- O. Groth, and Mark Nitzberg, *Solomon's Code: Humanity in a World of Thinking Machines* (chapter 8 on governance), New York: Pegasus Books, 2018.
- A. Mannes, "[Institutional Options for Robot Governance](#)," 1–40, in *We Robot 2016*, Miami, FL, April 1–2, 2016.
- G. E. Marchant, K. W. Abbott, and B. Allenby, *Innovative Governance Models for Emerging Technologies*. Cheltenham, U.K.: Edward Elgar Publishing, 2014.
- Y. H. Weng, Y. Sugahara, K. Hashimoto, and A. Takanishi. "Intersection of 'Tokku' Special Zone, Robots, and the Law: A Case Study on Legal Impacts to Humanoid Robots," *International Journal of Social Robotics* 7, no. 5, pp. 841–857, 2015.

Policy

Issue 5: Educate the public on the ethical and societal impacts of A/IS

Background

It is imperative for industry, academia, and government to communicate accurately to the public both the positive and negative potential of A/IS and the areas that require caution.⁶ Strategies for informing and engaging the public on A/IS benefits and challenges are critical to creating an environment conducive to effective decision-making.

Educating users of A/IS will help influence the nature of A/IS development. Educating policy makers and regulators on the technical and legal aspects of A/IS will help enable the creation of well-defined policies that promote human rights, safety, and economic benefits. Educating corporations, researchers, and developers of A/IS on the benefits and risks to individuals and societies will enhance the creation of A/IS that better serve human well-being.⁷

Another key requirement is that A/IS are sufficiently transparent regarding implicit and explicit values and algorithmic processes. This is necessary for the public understanding of A/IS accountability, predictions, decisions, biases, and mistakes.

Recommendations

Establish an international multi-stakeholder forum, to include commercial, governmental, and other civil society groups, to determine the best practices for using and developing A/IS. Codify the deliberations into international norms and standards. Many industries—in particular, system industries (automotive, air and space, defense, energy, medical systems, manufacturing)—will be changed by the growing use of A/IS. Therefore, we recommend governments to:

- Increase funding for interdisciplinary research and communication on topics ranging from basic research on intelligence to principles of ethics, safety, privacy, fairness, liability, and trustworthiness of A/IS. Societal aspects should be addressed both at an academic level and through the engagement of business, civil society, public authorities, and policy makers.
- Empower and enable independent journalists and media outlets to report on A/IS by providing access to technical expertise.
- Conduct educational outreach to inform the public on A/IS research, development, applications, risks and rewards, along with the policies, regulations, and testing that are designed to safeguard human rights and public safety.

Policy

Develop a broad range of A/IS educational programs. Undergraduate, professional degree, advanced degree, and executive education programs should offer instruction that ensures lawyers, legislators, and A/IS workers are well informed about issues arising from A/IS, including the need for measurable standards of A/IS performance, effects, and ethics, and the need to mature the still nascent capabilities to measure these elements of A/IS.

Further Resources

- Networking and Information Technology Research and Development (NITRD) Program, "[The National Artificial Intelligence Research and Development Strategic Plan](#)," Washington, DC: Office of Science and Technology Policy, 2016.
- J. Saunders, P. Hunt, and J. S. Hollywood, "[Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot](#)," *Journal of Experimental Criminology*, vol. 12, no. 347, pp. 347–371, 2016. [Online] Available: doi:10.1007/s11292-019272-0. [Accessed Nov. 10, 2018].
- B. Edelman and M. Luca, "[Digital Discrimination: The Case of Airbnb.com](#)," Harvard Business School Working Paper 14-054, Jan. 28, 2014.
- C. Garvie, A. Bedoya, and J. Frankle, "[The Perpetual Line-Up: Unregulated Police Face Recognition in America](#)," Washington, DC: Georgetown Law, Center on Privacy & Technology, 2016.
- M. Chui, and J. Manyika, "[Automation, Jobs, and the Future of Work](#)," Seattle, WA: McKinsey Global Institute, 2014.
- R. C. Arkin, "[Ethics and Autonomous Systems: Perils and Promises \[Point of View\]](#)," *Proceedings of the IEEE* 104, no. 10, pp. 1779–1781, Sept. 19, 2016.
- [European Commission, Eurobarometer Survey on Autonomous Systems](#) (DG Connect, June 2015), looks at Europeans' attitudes toward robots, driverless vehicles, and autonomous drones. The survey shows that those who have more experience with robots (at home, at work or elsewhere) are more positive toward their use.

Policy

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Policy Committee

- **Kay Firth-Butterfield** (Founding Co-Chair) – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Dr. Peter S. Brooks** (Co-Chair) – Institute for Defense Analyses
- **Mina Hanna** (Co-Chair) – Chair IEEE-USA Artificial Intelligence and Autonomous Systems Policy Committee, Vice Chair IEEE-USA Research and Development Policy Committee, Member of the Editorial Board of IEEE Computer Magazine
- **Chloe Autio** – Government & Policy Group, Intel Corporation
- **Stan Byers** – Frontier Markets Specialist
- **Corinne Cath-Speth** – PhD student at Oxford Internet Institute, The University of Oxford, Doctoral student at the Alan Turing Institute, Digital Consultant at ARTICLE 19
- **Michelle Finneran Dennedy** – Vice President, Chief Privacy Officer, Cisco; Author,

The Privacy Engineer's Manifesto: Getting from Policy to Code to QA to Value

- **Eileen Donahoe** – Executive Director of Stanford Global Digital Policy Incubator
- **Danit Gal** – Project Assistant Professor, Keio University; Chair, IEEE Standard P7009 on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- **Olaf J. Groth** – Professor of Strategy, Innovation, Economics & Program Director for Disruption Futures, HULT International Business School; Visiting Scholar, UC Berkeley BRIE/CITRIS; CEO, Cambrian.ai
- **Philip Hall** – (Founding Co-Chair) Co-Founder & CEO, RelmaTech; Member (and Immediate Past Chair), IEEE-USA Committee on Transportation & Aerospace Policy (CTAP); and Member, IEEE Society on Social Implications of Technology
- **John C. Havens** – Executive Director, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems; Executive Director, The Council on Extended Intelligence; Author, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*
- **Cyrus Hodes** – Senior Advisor, AI Office, UAE Prime Minister's Office; Co-Founded at Harvard Kennedy School the AI Initiative; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence.

Policy

- **Chihyung Jeon** – Assistant Professor, Graduate School of Science and Technology Policy (STP), Korea Advanced Institute of Science and Technology (KAIST)
- **Anja Kaspersen** – Former Head of International Security, World Economic Forum and head of strategic engagement and new technologies at the international committee of Red Cross (ICRC)
- **Nicolas Mialhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs.
- **Simon Mueller** – Executive Director, The AI Initiative; Vice President, The Future Society
- **Carolyn Nguyen** – Director, Microsoft's Technology Policy Group, responsible for policy initiatives related to data governance and personal data
- **Mark J. Nitzberg** – Executive Director, Center for Human-Compatible Artificial Intelligence at UC Berkeley; co-author, *Solomon's Code: Humanity in a World of Thinking Machines*
- **Daniel Schiff** – PhD Student, Georgia Institute of Technology; Chair, Sub-Group for Autonomous and Intelligent Systems Implementation, IEEE P7010: Well-being Metric for Autonomous and Intelligent Systems
- **Evangelos Simoudis** – Co-Founder and Managing Director, Synapse Partners. Author, *The Big Data Opportunity in our Driverless Future*
- **Brian W. Tang** – Founder and Managing Director, Asia Capital Markets Institute (ACMI); Founding executive director, LITE Lab@HKU at Hong Kong University Faculty of Law
- **Martin Tisné** – Managing Director, Luminate
- **Sarah Villeneuve** – Policy Analyst; Member, IEEE P7010: Well-being Metric for Autonomous and Intelligent Systems
- **Adrian Weller** – Senior Research Fellow, University of Cambridge; Programme Director for AI, The Alan Turing Institute
- **Yueh-Hsuan Weng** – Assistant Professor, Frontier Research Institute for Interdisciplinary Sciences (FRIS), Tohoku University; Fellow, Transatlantic Technology Law Forum (TTLF), Stanford Law School
- **Darrell M. West** – Vice President and Director, Governance Studies | Founding Director, Center for Technology Innovation | The Douglas Dillon Chair, Brookings Institution
- **Andreas Wolkenstein** – Researcher on neurotechnologies, AI, and political philosophy at LMU Munich (Germany)

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Policy

Endnotes

- ¹ This approach is rooted in internationally recognized economic, social, cultural, and political rights.
- ² This recommendation concurs with the multiple recommendations of the United States National Science and Technology Council, One Hundred Year Study of Artificial Intelligence, Japan's Cabinet Office Council, European Parliament's Committee on Legal Affairs, and others.
- ³ For example, American Civil Liberties Union, Article 19, the Center for Democracy & Technology, Canada.AI, or Privacy International. United Nations committees may also be useful in fostering knowledge exchanges.
- ⁴ This includes consideration regarding application of the precautionary principle, as used in environmental and health policy-making, where the possibility of widespread harm is high and extensive scientific knowledge or understanding on the matter is lacking.
- ⁵ Human rights-based approaches have been applied to development, education, and reproductive health. See the UN Practitioners' Portal on Human Rights Based Programming.
- ⁶ "(AI100)," Stanford University., August 2016.
- ⁷ Private sector initiatives are already emerging, such as the Partnership on AI; the AI for Good Foundation; and the Ethics and Governance of Artificial Intelligence Initiative, launched by Harvard's Berkman Klein Center for Internet & Society and the MIT Media Lab.

Law

The law affects and is affected by the development and deployment of autonomous and intelligent systems (A/IS) in contemporary life. Science, technological development, law, public policy, and ethics are not independent fields of activity that occasionally overlap. Instead, they are disciplines that are fundamentally tied to each other and collectively interact in the creation of a social order.

Accordingly, in studying A/IS and the law, we focus not only on how the law responds to the technological innovation represented by A/IS, but also on how the law guides and sets the conditions for that innovation. This interactive process is complex, and its desired outcomes can rest on particular legal and cultural traditions. While acknowledging this complexity and uncertainty, as well as the acute risk that A/IS may intentionally or unintentionally be misused or abused, we seek to identify principles that will steer this interactive process in a manner that leads to the improvement, prosperity, and well-being of everyone.

The fact that the law has a unique role to play in achieving this outcome is observed by Sheila Jasanoff, a preeminent scholar of science and technology studies:

Part of the answer is to recognize that science and technology—for all their power to create, preserve, and destroy—are not the only engines of innovation in the world. Other social institutions also innovate, and they may play an invaluable part in realigning the aims of science and technology with those of culturally disparate human societies. Foremost among these is the law.¹

The law can play its part in ensuring that A/IS, in both design and operation, are aligned with principles of ethics and human well-being.²

Comprehensive coverage of all issues within our scope of study is not feasible in a single chapter of *Ethically Aligned Design (EAD)*. Accordingly, aggregate coverage will expand as issues not yet studied are selected for treatment in future versions of *EAD*.

Law

EAD, First Edition includes commentary about how the law should respond to a number of specific ethical and legal challenges raised by the development and deployment of A/IS in contemporary life. It also focuses on the impact of A/IS *on the practice of law itself*. More specifically, we study both the potential benefits and the potential risks resulting from the incorporation of A/IS into a society's legal system—specifically, in law making, civil justice, criminal justice, and law enforcement. Considering the results of those inquiries, we endeavor to identify norms for the adoption of A/IS in a legal system that will enable the realization of the benefits while mitigating the risks.³

In this chapter of EAD, we include the following:

Section 1: Norms for the Trustworthy Adoption of A/IS in Legal Systems.

This section addresses issues raised by the potential adoption of A/IS in legal systems for the purpose of performing, or assisting in performing, tasks traditionally carried out by humans with specialized legal training or expertise. The section begins with the question of how A/IS, if properly incorporated into a legal system, can improve the functions of that legal system and thus enhance its ability to contribute to human well-being. The section then discusses challenges to the safe and effective incorporation of A/IS into a legal system and identifies **the chief challenge as an absence of informed trust**. The remainder of the section examines how societies can fill the trust gap by enacting policies and promoting practices that advance publicly accessible standards of **effectiveness, competence, accountability, and transparency**.

Section 2: Legal Status of A/IS.

This section addresses issues raised by the legal status of A/IS, including the potential assignment of certain legal rights and obligations to such systems. The section provides background on the issue and outlines some of the potential advantages and disadvantages of assigning some form of legal personhood to A/IS. Based on these considerations, the section concludes that extending legal personhood to A/IS is not appropriate at this time. It then considers alternatives and outlines certain future conditions that might warrant reconsideration of the section's central recommendation.

Law

Section 1: Norms for the Trustworthy Adoption of A/IS in Legal Systems⁴

"It's a day that is here."

John G. Roberts, Chief Justice of the Supreme Court of the United States, when asked in 2017 whether he could foresee a day when intelligent machines would assist with courtroom fact-finding or judicial decision-making.⁵

A/IS hold the potential to improve the functioning of a legal system and, thereby, to contribute to human well-being. That potential will be realized, however, only if both the use of A/IS and the avoidance of their use are grounded in solid information about the capabilities and limitations of A/IS, the competencies and conditions required for their safe and effective operation (including data requirements), and the lines along which responsibility for the outcomes generated by A/IS can be assigned. Absent that information, society risks both **uninformed adoption** of A/IS and **uninformed avoidance of adoption** of A/IS, risks that are particularly acute when A/IS are applied in an integral component of the social order, such as the law.

- **Uninformed adoption** poses the risk that A/IS will be applied to inform or replace the judgments of legal actors (legislators, judges, lawyers, law enforcement officers, and jurors) without controls to ensure their safe and effective operation. They may even be used

for purposes other than those for which the systems have been validated and vetted for legal use. In addition to actual harm to individuals, the result will be distrust, not only of the effectiveness of A/IS, but also of the fairness and effectiveness of the legal system itself.

- **Uninformed avoidance of adoption** poses the risk that a lack of understanding of what is required for the safe and effective operation of A/IS will result in blanket distrust of all forms and applications of A/IS, even those that are, when properly applied, safe and effective. The result will be a failure to realize the significant improvements in the legal system that A/IS can offer and a continuation of systems that are, even with the best of safeguards, still subject to human bias, inconsistency, and error.⁶

In this section, we consider how society can address these risks by developing norms for the adoption of A/IS in legal systems. The specific issues discussed follow. The first and second issues reflect the potential benefits of, and challenges to, trustworthy adoption of A/IS in the world's legal systems. The remaining issues discuss four principles,⁷ which, if adhered to, will enable trustworthy adoption.^{8 9}

Law

- **Issue 1: Well-being, Legal Systems, and A/IS**—How can A/IS improve the functioning of a legal system and, thereby, enhance human well-being?
- **Issue 2: Impediments to Informed Trust**—What are the challenges to adopting A/IS in legal systems and how can those impediments be overcome?
- **Issue 3: Effectiveness**—How can the collection and disclosure of evidence of effectiveness of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 4: Competence**—How can specification of the knowledge and skills required of the human operator(s) of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 5: Accountability**—How can the ability to apportion responsibility for the outcome of the application of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?
- **Issue 6: Transparency**—How can sharing information that explains how A/IS reach given decisions or outcomes foster informed trust in the suitability of A/IS for adoption in legal systems?

Issue 1: Well-Being, Legal Systems, and A/IS

How can A/IS improve the functioning of a legal system and, thereby, enhance human well-being?

Background

An effective legal system contributes to human well-being. The law is an integral component of social order; the nature of a legal system informs, in fundamental ways, the nature of a society, its potential for economic growth and technological innovation, and its capacity for advancing the well-being of its members.

If the law is a constitutive element of social order, it is not surprising that it also plays a key role in setting the conditions for well-being and economic growth. In part, this flows from the fact that a well-functioning legal system is an element of good governance. Good governance and a well-functioning legal system can help society and its members flourish, as measured by indicators of both economic prosperity¹⁰ and human well-being.¹¹ The attributes of good governance can be defined in several ways. Good governance can mean democracy; the observance of norms of human rights enshrined in conventions such as the Universal Declaration of Human Rights¹² and the Convention of the Rights of the Child;¹³ and constitutional constraints on government power. It can also

Law

mean bureaucratic competence, law and order, property rights, and contract enforcement.

The United Nations (UN) defines the rule of law as:

a principle of governance in which all persons, institutions and entities, public and private, including the State itself, are accountable to laws that are publicly promulgated, equally enforced and independently adjudicated. . . . It requires, as well, measures to ensure adherence to the principles of supremacy of law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness and procedural and legal transparency.¹⁴

Orderly systems of legal rules and institutions generally correlate positively with economic prosperity, social stability, and human well-being, including the protection of childhood.¹⁵ Studies from the World Bank suggest that legal reforms can lead to increased foreign investment, higher incomes, and greater wealth.¹⁶ Wealth, in turn, can enable policies that support improved education, health, environmental protection, equal opportunity, and, in democratic societies, greater individual freedom.

Law, moreover, can contribute to prosperity not only through its functional attributes, but also through its substantive content. Patent laws, for example, if well-designed, can encourage technological innovation, leading to increases in productivity and the economic growth that follows. Poorly designed patent laws, on the

other hand, may foster monopolistic markets and decrease competition, resulting in a decreased pace of technological innovation, fewer gains in productivity, and slower economic growth.¹⁷

While economic growth is a valuable benefit of a well-designed and well-functioning legal system, it is not the only benefit. Such a system can bring benefits to society and its members that, beyond economic prosperity, extend to mental and physical well-being. Specific benefits include the protection and advancement of an individual's dignity,¹⁸ human rights,¹⁹ liberty, stability, security, equality of treatment under the law, and ability to provide for the future.²⁰

In fact, recent thinking on the relationship between law and economic development has come to hold that a well-functioning legal system is not simply a *means* to development but *is* development, insofar as such a system is a constitutive element of a social order that protects and advances human dignity, rights, and well-being. As this position has been characterized by David Kennedy:

... the focal point for development policy was increasingly provided less by economics than from ideas about the nature of the good state themselves provided by literatures of political science, political economy, ethics, social theory, and law. In particular, "human rights" and the "rule of law"²¹ became substantive definitions of development. One should promote human rights not to *facilitate* development—but *as* development. The rule of law was not a development *tool*—it was itself a development

Law

objective. Increasingly, law—understood as a combination of human rights, courts, property rights, formalization of entitlements, prosecution of corruption, and public order—came to define development.²²

While this shift from considering law as a means to an end to considering law as an end in itself has been criticized on the grounds that it takes the focus off the difficult political choices that are inherent in any development policy,²³ it remains true that a well-functioning legal system is essential to the realization of a social order that protects and advances human dignity, rights, and well-being.

A/IS can contribute to the proper functioning of a legal system. A properly functioning legal system, one that is conducive to both economic prosperity and human well-being, will have a number of attributes. It should be:

- **Speedy:** enable quick resolution of civil and criminal cases;
- **Fair:** produce results that are just and proportionate to circumstance;²⁴
- **Free from undesirable bias:** operate without prejudice;
- **Consistent:** arrive at outcomes in a principled, consistent, and nonarbitrary manner;
- **Transparent:** be open to appropriate public examination and oversight;²⁵
- **Accessible:** be equally open to all citizens and residents in resolving disputes;
- **Effective:** achieve the ends intended by its laws and rules without negative collateral consequences;²⁶

- **Accurate:** achieve accurate results, minimizing both false positives (persons unjustly or incorrectly targeted, investigated, or sentenced for crimes) and false negatives (persons incorrectly *not* targeted, investigated, or sentenced for crimes);
- **Adaptable:** have the flexibility to adapt to changes in societal circumstances.

A/IS have the potential to alter the overall functioning of a legal system. A/IS, applied responsibly and appropriately, could improve the legislative process, enhance access to justice, accelerate judicial decision-making, provide transparent and readily accessible information on why and how decisions were reached, reduce bias, support uniformity in judicial outcomes, help society identify (and potentially correct) judicial errors, and improve public confidence in the legal system. By way of example:

- A/IS can make legislation and regulation more **effective** and **adaptable**. For lawmaking, A/IS could help legislators analyze data to craft more finely tuned, responsive, evidence-based laws and regulations. This could, potentially, offer self-correcting suggestions to legislators (and to the general public) to help inform dialogue on how to meet defined public policy objectives.
- A/IS can make the practice of law more effective and efficient. For example, A/IS can enhance the **speed, accuracy, and accessibility** of the process of fact-finding in legal proceedings. When used appropriately in legal fact-finding, particularly in jurisdictions that allow extensive discovery or disclosure, A/IS already make litigation and investigations more accessible by analyzing vast data

Law

collections faster, more efficiently, and potentially more effectively²⁷ than document analysis conducted solely by human attorneys. By making fact-finding in an era of big data progressively easier, faster, and cheaper, A/IS may facilitate access to justice for parties that otherwise may find using the legal system to resolve disputes cost-prohibitive. A/IS can also help ensure that justice is rendered based on better accounting of the facts, thus serving the central purpose of any legal system.

- In both civil and criminal proceedings, A/IS can be used to improve the **accuracy**, **fairness**, and **consistency** of decisions rendered during proceedings. A/IS could serve as an auditing function for both the civil and criminal justice systems, helping to identify and correct judicial and law enforcement errors.²⁸
- A/IS can increase the **speed**, **accuracy**, **fairness**, **freedom from bias**, and general **effectiveness** with which law enforcement resources are deployed to combat crime. A/IS could be used to reduce or prevent crime, respond more quickly to crimes in progress, and improve collaboration among different law enforcement agencies.²⁹
- A/IS can help ensure that determinations about the arrest, detention, and incarceration of individuals suspected of, or convicted of, violations of the law are **fair**, **free from bias**, **consistent**, and **accurate**. Automated risk assessment tools have the potential to address issues of systemic racial bias in sentencing, parole, and bail determination while also safely reducing incarceration and recidivism

rates by identifying individuals who are less likely to commit crimes if released.

- A/IS can help to ensure that the tools, procedures, and resources of the legal system are more **transparent** and **accessible** to citizens. For the ordinary citizen, A/IS can democratize access to legal expertise, especially in smaller matters, where they may provide effective, prompt, and low-cost initial guidance to an aggrieved party; for example, in landlord-tenant, product purchase, employment, or other contractual contexts where the individual often tends to find access to legal information and legal advice prohibitive, or where asymmetry of resources between the parties renders recourse to the legal system inequitable.³⁰

A/IS have the potential to improve how a legal system functions in fundamental ways. As is the case with all powerful tools, there are some risks. **A/IS should not be adopted in a legal system without due care and scrutiny;** they should be adopted after a society's careful reflection and proper examination of evidence that their deployment and operation can be trusted to advance human dignity, rights, and well-being (see Issues 2–6).

Recommendations³¹

1. Policymakers should, in the interest of improving the function of their legal systems and bringing about improvements to human well-being, explore, through a broad consultative dialogue with all stakeholders, how A/IS can be adopted for use in their legal systems. They should do

Law

so, however, only in accordance with norms for adoption that mitigate the risks attendant on such adoption (see Issues 2–6 in this section).

2. Governments, non-governmental organizations, and professional associations should support educational initiatives designed to create greater awareness among all stakeholders of the potential benefits and risks of adopting A/IS in the legal system, and of the ways of mitigating such risks. A particular focus of these initiatives should be the ordinary citizen who interacts with the legal system as a victim or criminal defendant.

Further Resources

- A. Brunetti, G. Kisunko, and B. Weder, "[Credibility of Rules and Economic Growth: Evidence from a Worldwide Survey of the Private Sector](#)," The World Bank Economic Review, vol. 12, no. 3, pp. 353-384, Sep. 1998.
- S. Jasanoff, "Governing Innovation: The Social Contract and the Democratic Imagination," Seminar, vol. 597, pp. 16-25, May 2009.
- D. Kennedy, "The 'Rule of Law,' Political Choices and Development Common Sense," in *The New Law and Economic Development: A Critical Appraisal*, D. M. Trubek and A. Santos, eds., Cambridge: Cambridge University Press, 2006, pp. 95-173.
- "[Artificial Intelligence](#)," National Institute of Standards and Technology.
- K. Schwab, "[The Global Competitiveness Report: 2018](#)," The World Economic Forum, 2018.
- A. Sen, *Development as Freedom*. New York, NY: Alfred A. Knopf, 1999.
- United Nations General Assembly, [Universal Declaration of Human Rights](#), Dec. 10, 1948.
- UNICEF, [Convention on the Rights of the Child](#), Nov. 4, 2014.
- United Nations Office of the High Commissioner: Human Rights, [The Vienna Declaration and Programme of Action](#), June 25, 1993.
- World Bank, [World Development Report 2017: Governance and the Law](#), Jan. 2017.
- World Justice Project, [Rule of Law Index](#), June 2018.

Law

Issue 2: Impediments to Informed Trust

What are the challenges to adopting A/IS in legal systems and how can those impediments be overcome?

Background

Although the benefits to be gained by adopting A/IS in legal systems are potentially numerous (see the discussion of Issue 1), there are also significant risks that must be addressed in order for the A/IS to be adopted in a manner that will realize those benefits. The risks sometimes mirror expected benefits:

- the potential for opaque decision-making;
- the intentional or unintentional biases and abuses of power;
- the emergence of nontraditional bad actors;
- the perpetuation of inequality;
- the depletion of public trust in a legal system;
- the lack of human capital active in judicial systems to manage and operate A/IS;
- the sacrifice of the spirit of the law in order to achieve the expediency that the letter of the law allows;
- the unanticipated consequences of the surrender of human agency to nonethical agents;

- the loss of privacy and dignity;
- and the erosion of democratic institutions.³²

By way of example:

- Currently, A/IS used in justice systems are not subject to uniform rules and norms and are often adopted piecemeal at the local or regional level, thereby creating a highly variable landscape of tools and adoption practices. Critics argue that, far from improving fact-finding in civil and criminal matters or eliminating bias in law enforcement, these tools have unproven accuracy, are error-prone, and may serve to entrench existing social inequalities. These tools' potential must be weighed against their pitfalls. These include unclear efficacy; incompetent operation; and potential impairment of a legal system's ability to adhere to principles of socioeconomic, racial, or religious equality, government transparency, and individual due process, to render justice in an informed, consistent, and fair manner.
- In the case of *State v. Loomis*, an important but not widely known case, the Wisconsin Supreme Court held that a trial court's use of an algorithmic risk assessment tool in sentencing did not violate the defendant's due process rights, despite the fact that the methodology used to obtain the automated assessment was not disclosed to either the court or the defendant.³³ A man received a lengthy sentence based in part on what an opaque algorithm thought of him. While the court considered many factors, and sought to balance competing societal values, this

Law

is just one case in a growing set of cases illustrating how criminal justice systems are being impacted by proprietary claims of trade secrets, opaque operation of A/IS, a lack of evidence of the effectiveness of A/IS, and a lack of norms for the adoption of A/IS in the extended legal system.

- More generally, humans tend to be subject to the cognitive bias known as “anchoring”, which can be described as the excessive reliance on an initial piece of information. This may lead to the progressive, unwitting, and detrimental reliance of judges and legal practitioners on assessments produced by A/IS. This risk is compounded by the fact that A/IS are (and shall remain in the foreseeable future) nonethical agents, incapable of empathy, and thus at risk of being unable to produce decisions aligned with not just the letter of the law, but also the spirit of the law and reasonable regard for the circumstances of each defendant.
- The required technical and scientific knowledge to procure, deploy, and effectively operate A/IS, as well as that required to measure the ability of A/IS to achieve a given purpose without adverse collateral consequences, represent significant hurdles to the beneficial long-term adoption of A/IS in a legal system. This is especially the case when—as is the case presently—actors in the civil and criminal justice systems and in law enforcement may lack the requisite specialized technological or scientific expertise.³⁴

Such risks must be addressed in order to ensure sustainable management and public oversight of what will foreseeably become an increasingly automated justice system.³⁵ The view expressed by the Organisation for Economic Co-operation and Development (OECD) in the domain of digital security that “robust strategies to [manage risk] are essential to establish the trust needed for economic and social activities to fully benefit from digital innovation”³⁶ applies equally to the adoption of A/IS in the world’s legal systems.

Informed trust. If we are to realize the benefits of A/IS, we must trust that they are safe and effective. People board airplanes, take medicine, and allow their children on amusement park rides because they trust that the tools, methods, and people powering those technologies meet certain safety and effectiveness standards that reduce the risks to an acceptable level given the objectives and benefits to be achieved. This need for trust is especially important in the case of A/IS used in a legal system. The “black box” nature and lack of trust in A/IS deployed in the service of a legal system could quickly translate into a lack of trust in the legal system itself. This, in turn, may lead to an undermining of the social order. Therefore, if we are to improve the functioning of our legal systems through the adoption of A/IS, **we must enact policies and promote practices that allow those technologies to be adopted on the basis of informed trust.** Informed trust rests on a reasoned evaluation of clear and accurate information about the effectiveness of A/IS and the competence of their operators.³⁷

Law

To formulate policies and standards of practice intended to foster informed trust, it is helpful, first, to identify principles applicable over the entire supply chain for the delivery of A/IS-enabled decisions and guidance, including design, development, procurement, deployment, operation, and validation of effectiveness, that, if adhered to, will foster trust. Once those general principles have been identified, specific policies and standards of practice can be formulated that encourage adherence to the principles in every aspect of a legal system, including lawmaking, civil and criminal justice, and law enforcement. Such principles, if they are to serve their intended purpose of informing effective policies and practices, must meet certain design criteria. Specifically, **the principles should be (a) individually necessary and collectively sufficient, (b) globally applicable but culturally flexible, and (c) capable of being operationalized in applicable functions of the legal system.** A set of principles that meets these criteria will provide an effective framework for the development of policies and practices that foster trust, while leaving considerable flexibility in the specific policies and standards of practice that a society chooses to implement in furthering adherence to the principles.

A set of four principles that we believe meets the design criteria just described are the following:

- **Effectiveness:** Adoption of A/IS in a legal system should be based on sound empirical evidence that they are fit for their intended purpose.
- **Competence:** A/IS should be adopted in a legal system only if their creators specify

the skills and knowledge required for their effective operation and if their operators adhere to those competency requirements.

- **Accountability:** A/IS should be adopted in a legal system only if all those engaged in their design, development, procurement, deployment, operation, and validation of effectiveness maintain clear and transparent lines of responsibility for their outcomes and are open to inquiries as may be appropriate.
- **Transparency:** A/IS should be adopted in a legal system only if the stakeholders in the results of A/IS have access to pertinent and appropriate information about their design, development, procurement, deployment, operation, and validation of effectiveness.

In the remainder of Section 1, we elaborate on each of these principles. Before turning to a specific discussion of each, we add two further considerations that should be kept in mind when applying them collectively.

Differences in emphasis. While all four of the aforementioned principles will contribute to the fostering of trust, each principle will *not* contribute equally in every circumstance. For example, in many applications of A/IS, a well-established measure of effectiveness, obtained by proven and accepted methods, may go a considerable way to creating conditions for trust in the given application. In such a case, the other principles may add to trust, but they may not be necessary to establish trust. Or, to take another example, in some applications the role of the human operator may be minimal, while in other applications there will be extensive scope for

Law

human agency where competence has a greater role to play. In finding the right emphasis and balance among the four principles, policymakers and practitioners will have to consider the specific circumstances of A/IS.

Flexibility in implementation. It should be noted that we have addressed the four principles above at a rather high level and have not offered specific prescriptions of how adherence to the principles should be implemented. This is by design. Although adherence to all four principles is important, it is also important that, at the operational level, flexibility be allowed for the selection and implementation of policies and practices that (a) are in harmony with a given society's traditions, norms, and values; (b) conform with the laws and regulations operative in a given jurisdiction; and (c) are consistent with the ethical obligations of legal practitioners.

Recommendations

1. Governments should set procurement and contracting requirements that encourage parties seeking to use A/IS in the conduct of business with or for the government, particularly with or for the court system and law enforcement agencies, to adhere to the principles of effectiveness, competence, accountability, and transparency as described in this chapter. This can be achieved through legislation or administrative regulation. All government efforts in this regard should be transparent and open to public scrutiny.
2. Professionals engaged in the practice, interpretation, and enforcement of the

law (such as lawyers, judges, and law enforcement officers), when engaging with or relying on providers of A/IS technology or services, should require, at a minimum, that those providers adhere to, and be able to demonstrate adherence to, the principles of effectiveness, competence, accountability, and transparency as described in this chapter. Likewise, those professionals, when operating A/IS themselves, should adhere to, and be able to demonstrate adherence to, the principles of effectiveness, competence, accountability, and transparency. Demonstrations of adherence to the requirements should be publicly accessible.

3. Regulators should permit insurers to issue professional liability and other insurance policies that consider whether the insured (either a provider or operator of A/IS in a legal system) adheres to the principles of effectiveness, competence, accountability, and transparency (as they are articulated in this chapter).

Further Resources

- [“Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing—State v. Loomis, 881 N.W.2d 749 \(Wis. 2016\),”](#) Harvard Law Review, vol. 130, no. 5, pp. 1530-1537, 2017.
- K. Freeman, [“Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis,”](#) North Carolina Journal of Law and Technology, vol. 18, no. 5, pp. 75-76, 2016.

Law

- [“Managing Digital Security and Privacy Risk: Background Report for Ministerial Panel 3.2,”](#) Organisation for Economic Co-operation and Development (OECD) Directorate for Science, Technology, and Innovation: Committee on Digital Economy Policy, June 1, 2016.
- *State v Loomis*, 881 N.W.2d 749 (Wis. 2016), *cert. denied* (2017).
- [“Global Governance of AI Roundtable: Summary Report 2018,”](#) World Government Summit, 2018.

Issue 3: Effectiveness

How can the collection and disclosure of evidence of effectiveness of A/IS foster informed trust in the suitability for adoption in legal systems?

Background

An essential component of trust in a technology is trust that it works and meets the purpose for which it is intended. We now turn to a discussion of the role that evidence of effectiveness, chiefly in the form of the results of a measurement exercise, can play in fostering informed trust in A/IS as applied in legal systems.³⁸ We begin with a general characterization of what we mean by *evidence of effectiveness*: what we are measuring, how we are measuring, what form our results take, and who the intended

consumers of the evidence are. We then identify the specific features of the practice of measuring effectiveness that will enable it to contribute to informed trust in A/IS as applied in a legal system.

What constitutes evidence of effectiveness?

What we are measuring. In gathering evidence of effectiveness, we are seeking to gather empirical data that will tell us whether a given technology or its application will serve as an effective solution to the problem it is intended to address. Serving as an effective solution means more than meeting narrow specifications or requirements; it means that **the A/IS are capable of addressing their target problems in the real world**, which, in the case of A/IS applied in a legal system, are problems in the making, administration, adjudication, or enforcement of the law. It also means remaining practically feasible once collateral concerns and potential unintended consequences are taken into account.³⁹ To take a non-A/IS example, under the definition of effectiveness we are considering, for an herbicide to be considered effective, it must be shown not only to kill the target weeds, but also to do so without causing harm to nontarget plants, to the person applying the agent, and to the environment in general.

Under the definition above, assessing the effectiveness of A/IS in accomplishing the target task (narrowly defined) is not sufficient; it may also be necessary to assess the extent to which the A/IS are aligned with applicable

Law

laws, regulations, and standards,⁴⁰ and whether (and to what extent) they impinge on values such as privacy, fairness, or freedom from bias.⁴¹ Whether such collateral concerns are salient will depend on the nature of the A/IS and on the particular circumstances in which they are to be applied.⁴² However, it is only from such a complete view of the impact of A/IS that a balanced judgment can be made of the appropriateness of their adoption.⁴³

Although the scope of an evaluation of effectiveness is broader than a narrowly focused verification that a specific requirement is met, it has its limits. There are measures of aspects of A/IS that one might find useful but that are outside the scope of effectiveness. For example, given frequently expressed concerns that A/IS will one day cross the limits of their intended purpose and overwhelm their creators and users, one might seek to define and obtain general measures of the autonomy of a system or of a system's capacity for artificial general intelligence (AGI). Although such measures could be useful—assuming they could be defined—they are beyond the scope of evaluations of effectiveness. Effectiveness is always tied to a target purpose, even if it includes consideration of the collateral effects of the manner of meeting that purpose.

What we are measuring is therefore a general “fitness for purpose”.

How we measure. Evidence of effectiveness is typically gathered in one of two types of exercises:⁴⁴

- **A single-system validation exercise** measures and reports on the effectiveness of a single system on a given task. In such an exercise, the system to be validated will typically have already carried out the target task on a given data set. The purpose of the validation is to provide empirical evidence of how successful the system has been in carrying out the task on that data set. Measurements are obtained by independent sampling and review of the data to which the system was applied. Once obtained, those metrics serve to corroborate or refute the hypothesis that the system operated as intended in the instance under consideration. An example of validation as applied to legal fact-finding would be a test of the effectiveness of A/IS that had been used to retrieve material relevant (as defined by the humans deploying the system) to a given legal inquiry from a collection of emails.
- **A multi-system (or benchmarking) evaluation** involves conducting a comparative study of the effectiveness of several systems designed to meet the same objective. Typically, in such a study, a test data set is identified, a task to be performed is defined (ideally, a task that models the real-world objectives and conditions for which the systems under evaluation have been designed⁴⁵), the systems to be evaluated are used to carry out the task, and the success of each system in carrying out the task is measured and reported. An example of this sort of evaluation applied to a specific

Law

real-world challenge in the justice system is the series of evaluations of the effectiveness of information retrieval systems in civil discovery, including A/IS, conducted as part of the US National Institute of Standards and Technology (NIST) Text REtrieval Conference (TREC) Legal Track initiative.⁴⁶

The measurements obtained by both types of evaluation exercises are valuable. The results of a single-system validation exercise are typically more specific, answering the question of whether a system *was* effective in a specific instance. The results of a multi-system evaluation are typically more generic, answering the question of whether a system *can* be effective in real-world circumstances. Both questions are important, hence both types of evaluations are valuable.⁴⁷

The form of results. The results of an evaluation typically take the form of a number—a quantitative gauge of effectiveness. This can be, for example, the decreased likelihood of developing a given medical condition; safety ratings for automobiles; recall measures for retrieving responsive documents; and so on. Certainly, qualitative considerations are not (and should not) be ignored; they often provide context crucial to interpreting the quantitative results.⁴⁸ Nevertheless, at the heart of the results of an evaluation exercise is a number, a metric that serves as a telling indicator of effectiveness.⁴⁹

In some cases, the research community engaged in developing any new system will have reached consensus on salient effectiveness metrics. In other cases, the research community may not

have reached a consensus, requiring further study. In the case of A/IS, given both their accelerating development and the fact that they are often applied to tasks for which the effectiveness of their human counterparts is seldom precisely gauged, we are often still at the stage of defining metrics. An example of an application of A/IS for which there is a general consensus around measures of effectiveness is legal electronic discovery,⁵⁰ where there is a working consensus around the use of the evaluation metrics referred to as “recall” and “precision”.⁵¹ Conversely, in the case of A/IS applied in support of sentencing decisions, a consensus on the operative effectiveness metrics does not yet exist.⁵²

The consumers of the results. In defining metrics, it is important to keep in mind the consumers of the results of an evaluation of effectiveness. Broadly speaking, it is helpful to distinguish between two categories of stakeholders who will be interested in measurements of effectiveness:

- **Experts** are the researchers, designers, operators, and advanced users with appropriate scientific or professional credentials who have a technical understanding of the way in which a system works and are well-versed in evaluation methods and the results they generate.
- **Nonexperts** are the legislators, judges, lawyers, prosecutors, litigants, communities, victims, defendants, and system advocates whose work or legal outcomes may, even if only indirectly, be affected by the results

Law

of a given system. These individuals, however, may not have a technical understanding of the way in which a system operates. Furthermore, they may have little experience in conducting scientific evaluations and interpreting their results.

Effectiveness metrics must meet the needs of *both* expert *and* nonexpert consumers.

- With respect to experts, the purpose of an effectiveness metric is *to advance both long-term research and more immediate product development, maintenance, and oversight*. To achieve that purpose, it is appropriate to define a fine-grained metric that may not be within the grasp of the nonexpert. Researchers and developers will be acting on the information provided by such a metric, so it should be tailored to their needs.
- With respect to nonexperts, including the general public, the purpose of an effectiveness metric is *to advance informed trust*, meaning trust that is based on sound evidence that the A/IS have met, or will meet, their intended objectives, taking into account both the immediate purpose and the contextual purpose of preserving and fostering important values such as human rights, dignity, and well-being. For this purpose, it will be necessary to define a metric that can serve as a readily understood summary measure of effectiveness. This metric must provide a simple, direct answer to the question of how effective a given system is. Automobile safety ratings are an example of this sort of metric. For automobile designers and engineers, the summary

metrics are not sufficiently fine-grained to give immediately actionable information; for consumers, however, the metrics, insofar as they are accurate, empower them to make better-informed buying decisions.

For the purpose of fostering informed trust in A/IS adopted in the legal system, the most important goal is to establish a clear measure of effectiveness that can be understood by nonexperts. However, significant obstacles to achieving this goal include (a) developer incentives that prioritize research and development, along with the metrics that support such efforts, and (b) market forces that inhibit, or do not encourage, consumer-facing metrics. For those reasons, it is important that the selection and definition of the operative metrics draw on input not only from the A/IS creators but from other stakeholders as well; only under these conditions will a consensus form around the meaningfulness of the metrics.

What measurement practices foster informed trust?

By equipping both experts and nonexperts with accurate information regarding the capabilities and limitations of a given system, measurements of effectiveness can provide society with information needed to adopt and apply A/IS in a thoughtful, carefully considered, beneficial manner.⁵³

In order for the practice of measuring effectiveness to realize its full potential for fostering trust and mitigating the risks of uninformed adoption and uninformed avoidance of adoption, it must have certain features:

Law

- **Meaningful metrics:** As noted above, an essential element of a measurement practice is a metric that provides an accurate and readily understood gauge of effectiveness. The metric should provide clear and actionable information as to the extent to which a given application has, or has not, met its objective so that potential users of the results of the application can respond accordingly. For example, in legal discovery, both recall and precision have done this well and have contributed to the acceptance of the use of A/IS for this purpose.⁵⁴
- **Sound methods:** Measures of effectiveness must be obtained by scientifically sound methods. If, for example, measures are obtained by sampling, those sample-based estimates must be the result of sound statistical procedures that hold up to objective scrutiny.
- **Valid data:** Data on which evaluations of effectiveness are conducted should accurately represent the actual data to which the given A/IS would be applied and should be vetted for potential bias. Any data sets used for benchmarking or testing should be collected, maintained, and used in accordance with principles for the protection of individual privacy and agency.⁵⁵
- **Awareness and consensus:** Measurement practices must not only be technically sound in terms of metrics, methods, and data, but they must also be widely understood and accepted as evidence of effectiveness.
- **Implementation:** Measurement practices must be both practically feasible and actually implemented, i.e., widely adopted by practitioners⁵⁶.
- **Transparency.** Measurement methods and results must be open to scrutiny by experts and the general public.⁵⁷ Without such scrutiny, the measurements will not be trusted and will be incapable of fulfilling their intended purpose.⁵⁸

In seeking to advance informed trust in A/IS, policymakers should formulate policies and promote standards that encourage sound measurement practices, especially those that incorporate the key features.

Additional note. While in all circumstances all four principles discussed in this chapter (Effectiveness, Competence, Accountability, Transparency) will have something to contribute to the fostering of informed trust, it is not the case that in every circumstance all four principles will contribute equally to the fostering of trust. In some circumstances, a well-established measure of effectiveness, obtained by proven and accepted methods, may go a considerable way, on its own, in fostering trust in a given application—or distrust, if that is what the measurements indicate. In such circumstances, the challenges presented by the other principles, e.g., the challenge of adhering to the principle of transparency while respecting intellectual property considerations, may become of secondary importance.

Law

Illustration—Effectiveness

The search for factual evidence in large document collections in US civil or criminal proceedings has traditionally involved page-by-page manual review by attorneys. Starting in the 1990s, the proliferation of electronic data, such as email, rendered manual review prohibitively costly and time-consuming. By 2008, A/IS designed to substantially automate review of electronic data (a task known as “e-discovery”) were available. Yet, adoption remained limited. Chief among the obstacles to adoption was a concern about the effectiveness, and hence defensibility in court, of A/IS in e-discovery. **Simply put, practitioners and courts needed a sound answer to a simple question: “Does it work?”**

Starting in 2006, the US NIST⁵⁹ conducted studies to assess that question.⁶⁰ The studies focused on, among others, two sound statistical metrics, both expressed as easy-to-understand percentages:^{61,62}

- **Recall**, which is a gauge of the extent to which all the relevant documents were retrieved. For example, if there were 1,000 relevant documents to be found in the collection, and the review process identified 700 of them, then it achieved 70% recall.
- **Precision**, which is a gauge of the extent to which the documents identified as relevant by a process were actually relevant. For example, if for every two relevant documents the system captured, it also captured a nonrelevant one (i.e., a false positive), then it achieved 67% precision.

The studies provided empirical evidence that some systems could achieve high scores (80%) according to both metrics.⁶³ In a seminal follow-up study, Maura R. Grossman and Gordon V. Cormack found that two automated systems did, in fact, “conclusively” outperform human reviewers.⁶⁴ Drawing on the results of that study, Magistrate Judge Andrew Peck, in an opinion with far-reaching consequences, gave court approval for the use of A/IS to conduct legal discovery.⁶⁵

The story of the TREC Legal Track’s role in facilitating the adoption of A/IS for legal fact-finding contains a few lessons:

- **Metrics:** By focusing on recall and precision, the TREC studies quantified the effectiveness of the systems evaluated in a way that legal practitioners could readily understand.
- **Benchmarks:** The TREC studies filled an important gap: independent, scientifically sound evaluations of the effectiveness of A/IS applied to the real-world challenge of legal e-discovery.
- **Collaboration:** The founders of the TREC studies and the most successful participants came from both scientific and legal backgrounds, demonstrating the importance of multidisciplinary collaboration.

The TREC studies are a shining example of how the truth-seeking protocols of science can be used to advance the truth-seeking protocols of the law. They can serve as a conceptual basis for future benchmarking efforts, as well as the development of standards and certification programs to support informed trust when it comes to effectiveness of A/IS deployed in legal systems.⁶⁶

Law

Recommendations

1. Governments should fund and support the establishment of ongoing benchmarking exercises designed to provide valid, publicly accessible measurements of the effectiveness of A/IS deployed, or potentially deployed, in the legal system. That support could take a number of forms, ranging from direct sponsorship and oversight—for example, by nonregulatory measurement laboratories such as the US NIST—to indirect support by the recognition of the results of a credible third-party benchmarking exercise for the purposes of meeting procurement and contracting requirements. All government efforts in this regard should be transparent and open to public scrutiny.
2. Governments should facilitate the creation of data sets that can be used for purposes of evaluating the effectiveness of A/IS as applied in the legal system. In assisting in the creation of such data sets, governments and administrative agencies will have to take into consideration potentially competing societal values, such as the protection of personal data, and arrive at solutions that maintain those values while enabling the creation of usable, real-world data sets. All government efforts in this regard should be transparent and open to public scrutiny.
3. Creators of A/IS to be applied to legal matters should pursue valid measures of the effectiveness of their systems, whether through participation in benchmarking exercises or through conducting single-system validation exercises. Creators should describe the procedures and results of the testing in clear language that is understandable to both experts and nonexperts, and should do so without disclosing intellectual property. Further, the descriptions should be open to examination by all stakeholders, including, when appropriate, the general public.
4. Researchers engaged in the study and development of A/IS for use in the legal system should seek to define meaningful metrics that gauge the effectiveness of the systems they study. In selecting and defining metrics, researchers should seek input from all stakeholders in the outcome of the given application of A/IS in the legal system. The metrics should be readily understandable by experts and nonexperts alike.
5. Governments and industry associations should undertake educational efforts to inform both those engaged in the operation of A/IS deployed in the legal system and those affected by the results of their operation of the salient measures of effectiveness and what they can indicate about the capabilities and limitations of the A/IS in question.
6. Creators of A/IS for use in the legal system should ensure that the effectiveness metrics defined by the research community are readily obtainable and accessible to all stakeholders, including, when appropriate, the general public. Creators should provide guidance on how to interpret and respond to the metrics generated by the system.
7. Operators of A/IS applied to a legal task should follow the guidance on the measurement of effectiveness provided for

Law

the A/IS being used. This includes guidance about which metrics to obtain, how and when to obtain them, how to respond to given results, when it may be appropriate to follow alternative methods of gauging effectiveness, and so on.

8. In interpreting and responding to measurements of the effectiveness of A/IS applied to legal problems or questions, allowance should be made by those interpreting the results for variation in the specific objectives and circumstances of a given deployment of A/IS. Quantitative results should be supplemented by qualitative evaluation of the practical significance of a given outcome and whether it indicates a need for remediation. This evaluation should be done by an individual with the technical expertise and pragmatic experience needed to make a sound judgment.
 9. Industry associations or other organizations should collaborate on developing standards for measuring and reporting on the effectiveness of A/IS. These standards should be developed with input from both the scientific and legal communities.
 10. Recommendation 1 under Issue 2, with respect to effectiveness.
 11. Recommendation 2 under Issue 2, with respect to effectiveness.
- C. Garvie, A. M. Bedoya, and J. Frankle, "[The Perpetual Line-Up: Unregulated Police Face Recognition in America](#)," Georgetown Law, Center on Privacy & Technology, Oct. 2016.
 - M. R. Grossman and G. V. Cormack, "[Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review](#)," Richmond Journal of Law and Technology, vol. 17, no. 3, 2011.
 - B. Hedin, D. Brassil, and A. Jones, "On the Place of Measurement in E-Discovery," in Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner, J. R. Baron, R. C. Losey, and M. D. Berman, Eds. Chicago: American Bar Association, 2016.
 - J. A. Kroll, "[The fallacy of inscrutability](#)," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
 - D. W. Oard, J. R. Baron, B. Hedin, D. Lewis, and S. Tomlinson, "[Evaluation of Information Retrieval for E-Discovery](#)," Artificial Intelligence and Law, vol. 18, no. 4, pp. 347-386, Aug. 2010.
 - The Sedona Conference, "The Sedona Conference Commentary on Achieving Quality in the E-Discovery Process," The Sedona Conference Journal, vol. 15, pp. 265-304, 2014.
 - M. T. Stevenson, "[Assessing Risk Assessment in Action](#)," Minnesota Law Review, vol. 103, June 2018.

Further Resources

- *Da Silva Moore v. Publicis Groupe*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012).

Law

- [“Global Governance of AI Roundtable: Summary Report 2018,”](#) World Government Summit, 2018.
- High-Level Expert Group on Artificial Intelligence, “DRAFT Ethics Guidelines for Trustworthy AI: Working Document for Stakeholders’ Consultation,” The European Commission. Brussels, Belgium: Dec. 18, 2018.

Issue 4: Competence

How can specification of the knowledge and skills required of the human operator(s) of A/IS foster informed in the suitability of A/IS for adoption in legal systems?

Background

An essential component of informed trust in a technological system, especially one that may affect us in profound ways, is confidence in the competence of the operator(s) of the technology. We trust surgeons or pilots with our lives because we have confidence that they have the knowledge, skills, and experience to apply the tools and methods needed to carry out their tasks effectively. We have that confidence because we know that these operators have met rigorous professional and scientific accreditation standards before being allowed to step into the

operating room or cockpit. This informed trust in operator competence is what gives us confidence that surgery or air travel will result in the desired outcome. No such standards of operator competence currently exist with respect to A/IS applied in legal systems, where the life, liberty, and rights of citizens can be at stake. That absence of standards hinders the trustworthy adoption of A/IS in the legal domain.

The human operator is an integral component of A/IS

Almost all current applications of A/IS in legal systems, like those in most other fields, require human mediation and likely will continue to do so for the near future. This human mediation, post design and post development, will take a number of forms, including decisions about (a) whether or not to use A/IS for a given purpose,⁶⁷ (b) the data used to train the systems, (c) settings for system parameters to be used in generating results, (d) methods of validating results, (e) interpretation and application of the results, and so on. Because these systems’ outcomes are a function of all their components, including the human operator(s), their effectiveness, and by extension trustworthiness, will depend on their human operator(s).

Despite this, there are few standards that specify how humans should mediate applications of A/IS in legal systems, or what knowledge qualifies a person to apply A/IS and interpret their results.⁶⁸ This reality is especially troubling for the instances in which the life, rights, or liberty of humans are at stake. Today, while professional codes of ethics for lawyers are beginning to include among their

Law

requirements an awareness and understanding of technologies with legal application,⁶⁹ the operators of A/IS in legal systems are essentially deemed to be capable of determining their own competence: lawyers or IT professionals operating in civil discovery, correctional officers using risk assessment algorithms, and law enforcement agencies engaging in predictive policing or using automated surveillance technologies. All are mostly able to use A/IS without demonstrating that they understand the operation of the system they are using or that they have any particular set of consensus competencies.⁷⁰

The lack of competency requirements or standards undermines the establishment of informed trust in the use of A/IS in legal systems. If courts, legal practitioners, law enforcement agencies, and the general public are to rely on the results of A/IS when applied to tasks traditionally carried out by legal professionals, they must have grounds for believing that those operating A/IS will possess the requisite knowledge and skill to understand the conditions and methods for operating the systems effectively, including evaluating the data on which the A/IS trained, the data to which they are applied, the results they produce, and the methods and results of measuring the effectiveness the systems. Applied incompetently, A/IS could produce the opposite intended effect. Instead of improving a legal system—and bringing about the gains in well-being that follow from such improvements—they may undermine both the fairness and effectiveness of a legal system and trust in its fairness and effectiveness, creating conditions for social disorder and the deterioration of human

well-being that would follow from that disorder. By way of illustration:

- A city council might misallocate funds for policing across city neighborhoods because it relies on the output of an algorithm that directs attention to neighborhoods based on arrest rates rather than actual crime rates.⁷¹
- In civil justice, A/IS applied in a search of documents to uncover relevant facts may fail to do so because an operator without sufficient competence in statistics may materially overestimate the accuracy of the system, thus ceasing vital fact-finding activities.⁷²
- In the money bail system, reliance on A/IS to reduce bias may instead perpetuate it. For example, if a judge does not understand whether an algorithm makes sufficient contextual distinctions between gradations of offenses,⁷³ that judge would not be able to probe the output of the A/IS and make a well-informed use of it.
- In the criminal justice system, an operator using A/IS in sentencing decision-support may fail to identify bias, or to assess the risk of bias, in the results generated by the A/IS,⁷⁴ unfairly depriving a citizen of his or her liberty or prematurely granting an offender's release, increasing the risk of recidivism.

More generally, without the confidence that A/IS operators will apply the technology as intended and supervise it appropriately, the general public will harbor fear, uncertainty, and doubt about the use of A/IS in legal systems and potentially about the legal systems themselves.

Law

Fostering informed trust in the competence of human operators

If negative outcomes such as those just described are to be avoided, **it will be necessary to include among norms for the adoption of A/IS in a legal system a provision for building informed trust in the operators of A/IS.** Building trust will require articulating standards and best practices for two groups of agents involved in the deployment of A/IS: creators and operators.

On the one hand, those engaged in the design, development, and marketing of A/IS must commit to specifying the knowledge, skills, and conditions required for the safe, ethical, and effective deployment and operation of the systems.⁷⁵ On the other hand, those engaged in actually operating the systems, including both legal professionals and experts acting in the service of legal professionals, must commit to adhering to these requirements in a manner consistent with other operative legal, ethical, and professional requirements. The precise nature of the competency requirements will vary with the nature and purpose of the A/IS and what is at stake in their effective operation. The requirements for the operation of A/IS designed to assist in the creation of contracts, for example, might be less stringent than those for the operation of A/IS designed to assess flight risk, which could affect the liberty of individual citizens.

A corollary of these provisions is that education and training in the requisite skills should be available and accessible to those who would operate A/IS, whether that training is provided

through professional schools, such as law school; through institutions providing ongoing professional training, such as, for federal judges in the United States, the Federal Judicial Center; through professional and industry associations, such as the American Bar Association; or through resources accessible by the general public.⁷⁶ Making sure such training is available and accessible will be essential to ensuring that the resources needed for the competent operation of A/IS are widely and equitably distributed.⁷⁷

It will take a combined effort of both creators and operators to ensure both that A/IS designed for use in legal systems are properly applied and that those with a stake in the effective functioning of legal systems—including legal professionals, of course, but also decision subjects, victims of crime, communities, and the general public—will have informed trust, or, for that matter, informed distrust (if that is what a competence assessment finds) in the competence of the operators of A/IS as applied to legal problems and questions.⁷⁸

Illustration—Competence

Included among the offerings of Amazon Web Services is an image and video analysis service known as Amazon Rekognition.⁷⁹ The service is designed to enable the recognition of text, objects, people, and actions in images and videos. The technology also enables the search and comparison of faces, a feature with potential law enforcement and national security applications, such as comparing faces identified in video taken by a security camera with those in a database of jail booking photos. Attracted by

Law

the latter feature, police departments in Oregon and Florida have undertaken pilots of Rekognition as a tool in their law enforcement efforts.⁸⁰

In 2018, the American Civil Liberties Union (ACLU), a frequent critic of the use of facial recognition technologies by law enforcement agencies,⁸¹ conducted a test of Rekognition. The test consisted of first constructing a database of 25,000 booking photos (“mugshots”) then comparing publicly available photos of all then-current members of the US Congress against the images in the database. The test found that Rekognition incorrectly matched the faces of 28 members of Congress with faces of individuals who had been arrested for a crime.⁸² The ACLU argues that the high number of false positives generated by the technology shows that police use of facial recognition technologies generally (and of Rekognition in particular) poses a risk to the privacy and liberty of law-abiding citizens. The ACLU has used the results of its test of Rekognition to support its proposal that Congress enact a moratorium on the use of facial recognition technologies by law enforcement agencies until stronger safeguards against their misuse, and potential abuse, can be put in place.⁸³

In response to the ACLU report, Amazon noted that the ACLU researchers, in conducting their study, had applied the technology utilizing a similarity threshold (a gauge of the likelihood of a true match) of 80%, a threshold that casts a fairly wide net for potential matches (and hence generates a high number of false positives). For applications in which there are greater costs associated with false positives (e.g., policing),

Amazon recommends utilizing a similarity threshold value of 99% or above to reduce accidental misidentification.⁸⁴ Amazon also noted that, in all law enforcement use cases, it would be expected that the results of the technology would be reviewed by a human before any actual police action would be undertaken.

The story of the ACLU’s testing of Rekognition and Amazon’s response to the test highlights the importance of specifying and adhering to guidelines for competent use.⁸⁵ Had a law enforcement agency used the technology in the way it was used in the ACLU test, it would, in most legitimate use cases, be guilty of incompetent use. At the same time, Amazon is not free of blame insofar as it did not specify prominently and clearly the competency guidelines for effective use of the technology in support of law enforcement efforts, as well as the risks that might be incurred if those guidelines are not followed. Competent use⁸⁶ follows both from the A/IS creator’s specification of well-grounded⁸⁷ competency guidelines and from the A/IS operator’s adherence to those guidelines.⁸⁸

Recommendations

1. Creators of A/IS for application in legal systems should provide clear and accessible guidance for the knowledge, skills, and experience required of the human operators of the A/IS if the systems are to achieve expected levels of effectiveness. Included in that guidance should be a delineation of the risks involved if those requirements are not met. Such guidance should be

Law

documented in a form that is accessible and understandable by both experts and the general public.

2. Creators and developers of A/IS for application in legal systems should create written policies that govern how the A/IS should be operated. In creating these policies, creators and developers should draw on input from the legal professionals who will be using the A/IS they are creating. The policies should include:

- the specification of the real-world applications for the A/IS;
- the preconditions for their effective use;
- the training and skills that are required for operators of the systems;
- the procedures for gauging the effectiveness of the A/IS;
- the considerations to take into account in interpreting the results of the A/IS;
- the outcomes that can be expected by both operators and other affected parties when the A/IS are operated properly; and
- the specific risks that follow from improper use.

The policies should also specify circumstances in which it might be necessary for the operator to override the A/IS. All such policies should be publicly accessible.

3. Creators and developers of A/IS to be applied in legal systems should integrate safeguards against the incompetent operation of their systems. Safeguards could include issuing notifications and warnings to operators in

certain conditions, requiring, as appropriate, acknowledgment of receipt; limiting access to A/IS functionality based on the operator's level of expertise; enabling system shut-down in potentially high-risk conditions; and more. These safeguards should be flexible and governed by context-sensitive policies set by competent personnel of the entity (e.g., the judiciary), utilizing the A/IS to address a legal problem.

4. Governments should provide that any individual whose legal outcome is affected by the application of A/IS should be notified of the role played by A/IS in that outcome. Further, the affected party should have recourse to appeal to the judgment of a competent human being.
5. Professionals engaged in the creation, practice, interpretation, and enforcement of the law, such as lawyers, judges, and law enforcement officers, should recognize the specialized scientific and professional expertise required for the ethical and effective application of A/IS to their professional duties. The professional associations to which such legal practitioners belong, such as the American Bar Association, should, through both educational programs and professional codes of ethics, seek to ensure that their members are well informed about the scientific and technical competency requirements for the effective and trustworthy application of A/IS to the law.⁸⁹
6. The operators of A/IS applied in legal systems—whether the operator is a specialist in A/IS or a legal professional—should

Law

understand the competencies required for the effective performance of their roles and should either acquire those competencies or identify individuals with those competencies who can support them in the performance of their roles. The operator does not need to be an expert in all the pertinent domains but should have access to individuals with the requisite expertise.

7. Recommendation 1 under Issue 2, with respect to competence.
8. Recommendation 2 under Issue 2, with respect to competence.

Further Resources

- C. Garvie, A. M. Bedoya, and J. Frankle, "[The Perpetual Line-Up: Unregulated Police Face Recognition in America](#)," Georgetown Law, Center on Privacy & Technology, Oct. 2016.
- International Organization for Standardization, *ISO/IEC 27050-3: Information technology—Security techniques—Electronic discovery—Part 3: Code of practice for electronic discovery*, Geneva, 2017.
- J. A. Kroll, "[The fallacy of inscrutability](#)," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
- A. G. Ferguson, "[Policing Predictive Policing](#)," Washington University Law Review, vol. 94, no. 5 2017.
- "[Global Governance of AI Roundtable: Summary Report 2018](#)," World Government Summit, 2018.

Issue 5: Accountability

How can the ability to apportion responsibility for the outcome of the application of A/IS foster informed trust in the suitability of A/IS for adoption in legal systems?

Background

Apportioning responsibility. An essential component of informed trust in a technological system is confidence that it is possible, if the need arises, to apportion responsibility among the human agents engaged along the path of its creation and application: from design through to development, procurement, deployment,⁹⁰ operation, and, finally, validation of effectiveness. Unless there are mechanisms to hold the agents engaged in these steps accountable, it will be difficult or impossible to assess responsibility for the outcome of the system under any framework, whether a formal legal framework or a less formal normative framework. A model of A/IS creation and use that does not have such mechanisms will also lack important forms of deterrence against poorly thought-out design, casual adoption, and inappropriate use of A/IS.

Simply put, a system that produces outcomes for which no one is responsible cannot be trusted. Those engaged in creating, procuring, deploying, and operating such a system will lack the discipline engendered by the clear

Law

assignment of responsibility. Meanwhile, those affected by the results of the system's operation will find their questions around a given result inadequately answered, and errors generated by the system will go uncorrected. In the case of A/IS applied in a legal system, where an individual's basic human rights may be at issue, these questions and errors are of fundamental importance. In such circumstances, the only options are either blind trust or blind distrust. Neither of those options is satisfactory, especially in the case of a technological system applied in a domain as fundamental to the social order as the law.

Challenges to accountability

In the case of A/IS, whether applied in a legal system or another domain, maintaining accountability can be a particularly steep challenge. This challenge to accountability is because of both the perceived "black box" nature of A/IS and the diffusion of responsibility it brings.

The perception of A/IS as a black box stems from the opacity that is an inevitable characteristic of a system that is a complex nexus of algorithms, computer code, and input data. As observed by Joshua New and Daniel Castro of the Information Technology and Innovation Foundation:

The most common criticism of algorithmic decision-making is that it is a "black box" of extraordinarily complex underlying decision models involving millions of data points and thousands of lines of code. Moreover, the model can change over time, particularly when using

machine learning algorithms that adjust the model as the algorithm encounters new data.⁹¹

This opacity of the systems makes it challenging to trace cause to effect,⁹² which, in turn, makes it difficult or even impossible, to draw lines of responsibility.

The diffuseness challenge stems from the fact that even the most seemingly straightforward A/IS can be complex, with a wide range of agents—systems designers, engineers, data analysts, quality control specialists, operators, and others—involved in design, development, and deployment. Moreover, some of these agents may not even have been engaged in the development of the A/IS in question; they may have, for example, developed open-source components that were intended for an entirely different purpose but that were subsequently incorporated into the A/IS. This diffuseness of responsibility poses a challenge to the maintenance of accountability.⁹³ As Matthew Scherer, a frequent writer and speaker on topics at the intersection of law and A/IS, observes:

The sheer number of individuals and firms that may participate in the design, modification, and incorporation of an AI system's components will make it difficult to identify the most responsible party or parties. Some components may have been designed years before the AI project had even been conceived, and the components' designers may never have envisioned, much less intended, that their designs would be incorporated into any AI system, still less the specific AI system that caused harm. In such circumstances, it may seem unfair to assign

Law

blame to the designer of a component whose work was far-removed in both time and geographic location from the completion and operation of the AI system.⁹⁴

Examples include the following:

- When a judge's ruling includes a long prison sentence, based in part on a flawed A/IS-enabled process that erroneously deemed a particular person to be at high risk of recidivism, who is responsible for the error? Is it the A/IS designer, the person who chose the data or weighed the inputs, the prosecution team who developed and delivered the risk profile to the court, or the judge who did not have the competence to ask the appropriate questions that would have enabled a clearer understanding of the limitations of the system? Or is responsibility somehow distributed among these various agents?⁹⁵
- When a lawyer engaged in civil or criminal discovery believes, erroneously, that all the relevant information was found when using A/IS in a data-intensive matter, who is responsible for the failure to gather important facts? The A/IS designer who typically would have had no ability to foretell the specific circumstances of a given matter, the legal or IT professional who operated the A/IS or erroneously measured its effectiveness, or the lawyer who made a representation to his or her client, to the court, or to investigatory agencies?
- When a law enforcement officer, relying on A/IS, erroneously identifies an individual as being more likely to commit a crime than

another, who is responsible for the resulting encroachment on the civil rights of the person erroneously targeted? Is it the A/IS designer, the individual who selected the data on which to train the algorithm, the individual who chose how the effectiveness of the A/IS would be measured,⁹⁶ the experts who provided training to the officer, or the officer himself or herself?

As a result of the challenges presented by the opacity and diffuseness of responsibility in A/IS, the present-day answer to the question, "Who is accountable?" is, in far too many instances, "It's hard to say." This is a response that, in practice, means "no one" or, equally unhelpful, "everyone". Such failure to maintain accountability will undermine efforts to bring A/IS (and all their potential benefits) into legal systems based on informed trust.

Maintaining accountability and trust in A/IS

Although maintaining accountability in complex systems can be a challenge, it is one that must be met in order to engender informed trust in the use of A/IS in the legal domain. "Blaming the algorithm" is not a substitute for taking on the challenge of maintaining transparent lines of responsibility and establishing norms of accountability.⁹⁷ This is true even if we allow that, given the complexity of the systems in question, some number of "systems accidents" is inevitable.⁹⁸ Informed trust in a system does not require a belief that zero errors will occur; however, it does require a belief that there are mechanisms in place for addressing errors when

Law

they do occur. Accountability is an essential component of those mechanisms.

In meeting the challenge, it should be recognized that there are existing norms and controls that have a role to play in ensuring that accountability is maintained. For example, contractual arrangements between the A/IS provider and a party acquiring and applying a system may help to specify who is (and is not) to be held liable in the event the system produces undesirable results. Professional codes of ethics may also go some way toward specifying the extent to which lawyers, for example, are responsible for the results generated by the technologies they use, whether they operate them directly or retain someone else to do so. Judicial systems may have procedures for assessing responsibility when a citizen's rights are improperly infringed. As illustrated by the cases described above, however, existing norms and controls, while helpful, are insufficient in themselves to meet the specific challenge represented by the opacity and diffuseness of A/IS. To meet the challenge further steps must be taken.⁹⁹

The first step is ensuring that all those engaged in the creation, procurement, deployment, operation, and testing of A/IS recognize that, if accountability is not maintained, these systems will not be trusted. In the interest of maintaining accountability, these stakeholders should take steps to clarify lines of responsibility throughout this continuum, and make those lines of responsibility, when appropriate, accessible to meaningful inquiry and audit.

The goal of clarifying lines of responsibility in the operation of A/IS is to implement a governing model that specifies who is responsible for what, and who has recourse to which corrective actions, i.e., a trustworthy model that ensures that it will admit actionable answers should questions of accountability arise. Arriving at an effective model will require the participation of those engaged in the creation and operation of A/IS, those affected by the results of their use, and those with the expertise to understand how such a model would be used in a given legal system. For example:

- Individuals responsible for the design of A/IS will have to maintain a transparent record of the sources of the various components of their systems, including identification of which components were developed in-house and which were acquired from outside sources, whether open source or acquired from another firm.
- Individuals responsible for the design of A/IS will have to specify the roles, responsibilities, and potential subsequent liabilities of those who will be engaged in the operation of the systems they create.
- Individuals responsible for the operation of a system will have to understand their roles, responsibilities, potential liabilities, and will have to maintain documentation of their adherence to requirements.
- Individuals affected by the results of the operation of A/IS, e.g., a defendant in a criminal proceeding, will have to be given access to information about the roles and responsibilities of those involved in relevant

Law

aspects of the creation, operation, and validation of the effectiveness of the A/IS affecting them.¹⁰⁰

- Individuals with legal and political training (e.g., jurists, regulators, as well as legal and political scholars) will have to ensure that any model that is created will provide information that is in fact actionable within the operative legal system.

A governing model of accountability that reflects the interests of all these stakeholders will be more effective both at deterring irresponsible design or use of A/IS before it happens and at apportioning responsibility for an undesirable outcome when it does happen.¹⁰¹

Pulling together the input from the various stakeholders will likely not take place without some amount of institutional initiative. Organizations that employ A/IS for accomplishing legal tasks—private firms, regulatory agencies, law enforcement agencies, judicial institutions—should therefore develop and implement policies that will advance the goal of clarifying lines of responsibility. Such policies could take the form of, for example, designating an official specifically charged with oversight of the organization's procurement, deployment, and evaluation of A/IS as well as the organization's efforts to educate people both inside and outside the organization on its use of A/IS. Such policies might also include the establishment of a review board to assess the organization's use of A/IS and to ensure that lines of responsibility for the outcomes of its use are maintained. In the case of agencies, such as police departments, whose use of A/IS could impact the general public,

such review boards would, in the interest of legitimacy, have to include participation from various citizens' groups, such as those representing defendants in the criminal system as well as those representing victims of crime.¹⁰²

The goal of opening lines of responsibility to meaningful inquiry is to ensure that an investigation into the use of A/IS will be able to isolate responsibility for errors (or potential errors) generated by the systems and their operation.¹⁰³ This means that all those engaged in the design, development, procurement, deployment, operation, and validation of the effectiveness of A/IS, as well as the organizations that employ them, must in good faith be willing to participate in an audit, whether the audit is a formal legal investigation or a less formal inquiry. They must also be willing to create and preserve documentation of key procedures, decisions, certifications,¹⁰⁴ and tests made in the course of developing and deploying the A/IS.¹⁰⁵

The combination of a governing model of accountability and an openness to meaningful audit will allow the maintenance of accountability, even in complex deployments of A/IS in the service of a legal system.

Additional note 1. The principle of accountability is closely linked with each of the other principles intended to foster informed trust in A/IS: effectiveness, competence, and transparency. With respect to effectiveness, evidence of attaining key metrics and benchmarks to confirm that A/IS are functioning as intended may put questions of where, among creators,

Law

owners, and operators, responsibility for the outcome of a system lies on a sound empirical footing. With respect to competence, operator credentialing and specified system handoffs enable a clear chain of responsibility in the deployment of A/IS.¹⁰⁶ With respect to transparency, providing a view into the general design and methods of A/IS, or even a specific explanation for a given outcome, can help to advance accountability.

Additional note 2. Closely related to accountability is the trust that follows from knowing that a human expert is guiding the A/IS and is capable of overriding them, if necessary. Subjecting humans to automated decisions not only raises legal and ethical concerns, both from a data protection¹⁰⁷ and fundamental rights perspective,¹⁰⁸ but also will likely be viewed with distrust if the human component, which can introduce circumstantial flexibility in the interest of realizing an ethically superior outcome, is missing. In addition to ensuring technical safety and reliability of A/IS used in the course of decision-making processes, the legal system should also, where appropriate, provide for the possibility of an appeal for review by a human judge. Careful attention must be paid to the design of corresponding appeal procedures.¹⁰⁹

Illustration—Accountability

Over the last two decades, criminal justice agencies have increasingly embraced predictive tools to assist in the determination for bail, sentencing, and parole. A mix of companies, government agencies, nonprofits, and universities have built and promoted tools that provide a likelihood that someone may fail to appear

or may commit a new crime or a new violent act. While math has played a role in these determinations since at least the 1920s,¹¹⁰ a new interest in accountability and transparency has brought novel legal challenges to these tools.

In 2013, Eric Loomis was arrested for a drive-by shooting in La Crosse, Wisconsin. No one was hit, but Loomis faced prison time. Loomis denied involvement in the shooting, but waived his right to trial and entered a guilty plea to two of the less severe offenses with which he was charged: attempting to flee a traffic officer and operating a motor vehicle without the owner's consent. The judge sentenced him to six years in prison, saying he was "high risk". The judge based this conclusion, in part, on the risk assessment score given by Compas, a secret and privately held algorithmic tool used routinely by the Wisconsin Department of Corrections.

On appeal, Loomis made three major arguments, two focused on accountability.¹¹¹ First, the tool's proprietary nature—the underlying code was not made available to the defense—made it impossible to test its scientific validity. Second, the tool inappropriately considered gender in making its determination.

A unanimous Wisconsin Supreme Court ruled against Loomis on both arguments.

The court reasoned that knowing the inputs and output of the tool, and having access to validating studies of the tool's accuracy, were sufficient to prevent infringement of Loomis' due process.¹¹² Regarding the use of gender—a protected class in the United States—the court said he did not show that there was a reliance on gender in making the output or sentencing decision.

Law

Without the ability to interrogate the tool and know how gender is used, the court created a paradox with its opinion.

The *Loomis* decision represents the challenges that judges have balancing accountability of “black boxed” A/IS and trade secret protections.¹¹³ Other decisions have sided against accountability of other risk assessments,¹¹⁴ probabilistic DNA analysis tools,¹¹⁵ and government remote hacking investigation software.¹¹⁶ Siding with accountability, a federal judge found that the underlying code of a probability software used in DNA comparisons was admissible and relevant to a pretrial hearing where the admissibility of expert testimony is challenged.¹¹⁷

These issues will continue to be litigated as A/IS tools continue to proliferate in judicial systems. To that end, as the *Loomis* court notes, “The justice system must keep up with the research and continuously assess the use of these tools.”

Recommendations

1. Creators of A/IS to be applied in a legal system should articulate and document well-defined lines of responsibility, among all those who would be engaged in the development and operation of the A/IS, for the outcome of the A/IS.
2. Those engaged in the adoption and operation of A/IS to be applied in a legal system should understand their specific responsibilities for the outcome of the A/IS as well as their potential liability should the A/IS produce an outcome other than that intended. In the case of A/IS, many questions of legal liability remain unsettled. Adopters and operators of A/IS should nevertheless understand to what extent they could, *potentially*, be held liable for an undesirable outcome.
3. When negotiating contracts for the provision of A/IS products and services for use in the legal system, providers and buyers of A/IS should include contractual terms specifying clear lines of responsibility for the outcomes of the systems being acquired.
4. Creators and operators of A/IS applied in a legal system, and the organizations that employ them, should be amenable to internal oversight mechanisms and inquiries (or audits) that have the objective of allocating responsibility for the outcomes generated by the A/IS. In the case of A/IS adopted and deployed by organizations that have direct public interaction (e.g., a law enforcement agency), oversight and inquiry could also be conducted by external review boards. Being prepared for such inquiries means maintaining clear documentation of all salient procedures followed, decisions made, and tests conducted in the course of developing and applying the A/IS.
5. Organizations engaged in the development and operation of A/IS for legal tasks should consider mechanisms that will create individual and collective incentives for ensuring both that the outcomes of the A/IS adhere to ethical standards and that accountability for those outcomes is maintained, e.g., mechanisms to ensure that speed and efficiency are not rewarded at the expense of a loss of accountability.

Law

6. Those conducting inquiries to determine responsibility for the outcomes of A/IS applied in a legal system should take into consideration all human agents involved in the design, development, procurement, deployment, operation, and validation of effectiveness of the A/IS and should assign responsibility accordingly.
7. Recommendation 1 under Issue 2, with respect to accountability.
8. Recommendation 2 under Issue 2, with respect to accountability.

Further Resources

- N. Diakopoulos, S. Friedler, M. Arenas, S. Barocas, M. Hay, B. Howe, H. V. Jagadish, K. Unsworth, A. Sahuguet, S. Venkatasubramanian, C. Wilson, C. Yu, and B. Zevenbergen, "[Principles for Accountable Algorithms and a Social Impact Statement for Algorithms](#)," FAT/ML.
- F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger, and A. Wood, "[Accountability of AI Under the Law: The Role of Explanation](#)," Berkman Center Research Publication Forthcoming; Harvard Public Law Working Paper, no. 18-07, Nov. 3, 2017.
- European Commission for the Efficiency of Justice. *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment*. Strasbourg, 2018.
- J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "[Accountable Algorithms](#)," University of Pennsylvania Law Review, vol. 165, pp. 633-705. Feb. 2017.
- J. New and D. Castro, "[How Policymakers Can Foster Algorithmic Accountability](#)," Information Technology and Innovation Foundation, May 21, 2018.
- M. U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law & Technology*, vol. 29. no. 2, pp. 369-373, 2016.
- J. Tashea, "Calculating Crime: Attorneys are Challenging the Use of Algorithms to Help Determine Bail, Sentencing and Parole," *ABA Journal*, March 2017.

Issue 6: Transparency

How can sharing information that explains how A/IS reached given decisions or outcomes foster informed trust in the suitability of A/IS for adoption in legal systems?

Background

Access to meaningful information.

An essential component of informed trust in a technological system is confidence that the information required for a human to understand why the system behaves a certain way in a specific circumstance (or would behave in

Law

a hypothetical circumstance) will be accessible. Without transparency, there is no basis for trusting that a given decision or outcome of the system can be explained, replicated, or, if necessary, corrected.¹¹⁸ Without transparency, there is no basis for informed trust that the system can be operated in a way that achieves its ends reliably and consistently or that the system will not be used in a way that impinges on human rights. In the case of A/IS applied in a legal system, such a lack of trust could undermine the credibility of the legal system itself.

Transparency and trust

Transparency, by prioritizing access to information about the operation and effectiveness of A/IS, serves the purpose of fostering informed trust in the systems. More specifically, transparency fosters trust that:

- the operation of A/IS and the results they produce are explainable;
- the operation and results of A/IS are fair;¹¹⁹
- the operation and results of A/IS are unbiased;
- the A/IS meet normative standards for operation and results;
- the A/IS are effective;
- the results of A/IS are replicable;¹²⁰ and
- those engaged in the design, development, procurement, deployment, operation, and validation of the effectiveness of A/IS can be held accountable, where appropriate, for negative outcomes, and that corrective or punitive action can be taken when warranted.

For A/IS used in a legal system to achieve their intended purposes, all those with a stake in the effective functioning of the legal system must have a well-grounded trust that the A/IS can meet these requirements. This trust can be fostered by transparency.

The elements of transparency

Transparency of A/IS in legal matters requires disclosing information about the design and operation of the A/IS to various stakeholders. In implementing the principle, however, we must, in the interest of both feasibility and effectiveness, be more precise both about the categories of stakeholders to whom the information will be disclosed, and about the categories of information that will be disclosed to those stakeholders.

Relevant stakeholders in a legal system include those who:

- operate A/IS for the purpose of carrying out tasks in civil justice, criminal justice, and law enforcement, such as a law enforcement officer who uses facial recognition tools to identify potential suspects;
- rely on the results of A/IS to make important decisions, such as a judge who draws on the results of an algorithmic assessment of recidivism risk in deciding on a sentence;
- are directly affected by the use of A/IS—a “decision subject”, such as a defendant in a criminal proceeding whose bail terms are influenced by an algorithmic assessment of flight risk;

Law

- are indirectly affected by the results of A/IS, such as the members of a community that receives more or less police attention because of the results of predictive policing technology; and
- have an interest in the effective functioning of the legal system, such as judges, lawyers, and the general public.

Different types of relevant information can be grouped into high-level categories. As illustrated below, a taxonomy of such high-level categories may, for example, distinguish between:

- nontechnical procedural information regarding the employment and development of a given application of A/IS;
- information regarding data involved in the development, training, and operation of the system;
- information concerning a system's effectiveness/performance;
- information about the formal models that the system relies on; and
- information that serves to explain a system's general logic or specific outputs.

These more granular distinctions matter because different sorts of inquiries will require different sorts of information, and it is important to match the information provided to the actual needs of the inquiry. For example, an inquiry into a predictive policing system that misdirected police resources may not be much advanced by information about the formal models on which the system relied, but it may well be advanced by an explanation for the specific outcome.

On the other hand, an inquiry, undertaken by a designer or operator, into ways to improve system performance may benefit from access to information about the formal models on which the system relies.¹²¹

These distinctions also matter because there may be circumstances in which it would be desirable to limit access to a given type of information to certain stakeholders. For example, there may be circumstances in which one would want to identify an agent to serve as a public interest steward. For auditing purposes, this individual would have access to certain types of sensitive information unavailable to others. Such restrictions on information access are necessary if the transparency principle is not to impinge on other societal values and goals, such as security, privacy, and appropriate protection of intellectual property.¹²²

The salience of the question, "*Who is given access to what information?*" is illustrated by Sentiment Meter, a technology developed by Elucd, a GovTech company that provides cities with near real-time understanding of how citizens feel about their government, in conjunction with the New York Police Department, to assist the NYPD in gauging citizens' views regarding police activity in their communities.¹²³ One of the stated goals of the program is to build public trust in the police department. In the interest of trust, should "the public" have access to all potentially relevant information, including how the system was designed and developed, what the input data are, who operates the system and what their qualifications are, how the system's effectiveness was tested, and why the public was not brought

Law

into the process of construction? If the answer is that the general public should not have access to all this information, then who should? How do we define “the public?” Is it the whole community represented in its elected officials? Or should certain communities have greater access, for example, those most affected by controversial police practices such as stop, question, and frisk? Such questions must be answered if the program is to achieve its stated goals.

Transparency in practice

As just noted, although transparency can foster informed trust in A/IS applied in a legal system, **its practical implementation requires careful thought.** Requiring public access to all information pertaining to the operation and results of A/IS is neither necessary nor feasible. What is required is a careful consideration of who needs access to what information for the specific purpose of building informed trust. The following table is an example of a tool that might be used to match type of information to type of information consumer for the purpose of fostering trust.¹²⁴

Law

Types of information that should be considered in determining transparency demands in relation to a given A/IS		Stakeholders whose interest in access to different types of information should be considered in determining the transparency demands in relation to a given application of A/IS			
High-level category	Specific type of information (examples) Disclosure of...	Operators	Decision-subjects	Public interest steward	General public
Procedural aspects regarding A/IS employment and development	the fact that a given context involves the employment of A/IS	N/A	?	?	?
	how the employment of the system was authorized	?	?	?	?
	who developed the system	?	?	?	?
	...				
Data involved in A/IS development and operation	the origins of training data and data involved in the operation of the system	?	?	?	?
	the kinds of quality checks that data was subject to and their results	?	?	?	?
	how data labels are defined and to what extent data involves proxy variables	?	?	?	?
	relevant data sets themselves	?	?	?	?
	...				
Effectiveness/performance	the kinds of effectiveness/performance measurement that have occurred	?	?	?	?
	measurement results	?	?	?	?
	any independent auditing or certification	?	?	?	?
	...				
Model specification	the input variables involved	?	?	?	?
	the variable(s) that the model optimizes for	?	?	?	?
	the complete model (complete formal representation, source code, etc.)	?	?	?	?
	...				
Explanation	information concerning the system's general logic or functioning	?	?	?	?
	information concerning the determinants of a particular output ¹²⁵	?	?	?	?
	...				

Law

When it comes to deciding whether a specific type of information should be made available and, if so, which types of stakeholders should have access to it, there are various considerations, for example:

- The release of certain types of information may conflict with data privacy concerns, commercial or public policy interests—such as the promotion of innovation through appropriate intellectual property protections—and security interests, e.g., concerns about gaming and adversarial attacks. At the same time, such competing interests should not be permitted to be used, without specific justification, as a blanket cover for not adhering to due process, transparency, or accountability standards. The tension between these interests is particularly acute in the case of A/IS applied in a legal system, where the dignity, security, and liberty of individuals are at stake.¹²⁶
- There is tension between the specific goal of explainability, which may argue for limits on system complexity, and system performance, which may be served by greater complexity, to the detriment of explainability.¹²⁷
- One must carefully consider the question that is being asked in an inquiry into A/IS and what information transparency can actually produce to answer that question. Disclosure of A/IS algorithms or training data is, itself, insufficient to enable an auditor to determine whether the system was effective in a specific circumstance.¹²⁸ By analogy, transparency into drug manufacturing processes does not, itself, provide information about the

actual effectiveness of a drug. Clinical trials provide that insight. In a legal system, an excessive focus on transparency-related information-gathering and assessment may overwhelm courts, legal practitioners, and law enforcement agencies. Meanwhile, other factors, such as measurement of effectiveness or operator competence, coupled with information on training data, may often suffice to ensure that there is a well-informed basis for trusting A/IS in a given circumstance.¹²⁹

Given these competing considerations, arriving at a balance that is optimal for the functioning of a legal system and that has legitimacy in the eyes of the public will require an inclusive dialogue, bringing together the perspectives of those with an immediate stake in the proper functioning of a given technology, including those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of the technology, as well as those directly affected by the results of the technology; the perspectives of communities that may be indirectly impacted by the technology; and the perspectives of those with specialized expertise in ethics, government, and the law, such as jurists, regulators, and scholars. How the competing considerations should be balanced will also vary from one circumstance to another. Rather than aiming for universal transparency standards that would be applicable to all uses of A/IS within a legal system, transparency standards should allow for circumstance-dependent flexibility, in the context of the four constitutive components of trust discussed in this section.

Law

Additional note 1. The goals of transparency, e.g., answering a question as to why A/IS reached a given decision, may, in some cases, be better served by modes of explanation that do not involve examining an algorithm's terms or opening the "black box". A counterfactual explanation taking the form of, for example, "You were denied a loan because your annual income was £30,000; if your income had been £45,000, you would have been offered a loan," may provide more insight sooner than the disclosure of an algorithm.¹³⁰

Additional note 2. The transparency principle intersects with other principles focused on fostering trust. More specifically, we note the following:

- **Transparency and effectiveness.** Information about the measurement of effectiveness can foster trust only if it is disclosed, i.e., only if there is transparency pertaining to the procedures and results of a measurement exercise.
- **Transparency and competence.** Transparency is essential in ensuring that the competencies required by the human operators of A/IS are known and met. At the same time, questions addressed by transparency extend beyond competence, while the questions addressed by competence extend beyond those answered by transparency.
- **Transparency and accountability.** Transparency is essential in determining accountability, but transparency serves purposes beyond accountability, while accountability seeks to answer questions not addressed directly by transparency.

Illustration—Transparency

In 2004, the city of Memphis, Tennessee, was experiencing an increase in crime rates that exceeded the national average. In response, in 2005, the city piloted a predictive policing program known as Blue CRUSH (Crime Reduction Utilizing Statistical History).¹³¹

Blue CRUSH, developed in conjunction with the University of Memphis,¹³² utilizes IBM's SPSS predictive analytics software to identify "hot spots": locations and times in which a given type of crime has a greater than average likelihood of occurring. The system generates its results through the analysis of a range of both historical data (type of crime, location, time of day, day of week, characteristics of victim, etc.) and live data provided by units on patrol. Equipped with the predictive crime map generated by the system, the Memphis Police Department can allocate resources dynamically to preempt or interrupt the target criminal activity. The precise response the department takes will vary with circumstance: deployment of a visible patrol car, deployment of an unmarked observer car, increasing vehicle stops in the area, undercover infiltration of the location, and so on.

The pilot program of Blue CRUSH focused on gang-related gun violence, which had been on the rise in Memphis prior to the pilot. The program showed an improvement, relative to incumbent methods, in the interdiction of such violence. Based on the success of the pilot, the scope of program was expanded, in 2007, for use throughout the city. By 2013, the policing efforts enabled by Blue CRUSH had helped to reduce overall crime in the city by over 30% and violent crime by 20%.¹³³ The program

Law

also enabled a dramatic increase in the rate at which crimes were solved: for cases handled by the department's Felony Assault Unit, the percentage of cases solved increased from 16% to nearly 70%.¹³⁴ And the program was cost effective: an analysis by Nucleus Research found that the program, when compared to the resources required to achieve the same results by traditional means, realized an annual benefit of approximately \$7.2 million at a cost of just under \$400,000.¹³⁵

The story of the deployment of Blue CRUSH in the metropolitan Memphis area is not just about the technology; it is equally about the police personnel utilizing the technology and about the communities in which the technology was deployed. As noted by former Memphis Police Department Director Larry Godwin: "You can have all the technology in the world but you've got to have leadership, you've got to have accountability, you've got to have boots on the streets for it to succeed."¹³⁶ Crucial to the program's success was public support. Blue CRUSH represents a variety of predictive policing technology that limits itself to identifying the "where", the "when", and the "what" of criminal activity; it does not attempt to identify the "who", and therefore avoids a number of the privacy questions raised by technologies that do attempt to identify individual perpetrators. The technology will still, however, prompt responses by the police that could include more intrusive police activity in identified hot spots. The public must be willing to accept that activity, and that acceptance is won by transparency. To that end, Godwin and Janikowski held more than 200 community and neighborhood

watch meetings to inform the public about the technology and how it would be used in policing their communities.¹³⁷ Without that level of transparency, it is doubtful that Blue CRUSH would have had the public support needed for its successful deployment.

Holding community meetings is an important step in building trust in a predictive policing program. As such programs become more widely implemented, however, and become more widely studied, trust may require more than town-hall meetings. Research into the programs has raised serious concerns about the ways in which they are implemented and their potential for perpetuating or even exacerbating historical bias.¹³⁸ Addressing these concerns will require more sophisticated and intrusive oversight than can be realized through community meetings.

Included among the questions that must be addressed are the following.

- In identifying hot spots, does the program rely primarily on arrest rates, which reflect (potentially biased) police activity, or does it rely on actual crime rates?
- What are the specific criteria for identifying a hot spot and are those criteria free of bias?¹³⁹
- How accessible are the input data used to identify hot spots? Are they open to analysis by an independent expert?
- What mechanisms for oversight, review, and remediation of the program have been put in place? Such oversight should have access to the data used to train the system, the models used to identify hot spots, tests of the

Law

effectiveness of the system, and steps taken to remediate errors (such as bias) when they are uncovered.

As the public becomes more aware of the potential negative impact¹⁴⁰ of predictive policing programs, law enforcement agencies hoping to build trust in such programs will have to put in place transparency mechanisms that go beyond town-hall meetings and that enable a sophisticated response to such questions.

Recommendations

1. Governments and professional associations should facilitate dialogue among stakeholders—those engaged in the design, development, procurement, deployment, operation, and validation of effectiveness of the technology; those who may be immediately affected by the results of the technology; those who may be indirectly affected by the results of the technology, including the general public; and those with specialized expertise in ethics, politics, and the law—on the question of achieving a balance between transparency and other priorities, e.g., security, privacy, appropriate property rights, efficient and uniform response by the legal system, and more. In developing frameworks for achieving such balance, policymakers and professional associations should make allowance for circumstantial variation in how competing interests may be reconciled.
2. Policymakers developing frameworks for realizing transparency in A/IS applied to legal tasks should require that any frameworks they develop are sensitive both to the distinctions among the types of information that might be disclosed and to the distinctions among categories of individuals who may seek information about the design, operation, and results of a given system.
3. Policymakers developing frameworks for realizing transparency in A/IS to be adopted in a legal system should consider the role of appropriate protection for intellectual property, but should not allow those concerns to be used as a shield to prevent duly limited disclosure of information needed to ascertain whether A/IS meet acceptable standards of effectiveness, fairness, and safety. In developing such frameworks, policymakers should make allowance that the level of disclosure warranted will be, to some extent, dependent on what is at stake in a given circumstance.
4. Policymakers developing frameworks for realizing transparency in A/IS to be adopted in a legal system should consider the option of creating a role for a specially designated “public interest steward”, or “trusted third party”, who would be given access to sensitive information not accessible to others. Such a public interest steward would be charged with assessing the information to answer the public interest questions at hand but would be under obligation not to disclose the specifics of the information accessed in arriving at those answers.
5. Designers of A/IS should design their systems with a view to meeting transparency requirements, i.e., so as to enable some

Law

- categories of information about the system and its performance to be disclosed while enabling other categories, such as intellectual property, to be protected.
6. When negotiating contracts for the provision of A/IS products and services for use in the legal system, providers and buyers of A/IS should include contractual terms specifying what categories of information will be accessible to what categories of individuals who may seek information about the design, operation, and results of the A/IS.
 7. In developing frameworks for realizing transparency in A/IS to be adopted in a legal system, policymakers should recognize that the information provided by other types of inquiries, e.g., examination of evidence of effectiveness or of operator competence, may in certain circumstances provide a more efficient means to informed trust in the effectiveness, fairness, and safety of the A/IS in question.
 8. Governments should, where appropriate, work together with A/IS developers, as well as other stakeholders in the effective functioning of the legal system, to facilitate the creation of error-sharing mechanisms to enable the more effective identification, isolation, and correction of flaws in broadly deployed A/IS in their legal systems, such as a systematic facial recognition error in policing applications or in risk assessment algorithms. In developing such mechanisms, the question of precisely what information gets shared with precisely which groups may vary from application to application. All government efforts in this regard should be transparent and open to public scrutiny.
 9. Governments should provide whistleblower protections to individuals who volunteer to offer information in situations where A/IS are not designed as claimed or operated as intended, or when their results are not interpreted correctly. For example, if a law enforcement agency is using facial recognition technology for a purpose that is illegal or unethical, or in a manner other than that in which it is intended to be used, an individual reporting that misuse should be given protection against reprisal. All government efforts in this regard should be transparent and open to public scrutiny.
 10. Recommendation 1 under Issue 2, with respect to transparency.
 11. Recommendation 2 under Issue 2, with respect to transparency.

Further Resources

- J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "[Accountable Algorithms](#)," University of Pennsylvania Law Review, vol. 165, Feb. 2017.
- J. A. Kroll, "[The fallacy of inscrutability](#)," Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences, vol. 376, no. 2133, Oct. 2018.
- W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, "[Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations](#)," The RAND Corporation, 2013.
- A. D. Selbst and S. Barocas, "[The Intuitive Appeal of Explainable Machines](#)," Fordham Law Review, vol. 87, no. 3, 2018.

Law

- S. Wachter, B. Mittelstadt, and L. Floridi, "[Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation](#)," International Data Privacy Law, vol. 7, no. 2, pp. 76-99, June 2017.
- S. Wachter, B. Mittelstadt, and C. Russell, "[Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR](#)," Harvard Journal of Law & Technology, vol. 31, no. 2, 2018.
- R. Wexler, "[Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System](#)," Stanford Law Review, vol. 70, no. 5, pp. 1342-1429, 2017.

Law

Section 2: Legal Status of A/IS

There has been much discussion about how to legally regulate A/IS-related technologies and the appropriate legal treatment of systems that deploy these technologies. Already, some lawmakers are wrestling with the issue of what status to apply to A/IS. Legal “[personhood](#)”—applied to humans and certain types of human organizations—is one possible option for framing such legal treatment, but granting that status to A/IS applications raises issues in multiple domains of human interaction.

Issue

What type of legal status (or other legal analytical framework) is appropriate for A/IS given (i) the legal issues raised by deployment of such technologies, and (ii) the desire to maximize the benefits of A/IS and minimize negative externalities?

Background

The convergence of A/IS and robotics technologies has led to the development of systems and devices resembling those of human

beings in terms of their autonomy, ability to perform intellectual tasks, and, in the case of some robots, their physical appearance. As some types of A/IS begin to display characteristics resembling those of human actors, some governmental entities and private commentators have concluded that it is time to examine how legal regimes should categorize and treat various types of A/IS, often with an eye toward according A/IS a legal status beyond that of mere property. These entities have posited questions such as whether the law should treat such systems as legal persons.¹⁴¹

While legal personhood is a multifaceted concept, the essential feature of “full” legal personhood is the ability to participate autonomously within the legal system by having the right to sue and the capacity to be sued in court.¹⁴² This allows legal “persons” to enter legally binding agreements, take independent action to enforce their own rights, and be held responsible for violations of the rights of others.

Conferring such status on A/IS seems initially remarkable until consideration is given to the long-standing legal personhood status granted to corporations, governmental entities, and the like—none of which are themselves human. Unlike these familiar legal entities, however, A/IS are not composed of—or necessarily controlled by—human beings. Recognizing A/IS as independent legal entities could therefore lead to abuses of that status, possibly by A/IS

Law

and certainly by the humans and legal entities who create or operate them, just as human shareholders and agents have abused the corporate form.¹⁴³ A/IS personhood is a significant departure from the legal traditions of both common law and civil law.¹⁴⁴

Current legal frameworks provide a number of categories of legal status, other than full legal personhood, that could be used as analogues for the legal treatment of A/IS and how to allocate legal responsibility for harm caused by A/IS. At one extreme, legal systems could treat A/IS as mere products, tools, or other form of personal or intellectual property, and therefore subject to the applicable regimes of property law. Such treatment would have the benefit of simplifying allocation of responsibility for harm. It would, however, not account for the fact that A/IS, unlike other forms of property, may be capable of making legally significant decisions autonomously. In addition, if A/IS are to be treated as a form of property, governments and courts would have to establish rules regarding ownership, possession, and use by third parties. Other legal analogues may include the treatment of pets, livestock, wild animals, children, prisoners, and the legal principles of agency, guardianship, and powers of attorney.¹⁴⁵ Or perhaps A/IS are something entirely without precedent, raising the question of whether one or more types of A/IS might be assigned a hybrid, intermediate, or novel type of legal status?

Clarifying the legal status of A/IS in one or more jurisdictions is essential in removing the uncertainty associated with the obligations and expectations for organization and operation of

these systems. Clarification along these lines will encourage more certain development and deployment of A/IS and will help clarify lines of legal responsibility and liability when A/IS cause harm. One of the problems of exploiting the existing status of legal personhood is that international treaties may bind multiple countries to follow the lead of a single legislature, as in the EU, making it impossible for a single country to experiment with the legal and economic consequences of such a strategy.

Recognizing A/IS as independent legal persons would limit or eliminate some human responsibility for subsequent decisions made by such A/IS. For example, under a theory of [intervening causation](#), a hammer manufacturer is not held responsible when a burglar uses a hammer to break the window of a house. However, if similar “relief” from responsibility was available to the designers, developers, and users of A/IS, it will potentially reduce their incentives to ensure the safety of A/IS they design and use. In this example, legal issues that are applied in similar [chain of causation](#) settings—such as [foreseeability](#), [complicity](#), [reasonable care](#), [strict liability](#) for unreasonably dangerous goods, and other precedential notions—will factor into the design process. Different jurisdictions may reach different conclusions about the nature of such causation chains, inviting future creative legal planners to consider how and where to pursue design, development, and deployment of future A/IS in order to receive the most beneficial legal treatment.

The legal status of A/IS thus intertwines with broader legal questions regarding how to ensure

Law

accountability and assign and allocate liability when A/IS cause harm. The question of legal personhood for A/IS, in particular, also interacts with broader ethical and practical questions on the extent to which A/IS should be treated as moral agents independent from their human designers and operators, whether recognition of A/IS personhood would enhance or detract from the purposes for which humans created the A/IS in the first place, and whether A/IS personhood facilitates or debilitates the widespread benefits of A/IS.

Some assert that because A/IS are at a very early stage of development, it is premature to choose a particular legal status or presumption in the many forms and settings in which those systems are and will be deployed. However, thoughtfully establishing a legal status early in the development could also provide crucial guidance to researchers, programmers, and developers. This uncertainty about legal status, coupled with the fact that multiple legal jurisdictions are already deploying A/IS—and each of them, as a sovereign entity, can regulate A/IS as it sees fit—suggests that there are multiple general frameworks that can and should be considered when assessing the legal status of A/IS.

Recommendations

1. While conferring full legal personhood on A/IS might bring some economic benefits, the technology has not yet developed to the point where it would be legally or morally appropriate to generally accord A/IS the rights and responsibilities inherent in the legal definition of personhood as it is now defined.
2. In determining what legal status, including granting A/IS legal rights short of full legal personhood, to accord to A/IS, government and industry stakeholders alike should:
(1) identify the types of decisions and operations that should never be delegated to A/IS; and (2) determine what rules and standards will most effectively ensure human control over those decisions.
3. Governments and courts should review various potential legal models—including agency, animal law, and the other analogues discussed in this section—and assess whether they could serve as a proper basis for assigning and apportioning legal rights and responsibilities with respect to the deployment and use of A/IS.
4. In addition, governments should scrutinize existing laws—especially those governing business organizations—for mechanisms that could allow A/IS to have legal autonomy. If ambiguities or loopholes create a legal method for recognizing A/IS personhood, the government should review and, if appropriate, amend the pertinent laws.
5. Manufacturers and operators should learn how each jurisdiction would categorize a given autonomous and/or intelligent system and how each jurisdiction would treat harm caused by the system. Manufacturers and operators should be required to comply with the applicable laws of all jurisdictions in

Law

which that system could operate. In addition, manufacturers and operators should be aware of standards of performance and measurement promulgated by standards development organizations and agencies.

6. Stakeholders should be attentive to future developments that could warrant reconsideration of the legal status of A/IS. For example, if A/IS were developed that displayed self-awareness and consciousness, it may be appropriate to revisit the issue of whether they deserve a legal status on par with humans. Likewise, if legal systems underwent radical changes such that human rights and dignity no longer represented the primary guiding principle, the concept of full personhood for artificial entities may not represent the radical departure it might today. If the development of A/IS were to go in the opposite direction, and mechanisms were introduced allowing humans to control and predict the actions of A/IS easily and reliably, then the dangers of A/IS personhood would not be any greater than for well-established legal entities, such as corporations.
7. In considering whether to accord or expand legal protections, rights, and responsibilities to A/IS, governments should exercise utmost caution. Before according full legal personhood or a comparable legal status on A/IS, governments and courts should carefully consider whether doing so might limit how widely spread the benefits of A/IS are or could be, as well as whether doing so would harm human dignity and uniqueness of human identity. Governments and decision-makers at every level must work closely with

regulators, representatives of civil society, industry actors, and other stakeholders to ensure that the interest of humanity—and not the interests of the autonomous systems themselves—remains the guiding principle.

Further Resources

- S. Bayern. "[The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems.](#)" Stanford Technology Law Review 19, no. 1, pp. 93-112, 2015.
- S. Bayern, et al., "[Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators.](#)" Hastings Science and Technology Law Journal, vol. 9, no. 2, pp. 135-162, 2017.
- D. Bhattacharyya. "[Being, River: The Law, the Person and the Unthinkable.](#)" Humanities and Social Sciences Online, April 26, 2017.
- B. A. Garner, Black's Law Dictionary, 10th Edition, Thomas West, 2014.
- J. Bryson, et al., "Of, for, and by the people: the legal lacuna of synthetic persons," Artificial Intelligence Law 25, pp. 273-91, 2017.
- D. J. Calverley, "[Android Science and Animal Rights, Does an Analogy Exist?](#)" Connection Science 18, no. 4, pp. 403-417, 2006.
- D. J. Calverley, "[Imagining a Non-Biological Machine as a Legal Person.](#)" AI & Society 22, pp. 403-417, 2008.
- R. Chatila, "Inclusion of Humanoid Robots in Human Society: Ethical Issues," in Springer Humanoid Robotics: A Reference, A. Goswami and P. Vadakkepat, Eds., Springer 2018.

Law

- European Parliament [Resolution of 16 February 2017 \(2015/2103\(INL\)\)](#) with recommendations to the Commission on Civil Law Rules on Robotics, 2017.
- L. M. LoPucki, "[Algorithmic Entities](#)", 95 Washington University Law Review 887, 2018.
- J. S. Nelson, "Paper Dragon Thieves." Georgetown Law Journal 105, pp. 871-941, 2017.
- M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." Nevada Law Journal 19, forthcoming 2018.
- M. U. Scherer, "[Is Legal Personhood for AI Already Possible Under Current United States Laws?](#)" Law and AI, May 14, 2017.
- L. B. Solum. "[Legal Personhood for Artificial Intelligences](#)." North Carolina Law Review 70, no. 4, pp. 1231–1287, 1992.
- J. F. Weaver. [Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws](#). Santa Barbara, CA: Praeger, 2013.
- L. Zyga. "[Incident of drunk man kicking humanoid robot raises legal questions](#)," Techxplore, October 2, 2015.

Thanks to the Contributors

We wish to acknowledge all of the people who contributed to this chapter.

The Law Committee

- **John Casey** (Co-Chair) – Attorney-at-Law, Corporate, Wilson Sonsini Goodrich & Rosati, P.C.
- **Nicolas Economou** (Co-Chair) – Chief Executive Officer, H5; Chair, Science, Law and Society Initiative at The Future Society; Chair, Law Committee, Global Governance of AI Roundtable; Member, Council on Extended Intelligence
- **Aden Allen** – Senior Associate, Patent Litigation, Wilson Sonsini Goodrich & Rosati, P.C.
- **Miles Brundage** – Research Scientist (Policy), OpenAI; Research Associate, Future of Humanity Institute, University of Oxford; PhD candidate, Human and Social Dimensions of Science and Technology, Arizona State University
- **Thomas Burri** – Assistant Professor of International Law and European Law, University of St. Gallen (HSG), Switzerland
- **Ryan Calo** – Assistant Professor of Law, the School of Law at the University of Washington
- **Clemens Canel** – Referendar (Trainee Lawyer) at Hanseatisches Oberlandesgericht, graduate of the University of Texas School of Law and Bucerius Law School
- **Chandramauli Chaudhuri** – Senior Data Scientist; Fractal Analytics
- **Danielle Keats Citron** – Lois K. Macht Research Professor & Professor of Law, University of Maryland Carey School of Law
- **Fernando Delgado** – PhD Student, Information Science, Cornell University.
- **Deven Desai** – Associate Professor of Law and Ethics, Georgia Institute of Technology, Scheller College of Business
- **Julien Durand** – International Technology Lawyer; Executive Director Compliance & Ethics, Amgen Biotechnology
- **Todd Elmer, JD** – Member of the Board of Directors, National Science and Technology Medals Foundation
- **Kay Firth-Butterfield** – Project Head, AI and Machine Learning at the World Economic Forum. Founding Advocate of AI-Global; Senior Fellow and Distinguished Scholar, Robert S. Strauss Center for International Security and Law, University of Texas, Austin; Co-Founder, Consortium for Law and Ethics of Artificial Intelligence and Robotics, University of Texas, Austin; Partner, Cognitive Finance Group, London, U.K.
- **Tom D. Grant** – Fellow, Wolfson College; Senior Associate of the Lauterpacht Centre for International Law, University of Cambridge, U.K.

Law

- **Cordel Green** – Attorney-at-Law; Executive Director, Broadcasting Commission—Jamaica
- **Maura R. Grossman** – Research Professor, David R. Cheriton School of Computer Science, University of Waterloo; Adjunct Professor, Osgoode Hall Law School, York University
- **Bruce Hedin** – Principal Scientist, H5
- **Daniel Hinkle** – Senior State Affairs Counsel for the American Association for Justice
- **Derek Jinks** – Marrs McLean Professor in Law, University of Texas Law School; Director, Consortium on Law and Ethics of Artificial Intelligence and Robotics (CLEAR), Robert S. Strauss Center for International Security and Law, University of Texas.
- **Nicolas Jupillat** – Adjunct Professor, University of Detroit Mercy School of Law
- **Marwan Kawadri** – Analyst, Founders Intelligence; Research Associate, The Future Society.
- **Mauricio K. Kimura** – Lawyer; PhD student, Faculty of Law, University of Waikato, New Zealand; LLM from George Washington University, Washington DC, USA; Bachelor of Laws from Sao Bernardo do Campo School of Law, Brazil
- **Irene Kitsara** – Lawyer; IP Information Officer, Access to Information and Knowledge Division, World Intellectual Property Organization, Switzerland
- **Timothy Lau, J.D., Sc.D.** – Research Associate, Federal Judicial Center
- **Mark Lyon** – Attorney-at-Law, Chair, Artificial Intelligence and Autonomous Systems Practice Group at Gibson, Dunn & Crutcher LLP
- **Gary Marchant** – Regents' Professor of Law, Lincoln Professor of Emerging Technologies, Law and Ethics, Arizona State University
- **Nicolas Mialhe** – Co-Founder & President, The Future Society; Member, AI Expert Group at the OECD; Member, Global Council on Extended Intelligence; Senior Visiting Research Fellow, Program on Science Technology and Society at Harvard Kennedy School. Lecturer, Paris School of International Affairs (Sciences Po); Visiting Professor, IE School of Global and Public Affairs
- **Paul Moseley** – Master's student, Electrical Engineering, Southern Methodist University; graduate of the University of Texas School of Law
- **Florian Ostmann** – Policy Fellow, The Alan Turing Institute
- **Pedro Pavón** – Assistant General Counsel, Global Data Protection, Honeywell
- **Josephine Png** – AI Policy Researcher and Deputy Project Manager, The Future Society; budding barrister; and BA Chinese and Law, School of Oriental and African Studies
- **Matthew Scherer** – Attorney at Littler Mendelson, P.C., and legal scholar based in Portland, Oregon, USA; Editor, LawAndAI.com
- **Bardo Schettini Gherardini** – Independent Legal Advisor on standardization, AI and robotics

Law

- **Jason Tashea** – Founder, Justice Codes and adjunct law professor at Georgetown Law Center
- **Yan Tougas** – Global Ethics & Compliance Officer, United Technologies Corporation; Adjunct Professor, Law & Ethics, University of Connecticut School of Business; Fellow, Ethics & Compliance Initiative; Kallman Executive Fellow, Bentley University Hoffman Center for Business Ethics
- **Sandra Wachter** – Lawyer and Research Fellow in Data Ethics, AI and Robotics, Oxford Internet Institute, University of Oxford
- **Axel Walz** – Lawyer; Senior Research Fellow at the Max Planck Institute for Innovation and Competition, Germany. (Member until October 31, 2018)
- **John Frank Weaver** – Lawyer, McLane Middleton, P.A.; Columnist for and Member of Board of Editors of *Journal of Robotics, Artificial Intelligence & Law*; Contributing Writer for *Slate*; Author, *Robots Are People Too*
- **Julius Weitzdörfer** – Affiliated Lecturer, Faculty of Law, University of Cambridge; Research Associate, Centre for the Study of Existential Risk, University of Cambridge
- **Yueh-Hsuan Weng** – Assistant Professor, Frontier Research Institute for Interdisciplinary Sciences (FRIS), Tohoku University; Fellow, Transatlantic Technology Law Forum (TTLF), Stanford Law School
- **Andrew Woods** – Associate Professor of Law, University of Arizona

For a full listing of all IEEE Global Initiative Members, visit standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf.

For information on disclaimers associated with EAD1e, see [How the Document Was Prepared](#).

Law

The Law Committee of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems would like to thank the following individuals for taking the time to offer valuable feedback and suggestions on Section 1 of the Law Chapter, “Norms for the Trustworthy Adoption of A/IS in Legal Systems”. Each of these contributors offered comments in an individual capacity, not in the name of the organization for which they work. The final version of the Section does not necessarily incorporate all comments or reflect the views of each contributor.

- **Rediet Abebe**, PhD Candidate, Department of Computer Science, Cornell University; cofounder, Mechanism Design for Social Good; cofounder, Black in AI.
- **Ifeoma Ajunwa**, Assistant Professor, Labor & Employment Law, Cornell Industrial and Labor Relations School; faculty Associate at Harvard Law, Berkman Klein Center.
- **Jason R. Baron**, of counsel, Drinker Biddle; co-chair, Information Governance Initiative; former Director of Litigation, United States National Archives and Records Administration.
- **Irakli Beridze**, Head, Centre for Artificial Intelligence and Robotics, United Nations (UNICRI).
- **Juan Carlos Botero**, Law Professor, Pontificia Universidad Javeriana, Bogota; former Executive Director, World Justice Project.
- **Anne Carblanc**, Principal Administrator, Information, Communications and Consumer Policy (ICCP) Division, Directorate for Science, Technology and Industry, OECD; former criminal investigations judge (juge d’instruction), Tribunal of Paris.
- **Gallia Daor**, Policy Analyst, OECD.
- **Lydia de la Torre**, Privacy Law Fellow, Santa Clara University.
- **Isabela Ferrari**, Federal Judge, Federal Court, Rio de Janeiro, Brazil.
- **Albert Fox Cahn**, Founder and Executive Director, Surveillance Technology Oversight Project; former Legal Director, CAIR-NY.
- **Paul W. Grimm**, United States District Judge, United States District Court for the District of Maryland.
- **Gillian Hadfield**, Professor of Law and Professor of Strategic Management, University of Toronto; Member, World Economic Forum Future Council for Agile Governance.
- **Sheila Jasanoff**, Pforzheimer Professor of Science and Technology Studies, Harvard Kennedy School of Government.
- **Baroness Beeban Kidron**, OBE, Member, United Kingdom House of Lords.
- **Eva Kaili**, Member, European Parliament; Chair, European Parliament Science and Technology Options Assessment body (STOA).
- **Mantelena Kaili**, cofounder, European Law Observatory on New Technologies.
- **Jon Kleinberg**, Tisch University Professor, Departments of Computer Science and Information Science, Cornell University; member of the National Academy of Sciences, the National Academy of Engineering, and the American Academy of Arts and Sciences.
- **Shuang Lu Frost**, Teaching Fellow, PhD candidate, Department of Anthropology, Harvard University.

Law

- **Arthur R. Miller CBE**, University Professor, New York University; former Bruce Bromley Professor of Law, Harvard Law School.
- **Manuel Muñiz**, Dean and Rafael del Pino Professor of Practice of Global Leadership, IE School of Global and Public Affairs, Madrid; Senior Associate, Belfer Center, Harvard University.
- **Erik Navarro Wolkart**, Federal Judge, Federal Court, Rio de Janeiro, Brazil.
- **Aileen Nielsen**, chair, Science and Law Committee, New York City Bar Association.
- **Michael Philips**, Assistant General Counsel, Microsoft.
- **Dinah PoKempner**, General Counsel, Human Rights Watch.
- **Irina Raicu**, Director, Internet Ethics Program, Markkula Center for Applied Ethics, Santa Clara University.
- **David Robinson**, Visiting Scientist, AI Policy and Practice Initiative, Cornell University; Adjunct Professor of Law, Georgetown University Law Center; Managing Director (on leave), Upturn.
- **Alanna Rutherford**, Vice President, Global Litigation & Competition, Visa.
- **George Socha, Esq.**, Consulting Managing Director, BDO USA; co-founder, Electronic Discovery Reference Model (EDRM) and Information Governance Reference Model (IGRM).
- **Lee Tiedrich**, Partner, IP/Technology Transactions, and Co-Chair, Artificial Intelligence Initiative, Covington & Burling LLP.
- **Darrell M. West**, VP, Governance Studies, Director, Center for Technology Innovation, Douglas Dillon Chair in Governance Studies, Brookings Institution.
- **Bendert Zevenbergen**, Research Fellow, Center for Information Technology Policy, Princeton University; Researcher, Oxford Internet Institute.
- **Jiyu Zhang**, Associate Professor and Executive Director of the Law and Technology Institute, Renmin University of China School of Law.
- **Peter Zimroth**, Director, New York University Center on Civil Justice; retired partner, Arnold & Porter; former Assistant US Attorney, Southern District of New York.

Endnotes

¹ See S. Jasanoff, "Governing Innovation: The Social Contract and the Democratic Imagination," Seminar, vol. 597, pp. 16-25, May 2009.

² As articulated in *EAD* General Principles 1 (Human Rights), 2 (Well-Being), and 3 (Data Agency). See also *EAD* Chapter, "Classical Ethics in A/IS," In applying A/IS in pursuit of these goals, tradeoffs are inevitable. Some applications of predictive policing, for example, may reduce crime, and so enhance well-being, but may do so at the cost of impinging on a right to privacy or weakening protections against unwarranted search and seizure. How these tradeoffs are negotiated may vary with cultural and legal traditions.

³ Risks and benefits, and their perception, are neither always well-defined at the outset nor static over time. Social expectations and even ideas of lawfulness constantly evolve. For example, if younger generations, accustomed to the use of social networking technologies, have lower expectations of privacy than older generations, should this be deemed to be a benefit to society, a risk, or neither?

⁴ Regarding the nature of the guidance provided in this section: Artificial intelligence, like many other domains relied on by the legal realm (e.g., medical and accounting forensics, ballistics, or economic analysis), is a scientific discipline distinct from the law. Its effective and safe design and operation have underpinnings in academic

and professional competencies in computer science, linguistics, data science, statistics, and related technical fields. Lawyers, judges, and law enforcement officers increasingly draw on these fields, directly or indirectly, as A/IS are progressively adopted in the legal system. This document does not seek to offer legal advice to lawyers, courts, or law enforcement agencies on how to practice their professions or enforce the law in their jurisdictions around the globe. Instead, it seeks to help ensure that A/IS and their operators in a given legal system can be trusted by lawyers, courts, and law enforcement agencies, and civil society at large, to perform effectively and safely. Such effective and safe operation of A/IS holds the potential of producing substantial benefits for the legal system, while protecting all of its participants from the ethical, professional, and business risks, or personal jeopardy, that may result from the intentional, unintentional, uninformed, or incompetent procurement and operation of artificial intelligence.

⁵ See Rensselaer Polytechnic Institute, "A Conversation with Chief Justice John G. Roberts, Jr.," April 11, 2017. YouTube video, 40:12. April 12, 2017. [Online]. Available: <https://www.youtube.com/watch?v=TuZEKlRgDEg>.

⁶ "Uninformed avoidance of adoption" can be one of two types: (a) avoidance of adoption when the information needed to enable sound decisions is available but is not taken into

Law

consideration, and (b) avoidance of adoption when the information needed to enable sound decisions is simply not available. Unlike the former type of avoidance, the latter type is a prudent and well-reasoned avoidance of adoption and, pending better information, is the course recommended by a number of experts and nonexperts.

⁷ For purposes of this chapter, we have made the deliberate choice to focus on these four principles without taking a prior position on where the deployment of A/IS may or may not be acceptable in legal systems. Where these principles cannot be adequately operationalized, it would follow that the deployment of A/IS in a legal system cannot be trusted. Where A/IS can be evidenced to meet desired thresholds for each duly operationalized principle, it would follow that their deployment can be trusted. Such information is intended to facilitate, not preempt, the indispensable public policy dialogue on the extent to which A/IS should be relied upon to meet the specific needs of the legal systems of societies around the world.

⁸ It is beyond the scope of this chapter to discuss the process through which such adherence may become institutionalized in the complex legal, technological, political, and cultural dynamics in which sociotechnical innovation occurs. It is worth noting, however, that this process typically involves four steps. First, a wide range of market and culture-driven practices emerge. Second, a set of best practices arises, reflecting a group's willingness to adopt certain rules. Third, some of these best practices are formulated into standards, which

enable enforcement (through private contracts, professional codes of practice, or legislation). Finally, those enforceable standards render the performance of some activities sufficiently reliable and predictable to enable trustworthy operation at the scale of society. Where these elements (rulemaking, enforcement, scalable operation) are present, new institutions are born.

⁹ For a discussion of the definition of A/IS, see the Terminology Update in the Executive Summary of EAD. The principles outlined in this section as constitutive of "informed trust" do not depend on a precise, consensus definition of A/IS and are, in fact, designed to enable successful operationalization under a broad range of definitions.

¹⁰ Such as Gross Domestic Product (GDP), Gross National Income (GNI) per capita, the WEF Global Competitiveness Index, and others.

¹¹ Such as life expectancy, infant mortality rate, and literacy rate, as well as composite indices such as the Human Development Index, the Inequality-Adjusted Human Development Index, the OECD Framework for Measuring Well-being and Progress, and others. For more on measures of well-being, see the *EAD* chapter on "Well-being".

¹² See United Nations General Assembly, Universal Declaration of Human Rights, Dec. 10, 1948, available: <http://www.un.org/en/universal-declaration-human-rights/index.html>; see also United Nations Office of the High Commissioner: Human Rights, The Vienna Declaration and Programme of Action, June 25, 1993, available: <https://www.ohchr.org/en/professionalinterest/pages/vienna.aspx>.

Law

¹³ See UNICEF, Convention on the Rights of the Child, Nov. 4, 2014, available: https://www.unicef.org/crc/index_30160.html.

¹⁴ See United Nations Security Council, “The Rule of Law and Transitional Justice in Conflict and Post-conflict Societies: Report of the Secretary General,” *Report S/2004/616* (2004).

¹⁵ See The World Economic Forum, *The Global Competitiveness Report: 2018*, ed. K. Schwab (2018), pp. 12ff.

¹⁶ See A. Brunetti, G. Kisunko, and B. Weder, “Credibility of Rules and Economic Growth: Evidence from a Worldwide Survey of the Private Sector,” *The World Bank Economic Review*, vol. 12, no. 3, pp. 353–384, 1998. Available: <https://doi.org/10.1093/wber/12.3.353>; see also World Bank, *World Development Report 2017: Governance and the Law*, Jan. 2017. Available: doi.org/10.1596/978-1-4648-0950-7.

¹⁷ The question of intellectual property law in an era of rapidly advancing technology (both A/IS and other technologies) is a complex and often contentious one involving legal, economic, and ethical considerations. We have not yet studied the question in sufficient depth to reach a consensus on the issues raised. We may examine the issues in depth in a future version of *EAD*. For a forum in which such issues are discussed, see the Berkeley-Stanford Advanced Patent Law Institute. See also The World Economic Forum, “Artificial Intelligence Collides with Patent Law.” April 2018. Available: http://www3.weforum.org/docs/WEF_48540_WP_End_of_Innovation_Protecting_Patent_Law.pdf.

¹⁸ A component of human dignity is privacy, and a component of privacy is protection and control of one’s data; in this regard, frameworks such as the EU’s General Data Protection Regulation (GDPR) and the Council of Europe’s “Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data” have a role to play in setting standards for how legal systems can protect data privacy. See also *EAD* General Principle 3 (Data Agency).

¹⁹ Frameworks such as the Universal Declaration of Human Rights and the Vienna Declaration and Programme of Action (VDPA) have a role to play in articulating human-rights standards to which legal systems should adhere. See also *EAD* General Principle 1 (Human Rights).

²⁰ For more on the importance of measures of well-being beyond GDP, see *EAD* General Principle 2 (Well-being).

²¹ For a conceptual framework enabling the country-by-country assessment of the Rule of Law, see World Justice Project, *Rule of Law Index*. 2018. url: https://worldjusticeproject.org/sites/default/files/documents/WJP-ROLI-2018-June-Online-Edition_0.pdf.

²² See D. Kennedy, “The ‘Rule of Law,’ Political Choices and Development Common Sense,” in *The New Law and Economic Development: A Critical Appraisal*, D. M. Trubek and A. Santos, Ed. Cambridge: Cambridge University Press, 2006, pp. 156-157; see also A. Sen, *Development as Freedom*. New York: Alfred A. Knopf, 1999.

Law

²³ See Kennedy (2006): pp. 168-169. “The idea that building ‘the rule of law’ might *itself* be a development strategy encourages the hope that choosing law *in general* could substitute for all the perplexing political and economic choices that have been at the center of development policy making for half a century. The politics of allocation is submerged. Although a legal regime offers an arena to contest those choices, it cannot substitute for them.”

²⁴ *Fairness* (as well as *bias*) can be defined in more than one way. For purposes of this chapter, a commitment is not made to any one definition—and indeed, it may not be either desirable or feasible to arrive at a single definition that would be applied in all circumstances. The trust principles proposed in the chapter (Effectiveness, Competence, Accountability, and Transparency) are defined such that they will provide information that will allow the testing of an application of A/IS against any fairness criteria.

²⁵ The confidentiality of jury deliberations, certain sensitive cases, and personal data are some of the considerations that influence the extent of appropriate public examination and oversight mechanisms.

²⁶ The avoidance of negative consequences is important to note in relation to effectiveness. The law can be used for malevolent or intensely disputed purposes (for example, the quashing of dissent or mass incarceration). The instruments of the law, including A/IS, can render the advancement of such purposes more effective to the detriment of democratic values, human rights, and human well-being.

²⁷ Studies conducted by the US National Institute of Standards and Technology (NIST) between 2006 and 2011, known as the US NIST Text REtrieval Conference (TREC) Legal Track, suggest that some A/IS-enabled processes, if operated by trained experts in the relevant scientific fields, can be more effective (or accurate) than human attorneys in correctly identifying case-relevant information in large data sets. NIST has a long-standing reputation for cultivating trust in technology by participating in the development of standards and metrics that strengthen measurement science and make technology more secure, usable, interoperable, and reliable. This work is critical in the A/IS space to ensure public trust of rapidly evolving technologies so that we can benefit from all that this field has to promise.

²⁸ In describing the potential A/IS have for aiding in the auditing of decisions made in the civil and criminal justice systems, we are envisioning them acting as aids to a competent human auditor (see Issue 4) in the context of internal or judicial review.

²⁹ Of course, the use of A/IS in improving the effectiveness of law enforcement may raise concerns about other aspects of well-being, such as privacy and the rise of the surveillance state, cf. Minority Report (2002). If A/IS are to be used for law enforcement, steps must be taken to ensure that they are used, and that citizens trust that they will be used, in ways that are conducive to ethical law enforcement and individual well-being (see Issue 2).

Law

³⁰ A/IS may also provide assistance in carrying out legal tasks associated with larger transactions, such as evaluating contracts for risk in connection with a M&A transaction or reporting exposure to regulators.

³¹ The recommendations provided in this chapter (both under this issue and under the other issues discussed in the chapter) are intended to give general guidance as to how those with a stake in the just and effective operation of a legal system can develop norms for the trustworthy adoption of A/IS in the legal system. The specific ways in which the recommendations are operationalized will vary from society to society and from jurisdiction to jurisdiction.

³² See “Global Governance of AI Roundtable: Summary Report 2018,” World Government Summit, 2018: p. 32. Available: <https://www.worldgovernmentsummit.org/api/publications/document?id=ff6c88c5-e97c-6578-b2f8-ff0000a7ddb6>. (The February 2018 Dubai Global Governance of AI Roundtable brought together ninety leading thinkers on AI governance.)

³³ See *State v Loomis*, 881 N.W.2d 749 (Wis. 2016), *cert. denied* (2017); see also “Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing—*State v. Loomis*, 881 N.W.2d 749 (Wis. 2016),” Harvard Law Review, vol. 130, no. 5, pp. 1535-1536, 2017. Available: http://harvardlawreview.org/wp-content/uploads/2017/03/1530-1537_online.pdf; see also K. Freeman, “Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in *State v. Loomis*,” North Carolina Journal of Law and Technology,

vol. 18, no. 5, pp. 75-76, 2016. Available: <https://scholarship.law.unc.edu/ncjolt/vol18/iss5/3/>.

³⁴ An example of an initiative that seeks to bridge the gap between technical and legal expertise is the Artificial Intelligence Legal Challenge, held at Ryerson University and sponsored by Canada’s Ministry of the Attorney General: http://www.legalinnovationzone.ca/press_release/ryersons-legal-innovation-zone-announces-winners-of-ai-legal-challenge/.

³⁵ And, in addressing the challenges, consideration must be given to existing modes of proposing and approving innovation in the legal system. Trust in A/IS will be undermined if they are viewed as not having been vetted via established processes.

³⁶ For an overview of risk and risk management, see Working Party on Security and Privacy in the Digital Economy, Background Report for Ministerial Panel 3.2, Directorate for Science, Technology and Innovation, Committee on Digital Economy Policy, Managing Digital Security and Privacy Risk, OECD, June 1, 2016; see p. 5.

³⁷ It is worth emphasizing the “informed” qualifier we attach to trust here. Far from advocating for a “blind trust” in A/IS, we argue that A/IS should be adopted only when we have sound evidence of their effectiveness, when we can be confident of the competence of their operators, when we have assurances that these systems allow for the attribution of responsibility for outcomes (both positive and negative), and when we have clear views into their operation. Without those conditions, we would argue that *A/IS should not be adopted* in the legal system.

Law

³⁸ The importance of testing the effectiveness of advanced technologies, including A/IS, in the legal system (and beyond) is not new: it was highlighted by Judge Paul W. Grimm in an important early ruling on legal fact-finding, *Victor Stanley v. Creative Pipe, Inc.*, 250 F.R.D. 251, 257 (D. Md. 2008), followed, among others, by the influential research and educational institute The Sedona Conference as well as the International Organization for Standardization (ISO). See *An Open Letter to Law Firms and Companies in the Legal Tech Sector*, The Sedona Conference (2009), and *Commentary on Achieving Quality in the E-Discovery Process* (2013): 7; ISO standard on electronic discovery (ISO/IEC 27050-3:2017): 19. Most recently, in the summary report of the Global Governance of AI Roundtable at the 2018 World Government Summit, Omar bin Sultan Al Olama, Minister of State for Artificial Intelligence of the UAE, highlighted the importance of “empirical information” in assessing the suitability of A/IS.

³⁹ In the terminology of software development, *verification* is a demonstration that a given application meets a narrowly defined requirement; *validation* is a demonstration that the application answers its real-world use case. When we speak of gathering evidence of the effectiveness of A/IS, we are speaking of validation.

⁴⁰ Standards may include compliance with defined professional competence or other ethical requirements, but also other types of standards, such as data standards. Data standards may serve as “a digital lingua franca” with the potential of both supporting broad-based technological innovation (including A/IS innovation) in a legal

system and facilitating access to justice. As part of interactive technology solutions, appropriate data standards may help connect the ordinary citizen to the appropriate resources and information for his or her legal needs. For a discussion of open data standards in the context of the US court system, see D. Colarusso and E. J. Rickard, “Speaking the Same Language: Data Standards and Disruptive Technologies in the Administration of Justice,” *Suffolk University Law Review*, vol. L387, 2017.

⁴¹ For measurement of bias in facial recognition software, see C. Garvie, A. M. Bedoya, and J. Frankle, “The Perpetual Line-Up: Unregulated Police Face Recognition in America,” *Georgetown Law, Center on Privacy & Technology*, Oct. 2016. Available: <https://www.perpetuallineup.org/>.

⁴² The inclusion of such collateral effects in assessing effectiveness is an important element in overcoming the apparent “black box” or inscrutable nature of A/IS. See, for example, J. A. Kroll, “The fallacy of inscrutability,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, vol. 376, no. 2133, Oct. 2018. Available: doi.org/10.1098/rsta.2018.0084. The study addresses, among other questions, “how measurement of a system beyond understanding of its internals and its design can help to defeat inscrutability.”

⁴³ The question of the salience of collateral impact will vary with the specific application of A/IS. For example, false positives in document review related to fact-finding will generally not raise acute ethical issues, but false positives

Law

in predictive policing or sentencing will. In these latter domains, complex and sometimes unsettled issues of fairness arise, particularly when social norms of fairness change regionally and over time (sometimes rapidly). Any A/IS that was designed to replicate some notion of fairness would need to demonstrate its effectiveness, first, at replicating prevailing notions of fairness that have legitimacy in society, and second, at responding to evolutions in such notions of fairness. In the current state of A/IS, in which no system has been able to demonstrate consistent effectiveness in either of the above regards, it is essential that great discretion be exercised in considering any reliance on A/IS in domains such as sentencing and predictive policing.

⁴⁴ These exercises go by various names in the literature: *effectiveness evaluations*, *benchmarking exercises*, *validation studies*, and so on. See, for example, the definition of *validation study* in AINOW's 2018 *Algorithmic Accountability Toolkit* (<https://ainowinstitute.org/aap-toolkit.pdf>), p. 29. For our purposes, what matters is that the exercise be one that collects, in a scientifically sound manner, evidence of how “fit for purpose” any given A/IS are.

⁴⁵ This feature of evaluation design is important, as only tasks that accurately reflect real-world conditions and objectives (which may include the avoidance of unintended consequences, such as racial bias) will provide compelling guidance as to the suitability of an application for adoption in the real world.

⁴⁶ For TREC generally, see: <https://trec.nist.gov/>. For the TREC Legal Track specifically, see: <https://trec-legal.umiacs.umd.edu/>.

⁴⁷ When a complex system can be broken down into separate component systems, it may be appropriate to assess either the effectiveness of each component, or that of the end-to-end application as a whole (including human operators), depending on the specific question to be answered.

⁴⁸ Qualitative considerations may also help counter attempts to “game the system” (i.e., attempts to use bad-faith methods to meet a specific numerical target); see B. Hedin, D. Brassil, and A. Jones, “On the Place of Measurement in E-Discovery,” in *Perspectives on Predictive Coding and Other Advanced Search Methods for the Legal Practitioner*, ed. J. R. Baron, R. C. Losey, and M. D. Berman. Chicago: American Bar Association, 2016, p. 415f.

⁴⁹ Even in fact-finding, accurate extraction of facts does not eliminate the need for reasoned judgment as to the significance of the facts in the context of specific circumstances and cultural considerations. Used properly, A/IS will advance the spirit of the law, not just the letter of the law.

⁵⁰ Electronic discovery is the task of searching through large collections of electronically stored information (ESI) for material relevant to civil and criminal litigation and investigations. Among applications of A/IS to legal tasks and questions, the application to legal discovery is probably the most “mature,” as measured against the criteria of having been tested, assessed and approved by courts, and adopted fairly widely across various jurisdictions.

Law

⁵¹ While there is general consensus about the importance of these metrics in gauging effectiveness in legal discovery, there is not a consensus around the precise values for those metrics that must be met for a discovery effort to be acceptable. That is a good thing, as the precise value that should be attained, and demonstrated to have been attained, in any given matter will be dependent on, and proportional to, the specific facts and circumstances of that matter.

⁵² Different domains of application of A/IS to legal matters will vary not only with regard to the availability of consensus metrics of effectiveness, but also with regard to conditions that affect the challenge of measuring effectiveness: availability of data, impact of social bias, and sensitivity to privacy concerns all affect how difficult it may be to arrive at consensus protocols for gauging effectiveness. In the case of defining an effectiveness metric for A/IS used in support of sentencing decisions, one challenge is that, while it is easy to find when an individual who has been released commits a crime (or is convicted of committing a crime), it is difficult to assess when an individual who was not released would have committed a crime. For a discussion of the challenges in measuring the effectiveness of tools designed to assess flight risk, see M. T. Stevenson, "Assessing Risk Assessment in Action." *Minnesota Law Review*, vol. 103, 2018. Available: doi.org/10.2139/ssrn.3016088.

⁵³ Sound measurement may also serve as an effective antidote to the unsubstantiated claims sometimes made regarding the effectiveness of certain applications of A/IS to legal matters

(e.g., flight risk assessment technologies); see Stevenson, "Assessing Risk Assessment". Unsubstantiated claims are an appropriate source of an *informed distrust* in A/IS. Such well-founded distrust can be addressed only with truly meaningful and sound measures that provide accurate information regarding the capabilities and limitations of a given system.

⁵⁴ See the discussion under "Illustration—Effectiveness" in this chapter.

⁵⁵ For more on principles for data protection, see the EAD chapter "Personal Data and Individual Agency".

⁵⁶ The importance of validation by practitioners is reflected in The European Commission's High-Level Expert Group on Artificial Intelligence Draft Ethics Guidelines for Trustworthy AI: "Testing and validation of the system should thus occur as early as possible and be iterative, ensuring the system behaves as intended throughout its entire life cycle *and especially after deployment.*" (Emphasis added.) See High-Level Expert Group on Artificial Intelligence, "DRAFT Ethics Guidelines for Trustworthy AI: Working Document for Stakeholders' Consultation," The European Commission. Brussels, Belgium: Dec. 18, 2018.

⁵⁷ That scrutiny need not extend to IP or other protected information (e.g., attorney work product). Validation methods and results are a matter of numbers and procedures for obtaining the numbers, and their disclosure would not impinge on safeguards against the disclosure of legitimately protected information.

Law

⁵⁸ A recent matter from the US legal system illustrates how a failure to disclose the results of a validation exercise can limit the exercise's ability to achieve its intended purpose. In *Winfield v. City of New York* (Opinion & Order. 15-CV-05236 [LTS] [KHP]. SDNY 2017), a party had utilized the A/IS-enabled system to conduct a review of documents for relevance to the matter being litigated. When the accuracy and completeness of the results of that review were challenged by the requesting party, the producing party disclosed that it had, in fact, conducted validation of its results. Rather than requiring that the producing party simply disclose the results of the validation to the requesting party, the judge overseeing the dispute chose to review the results herself *in camera*, without providing access to the requesting party. Although the judge then said that the evidence she was provided supported the accuracy and completeness of the review, the requesting party could not itself examine either the evidence or the methods whereby it was obtained, and so could not gain confidence in the results. That confidence comes only from examining the metrics and the procedures followed in obtaining them. Moreover, the results of a validation exercise, which are usually simple numbers that reflect sampling procedures, can be disclosed without revealing the content of any documents, any proprietary tools or methods, or any attorney work product. If the purpose of conducting a validation exercise is to gather evidence of the effectiveness of a process, in the event that the process is challenged, keeping that evidence hidden from those who would challenge the process limits the ability of the validation exercise to achieve its intended purpose.

⁵⁹ <https://www.nist.gov/>.

⁶⁰ TREC Legal Track (2006-2011): <https://trec-legal.umiacs.umd.edu/>.

⁶¹ The statistical evidence in question here is statistical evidence of the effectiveness of A/IS applied to the task of discovery; it is not statistical evidence of facts actually at issue in litigation. Courts may have different rules for the admissibility of the two kinds of statistical evidence (and there will be jurisdictional differences on these questions).

⁶² It is important to underscore that, whereas developers and operators of A/IS should be able to derive sound measurements of effectiveness, the courts should determine what level of effectiveness—what score—should be demonstrated to have been achieved, based on the facts and circumstances of a given matter. In some instances, the cost (in terms of sample sizes, resources required to review the samples, and so on) of demonstrating the achievement of a high score will be disproportionate to the stakes of a given matter. In others, for example, a major securities fraud claim that potentially affects thousands of citizens, a court might justifiably demand a demonstration of the achievement of a very high score, irrespective of cost. Demonstrations of the effectiveness of A/IS (and of their operators) are instruments in support of, not in substitution of, judicial decision-making.

⁶³ See, for example, B. Hedin, S. Tomlinson, J. R. Baron, and D. W. Oard, "Overview of the TREC 2009 Legal Track," in *NIST Special Publication: SP 500-278, The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings* (2009).

Law

⁶⁴ See M. R. Grossman and G. V. Cormack, "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review," *Richmond Journal of Law and Technology*, vol. 17, no. 3, 2011. Available: <http://jolt.richmond.edu/jolt-archive/v17i3/article11.pdf>. Note that the two systems that conclusively demonstrated "better than human" performance took methodologically distinct approaches, but they shared the characteristic of having been designed, operated, and measured for accuracy by scientifically trained experts.

⁶⁵ *Da Silva Moore v. Publicis Groupe*, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012). See also A. Peck, "Search, Forward," *Legaltech News*. Oct. 1, 2011. Available: <https://www.law.com/legaltechnews/almID/1202516530534Search-Forward/>.

⁶⁶ The fact that NIST has as important role to play in developing standards for the measurement of the safety and security of A/IS was recognized in a recent (September, 2018) report from the U.S. House of Representatives: "At minimum, a widely agreed upon standard for measuring the safety and security of AI products and applications should precede any new regulations. ... The National Institute of Standards and Technology (NIST) is situated to be a key player in developing standards." (Will Hurd and Robin Kelly, "Rise of the Machines: Artificial Intelligence and its Growing Impact on U.S. Policy," U.S. House of Representatives—Committee on Oversight and Government Reform—Subcommittee on Information Technology, September, 2018).

⁶⁷ The competence principle is intended to apply to the post design operation of A/IS. Of course, that does not mean that designers and developers of A/IS are free of responsibility for their systems' outcomes. As discussed in the background to this issue, it is incumbent on designers and developers to assess the risks associated with the operation of their systems and to specify the operator competencies needed to mitigate those risks. For more on the question of designer incompetence or negligence, see the discussion of "software malpractice" in Kroll (2018).

⁶⁸ The ISO standard on e-discovery, ISO/IEC 27050-3, does recognize the importance of expertise in applying advanced technologies in a search for documents responsive to a legal inquiry; see ISO/IEC 27050-3: *Information technology — Security techniques — Electronic discovery — Part 3: Code of practice for electronic discovery*, Geneva (2017), pp. 19-20.

⁶⁹ See, for example, ABA Model Rule 1, comment 8: "To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject." Available: https://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1/competence/comment_on_rule_1_1/. See also, The State Bar of California Standing Committee on Professional Responsibility and Conduct, Formal Opinion No. 2015-193. Available:

Law

[https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL%202015-193%20%5B11-0004%5D%20\(06-30-15\)%20-%20FINAL.pdf](https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL%202015-193%20%5B11-0004%5D%20(06-30-15)%20-%20FINAL.pdf).

⁷⁰ In the deliberations of the Law Committee of the 2018 Global Governance of AI Roundtable, the question of the competencies needed “in order to effectively operate and measure the efficacy of AI systems in legal functions that affect the rights and liberty of citizens” was cited as one of the considerations that “appear to be most overlooked in the current public dialogue.” See “Global Governance of AI Roundtable: Summary Report 2018,” World Government Summit, 2018: p. 7. Available: <https://www.worldgovernmentsummit.org/api/publications/document?id=ff6c88c5-e97c-6578-b2f8-ff0000a7ddb6>.

⁷¹ See A. G. Ferguson, “Policing Predictive Policing,” *Washington University Law Review*, vol. 94, no. 5, 2017: 1109, 1172. Available: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/.

⁷² In addition, a lack of competence in interpreting the results of a statistical exercise can (and often does) result in an incorrect conclusion (on the part of a party to a dispute or of a judge seeking to resolve a dispute). For example, in *In re: Biomet*, a judge addressing a discovery dispute interpreted the statistical data provided by the producing party as indicating that the producing party’s retrieval process had left behind “a comparatively modest number” of responsive documents, when the statistical evidence showed, in fact, that a substantial number of responsive documents had been left behind.

See *In re: Biomet M2a Magnum Hip Implant Prods. Liab. Litig.* No. 3:12-MD-2391 (N.D. Ind. April 18, 2013).

⁷³ For example, a prior violent conviction may be weighted equally, whether the violent act was a shove or a knife attack. See Human Rights Watch. “Q & A: Profile Based Risk Assessment for US Pretrial Incarceration, Release Decisions,” June 1, 2018. Available: <https://www.hrw.org/news/2018/06/01/q-profile-based-risk-assessment-us-pretrial-incarceration-release-decisions>.

⁷⁴ Bias can be introduced in a number of ways: via the features taken into consideration by the algorithm, via the nature and composition of the training data, via the design of the validation protocol, and so on. A competent operator will be alert to and assess such potential sources of bias.

⁷⁵ Among the conditions may be, for example, that the results of the system are to be used only to provide guidance to the human decision maker (e.g., judge) and should not be taken as, in themselves, dispositive.

⁷⁶ Given that the effective functioning of a legal system is a matter of interest to the whole of society, it is important that all members of a society be provided with access to the resources needed to understand when and how A/IS are applied in support of the functioning of a legal system.

⁷⁷ Among the topics covered by such training should be the potential for “automation bias” and ways to mitigate it. See L. J. Skitka, K. Mosier, and M. D. Burdick, “Does automation

Law

bias decision-making?" *International Journal of Human-Computer Studies*, vol. 51, no. 5, pp. 991-1006, 1999. Available: <https://doi.org/10.1006/ijhc.1999.0252>; L. J. Skitka, K. Mosier, and M. D. Burdick, "Accountability and automation bias," *International Journal of Human-Computer Studies*, vol. 52, no. 4, pp. 701-717, 2000. Available: <https://doi.org/10.1006/ijhc.1999.0349>.

⁷⁸ Some government agencies are working toward creating a more effective partnership between the skills found in technology start-ups and the skills required of legal practitioners. See Legal Innovation Zone. "Ryerson's Legal Innovation Zone Announces Winners of AI Legal Challenge," March 26, 2018. Available: http://www.legalinnovationzone.ca/press_release/ryersons-legal-innovation-zone-announces-winners-of-ai-legal-challenge/.

⁷⁹ See Amazon. "Amazon Rekognition." <https://aws.amazon.com/rekognition/> (2018).

⁸⁰ See E. Dwoskin, "Amazon is selling facial recognition to law enforcement—for a fistful of dollars." *Washington Post*, May 22, 2018. Available: https://www.washingtonpost.com/news/the-switch/wp/2018/05/22/amazon-is-selling-facial-recognition-to-law-enforcement-for-a-fistful-of-dollars/?noredirect=on&utm_term=.07d9ca13ab77.

⁸¹ See, for example, J. Stanley, "FBI and Industry Failing to Provide Needed Protections for Face Recognition." *ACLU—Free Future*, June 15, 2016. Available: <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/fbi-and-industry-failing-provide-needed>.

⁸² It is also the case that, among the false positives, nonwhite members of Congress were overrepresented relative to their proportion in Congress as a whole, perhaps indicating that the accuracy of the technology is, to some degree, race-dependent. Without knowing more about the composition of the mugshot database, however, we cannot assess the significance of this result.

⁸³ See J. Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots." *ACLU—Free Future*, July 26, 2018. Available: <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28>. See also R. Brandom, "Amazon's facial recognition matched 28 members of Congress to criminal mugshots." *The Verge*, July 26, 2018. Available: <https://www.theverge.com/2018/7/26/17615634/amazon-rekognition-aclu-mug-shot-congress-facial-recognition>.

⁸⁴ See "Amazon Rekognition Developer Guide." Amazon, p. 131, 2018. Available: <https://docs.aws.amazon.com/rekognition/latest/dg/rekognition-dg.pdf>. Also see K. Tenbarge, "Amazon Responds to ACLU's Highly Critical Report of Rekognition Tech," *Inverse*, July 26, 2018. Available: <https://www.yahoo.com/news/amazon-responds-aclu-aapos-highly-160000264.html>.

⁸⁵ The story also highlights the question of accountability, illustrating how the principles discussed in this report intersect with and complement each other.

Law

⁸⁶ Of course, competent use does not preclude use for bad ends (e.g., government surveillance that impinges on human rights). The principle of competence is one principle in a set that, collectively, is designed to ensure the ethical application of A/IS. See the EAD chapter “General Principles”.

⁸⁷ Developing “well grounded” guidelines will typically require that the creators of A/IS gather input from both those operating the technology and those affected by the technology’s operation.

⁸⁸ The use of facial recognition technologies by security and law enforcement agencies raises issues that extend beyond the question of operator competence. For further discussion of such issues, see C. Garvie, A. M. Bedoya, and J. Frankle, “The Perpetual Line-Up: Unregulated Police Face Recognition in America,” *Georgetown Law, Center on Privacy & Technology*, October 18, 2016, Available: <https://www.perpetuallineup.org/>.

⁸⁹ As noted above, some professional organizations, such as the ABA, have begun to recognize in their codes of ethics the importance of technological competence, although the guidance does not yet address A/IS specifically.

⁹⁰ Including those engaged in the procurement and deployment of a system means that those acquiring and authorizing the use of a system can share in the responsibility for its results. For example, in the case of A/IS deployed in the service of the courts, this could be the judiciary; in the case of A/IS deployed in the service of law enforcement, this could be the agency responsible for the enforcement of the law and

the administration of justice; in the case of A/IS used by a party to legal proceedings, this could be the party’s counsel.

⁹¹ J. New and D. Castro, “How Policymakers Can Foster Algorithmic Accountability,” *Information Technology & Innovation Foundation*, p. 5, 2018. Available: <https://www.itif.org/publications/2018/05/21/how-policymakers-can-foster-algorithmic-accountability>.

⁹² Included among possible “causes” for an effect are not only the decision-making pathways of algorithms but also, importantly, the decisions made by humans involved in the design, development, procurement, deployment, operation, and validation of effectiveness of A/IS.

⁹³ The challenge, moreover, is one not only of assigning responsibility, but of assigning levels of responsibility (a task that could benefit from a neutral model that could consider how much interaction and influence each stakeholder has in every decision).

⁹⁴ Scherer (2016): 372. In addition to diffuseness, Scherer identifies discreteness, discreteness, and opacity as features of the design and development of A/IS that make apportioning responsibility for their outcomes a challenge for regulators and courts.

⁹⁵ In answering these questions, it will be important to keep in mind the distinction between responsibility (a factual question) and ultimate accountability (a normative question). In the case of the example under discussion, there may be multiple individuals who have

Law

some practical responsibility for the sentence given, but the normative framework may place ultimate accountability on the judge. Before normative accountability can be assigned, however, pragmatic responsibilities must be clarified and understood. Hence the focus, in this section, on clarifying lines of responsibility so that ultimate accountability can be determined.

⁹⁶ If effectiveness is measured against statistics that themselves may represent human bias (e.g., arrest rates), then the effectiveness measures may just reflect and reinforce that bias.

⁹⁷ “The algorithm did it’ is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.” See “Principles for Accountable Algorithms and a Social Impact Statement for Algorithms.” FAT/ML Resources. www.fatml.org/resources/principles-for-accountable-algorithms.

⁹⁸ See Langewiesche, W. 1998. “The Lessons of ValuJet 592”. *Atlantic Monthly*. 281: 81-97; S. D. Sagan. *Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press, 1995.

⁹⁹ For a discussion of the role of explanation in maintaining accountability for the results of A/IS and of the question of whether the standards for explanation should be different for A/IS than they are for humans, see F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. J. Gershman, D. O’Brien, S. Shieber, J. Waldo, D. Weinberger, and A. Wood, Accountability of AI Under the Law: The Role of Explanation (November 3, 2017). Berkman Center Research Publication Forthcoming; Harvard Public Law Working

Paper No. 18-07. Available: <https://ssrn.com/abstract=3064761> or <http://dx.doi.org/10.2139/ssrn.3064761>.

¹⁰⁰ Also, gaining access to that information should not be unduly burdensome.

¹⁰¹ Those developing a model for accountability for A/IS may find helpful guidance in considering models of accountability used in other domains (e.g., data protection).

¹⁰² For a discussion of how such policies might be implemented in accordance with protocols for information governance, see J. R. Baron and K. E. Armstrong, “The Algorithm in the C-Suite: Applying Lessons Learned and Information Governance Best Practices to Achieve Greater Post-GDPR Algorithmic Accountability,” in *The GDPR Challenge: Privacy, Technology, and Compliance In An Age of Accelerating Change*, A. Taal, Ed. Boca Raton, FL: CRC Press, forthcoming.

¹⁰³ These inquiries can be supported by technological tools that may provide information essential to answering questions of accountability but that do not require full transparency into underlying computer code and may avoid the necessity of an intrusive audit; see Kroll et al. (2017). Among the tools identified by Kroll and his colleagues are: software verification, cryptographic commitments, zero-knowledge proofs, and fair random choices. While the use of such tools may avoid the limitations of solutions such as transparency and audit, they do require that creators of A/IS design their systems so that they will be compatible with the application of such tests.

Law

¹⁰⁴ Certifications may include, for example, professional certifications of competence, but also certifications of compliance of processes with standards. An example of a certification program specifically addressing A/IS is *The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)*, <https://standards.ieee.org/industry-connections/ecpais.html>.

¹⁰⁵ This means that A/IS used in legal systems will have to be defensible in courts. The margin of error will have to be low or the use of A/IS will not be permitted.

¹⁰⁶ It is also the case that evidence produced by A/IS will be subject to chain-of-custody rules, as are other types of forensic evidence, to ensure integrity, confidentiality, and authenticity.

¹⁰⁷ See for instance Art. 22(1) Regulation (EU) 2016/679.

¹⁰⁸ Human dignity, as a core value protected by the United Nations Universal Declaration of Human Rights, requires us to fully respect the personality of each human being and prohibits their objectification.

¹⁰⁹ This concern is reflected in Principle 5 of the European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment, recently published by the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ). Principle 5 ("Principle 'Under User Control': preclude a prescriptive approach and ensure that users are informed actors and in control of the choices made") states, with regard to professionals in the justice system that they should "at any moment, be able to review judicial decisions and the data used to produce a result

and continue not to be necessarily bound by it in the light of the specific features of that particular case," and, with regard to decision subjects, that he or she must "be clearly informed of any prior processing of a case by artificial intelligence before or during a judicial process and have the right to object, so that his/her case can be heard directly by a court." See CEPEJ, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment* (Strasbourg, 2018), p. 10.

¹¹⁰ J. Tashea, [Calculating Crime: Attorneys are Challenging the Use of Algorithms to Help Determine Bail, Sentencing and Parole](#), ABA Journal (March 2017).

¹¹¹ [Loomis v. Wisconsin](#), 68 WI. (2016).

¹¹² *Id.* at pp. 46-66.

¹¹³ R. Wexler, [Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System](#), Stanford Law Review, 2018.

¹¹⁴ [Malenchik v. State](#), 928 N.E.2d 564, 574 (Ind. 2010).

¹¹⁵ [People v. Chubbs](#) CA2/4, B258569 (Cal. Ct. App. 2015).

¹¹⁶ *U.S. v. Ocasio*, No. 3:11-cr-02728-KC, slip op. at 1-2, 11-12 (W.D. Tex. May 28, 2013).

¹¹⁷ *U.S. v. Johnson*, No. 1:15-cr-00565-VEC, order (S.D.N.Y., June 7, 2016).

¹¹⁸ Indeed, without transparency, there may, in some circumstances, be no means for even knowing whether an error that needs to be corrected was committed. In the case of A/IS

Law

applied in a legal system, an “error” can mean real harm to the dignity, liberty, and life of an individual.

¹¹⁹ *Fairness* (as well as *bias*) can be defined in more than one way. For purposes of this discussion, a commitment is not made to any one definition—and indeed, it may not be either desirable or feasible to arrive at a single definition that would be applied in all circumstances. For purposes of this discussion, the key point is that transparency will be essential in building informed trust in the fairness of a system, regardless of the specific definition of *fairness* that is operative.

¹²⁰ To the extent permitted by the normal operation of the A/IS: allowing for, for example, variation in the human inputs to a system that may not be eliminated in any attempt at replication.

¹²¹ With regard to information explaining how a system arrived at a given output, GDPR makes provision for a decision subject’s right to an explanation of algorithmic decisions affecting him or her: automated processing of personal data “should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.” GDPR, Recital 71.

¹²² Even among sensitive data, some data may be more sensitive than others. See I. Ajunwa, “Genetic Testing Meets Big Data: Tort and Contract Law Issues,” 75 *Ohio St. L. J.* 1225 (2014). Available: <https://ssrn.com/abstract=2460891>.

¹²³ See A. Baker, “Updated N.Y.P.D. Anti-Crime System to Ask: ‘How We Doing?’” *New York Times*, May 8, 2017, <https://www.nytimes.com/2017/05/08/nyregion/nypd-compstat-crime-mapping.html>; S. Weichselbaum, “How a ‘Sentiment Meter’ Helps Cops Understand Their Precincts,” *Wired*, July 16, 2018. Available: <https://www.wired.com/story/elucd-sentiment-meter-helps-cops-understand-precincts/>.

¹²⁴ This table is a preliminary draft and is meant only to illustrate a useful tool for facilitating reasoning about who should have access to what information. Other categories of stakeholder and other categories of information (e.g., the identity and nature of the designer/manufacture of the A/IS, the identity and nature of the investors backing a particular system or company) could be added as needed.

¹²⁵ For discussions of these two dimensions of explanation, see S. Wachter, et al. (2017). “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”; A. Selbst, and S. Barocas, *The Intuitive Appeal of Explainable Machines*.

¹²⁶ Wexler, Rebecca. 2018. “Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System”. *Stanford Law Review*. 70 (5): 1342-1429; Tashea, Jason. “Federal judge releases DNA software source code that was used by New York City’s crime lab.” *ABA Journal* (2017). http://www.abajournal.com/news/article/federal_judge_releases_dna_software_source_code.

¹²⁷ Or, if two approaches are found to be, for practical purposes, equally effective, the simpler, more easily explained approach may be preferred.

Law

¹²⁸ For a discussion of the limits of transparency and of alternative modes of gaining actionable answers to questions of verification and accountability, see J.A. Kroll, J. Huey, S. Barocas, E.W. Felten, J.R. Reidenberg, D.G. Robinson, H. Yu, "Accountable Algorithms" (March 2, 2016). *University of Pennsylvania Law Review*, Vol. 165, 2017 Forthcoming; Fordham Law Legal Studies Research Paper No. 2765268. Available at SSRN: <https://ssrn.com/abstract=2765268>. See also J.A. Kroll, The fallacy of inscrutability, *Phil. Trans. R. Soc. A* 376: 20180084. <http://dx.doi.org/10.1098/rsta.2018.0084> (Note p. 9: "While transparency is often taken to mean the disclosure of source code or data, possibly to a trusted entity such as a regulator, this is neither necessary nor sufficient for improving understanding of a system, and it does not capture the full meaning of transparency.")

¹²⁹ In particular with respect to due process, the current dialogue on the use of A/IS centers on the tension between the need for transparency and the need for the protection of intellectual property rights. Adhering to the principle of Effectiveness as articulated in this work can substantially help in defusing this tension. Reliable empirical evidence of the effectiveness of A/IS in meeting specific real-world objectives may foster informed trust in such A/IS, without disclosure of proprietary or trade secret information.

¹³⁰ S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," SSRN Electronic Journal, p. 5, 2017 for the example cited.

¹³¹ W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations," The RAND Corporation, pp. 67-69, 2013.

¹³² Support from the University of Memphis was led by Richard Janikowski, founding Director of the Center for Community Criminology and Research (School of Urban Affairs and Public Policy, the University of Memphis) and the Shared Urban Data System (The University of Memphis).

¹³³ E. Figg, "The Legacy of Blue CRUSH," High Ground, March 19, 2014.

¹³⁴ Figg, "Legacy."

¹³⁵ Nucleus Research, *ROI Case Study: IBM SPSS—Memphis Police Department*, Boston, Mass., Document K31, June 2010. Perry et al., *Predictive Policing*, 69.

¹³⁶ Figg, "Legacy."

¹³⁷ Figg, "Legacy."

¹³⁸ See: AI Now, *Algorithmic Accountability Policy Toolkit*, p. 12, Oct. 2018. Available: <https://ainowinstitute.org/aap-toolkit.pdf>; D. Robinson and L. Koepke, *Stuck in a Pattern: Early evidence on "predictive policing" and civil rights*, Upturn, Aug. 2016. Available: <https://www.upturn.org/reports/2016/stuck-in-a-pattern/>; S. Brayne, "Big Data Surveillance: The Case of Policing," *American Sociological Review*, 2016. Available: <https://journals.sagepub.com/doi/10.1177/0003122417725865>; A. G. Ferguson, "Policing Predictive Policing,"

Law

Washington University Law Review, vol. 94, no. 5, 2017. Available: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/; K. Lum and W. Isaac, "To predict and serve?" *Significance* 2016. Available: <https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2016.00960.x>; B. J. Jefferson, "Predictable Policing: Predictive Crime Mapping and Geographies of Policing and Race," *Annals of the American Association of Geographers*, vol. 108, no. 1, pp. 1-16, 2018. Available: <https://doi.org/10.1080/24694452.2017.1293500>.

¹³⁹ For a discussion of the criteria that may define a "high-crime area," and so potentially license more intrusive policing, see A. G. Ferguson and D. Bernache, "The 'High-Crime Area' Question: Requiring Verifiable and Quantifiable Evidence for Fourth Amendment Reasonable Suspicion Analysis," *American University Law Review*, vol. 57, pp. 1587-1644.

¹⁴⁰ While A/IS, if misapplied, may perpetuate bias, it holds at least the potential, if applied with appropriate controls, to reduce bias. For a study of how an impersonal technology such as a red light camera may reduce bias, see R. J. Eger, C. K. Fortner, and C. P. Slade, "The Policy of Enforcement: Red Light Cameras and Racial Profiling," *Police Quarterly*, pp. 1-17, 2015. Available: <http://hdl.handle.net/10945/46909>.

¹⁴¹ See, for example: J. Tashea, "Estonia considering new legal status for artificial intelligence," *ABA Journal*, Oct. 20, 2017, and European Parliament [Resolution of Feb. 16, 2017](#).

¹⁴² See Legal Entity, Person, in B. Bryan A. Garner, *Black's Law Dictionary*, 10th Edition. Thomas West, 2014.

¹⁴³ J. S. Nelson, "Paper Dragon Thieves." *Georgetown Law Journal* 105 (2017): 871-941.

¹⁴⁴ M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." *Nevada Law Journal* 19, forthcoming 2018.

¹⁴⁵ See M. U. Scherer, "Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems." *Nevada Law Journal* 19, forthcoming 2018; J. F. Weaver. [Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws](#). Santa Barbara, CA: Praeger, 2013; L. B. Solum. "[Legal Personhood for Artificial Intelligences](#)." *North Carolina Law Review* 70, no. 4 (1992): 1231-1287.

About *Ethically Aligned Design*

The Mission and Results of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.

To advance toward this goal, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems brought together more than a thousand participants from six continents who are thought leaders from academia, industry, civil society, policy, and government in the related technical and humanistic disciplines to identify and find consensus on timely issues surrounding autonomous and intelligent systems.

By “stakeholder” we mean anyone involved in the research, design, manufacture, or messaging around intelligent and autonomous systems—including universities, organizations, governments, and corporations—all of which are making these technologies a reality for society.



About *Ethically Aligned Design*

From Principles to Practice— Results from our Work to Date

In addition to the creation of *Ethically Aligned Design*, The IEEE Global Initiative, independently or through the IEEE Standards Association, has directly inspired the following works:

- **The launch of the IEEE P7000™ series of approved standardization projects**

This is the first series of standards in the history of the IEEE Standards Association that explicitly focuses on societal and ethical issues associated with a certain field of technology

More information can be found at:
ethicsinaction.ieee.org

- **Artificial Intelligence and Ethics in Design**

These ten courses are designed for global professionals, as well as their managers, working in engineering, IT, computer science, big data, artificial intelligence, and related fields across all industries who require up-to-date information on the latest technologies. The courses explicitly mirror content from *Ethically Aligned Design*, and feature numerous experts as instructors who helped create *Ethically Aligned Design*.

More information can be found at:
innovationatwork.ieee.org/courses/artificial-intelligence-and-ethics-in-design

- **The creation of an A/IS Ethics Glossary**

The Glossary features more than two hundred pages of terms that help to define the context of A/IS ethics for multiple stakeholder groups, specifically: engineers, policy makers, philosophers, standards developers, and computational disciplines experts. It is currently in its second iteration and has also been informed by the IEEE P7000™ standards working groups.

Download the Glossary at:
standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_glossary.pdf

- **The launch of OCEANIS**

The IEEE Standards Association, inspired by the work of The IEEE Global Initiative, has contributed significantly to the establishment of The Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS). It is a global forum for discussion, debate, and collaboration for organizations interested in the development and use of standards to further the creation of autonomous and intelligent systems. OCEANIS members are working together to enhance the understanding of the role of standards in facilitating innovation, while addressing problems that expand beyond technical solutions to addressing ethics and values.

More information can be found at:
ethicsstandards.org

About *Ethically Aligned Design*

- **The launch of ECPAIS**

The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) has the goal to create specifications for certification and marking processes that advance transparency, accountability, and reduction in algorithmic bias in autonomous and intelligent systems. ECPAIS intends to offer a process and define a series of marks by which organizations can seek certifications for their processes around the A/IS products, systems, and services they provide.

More information can be found at:

standards.ieee.org/industry-connections/ecpais.html

- **The launch of CXI**

The Council on Extended Intelligence (CXI) was directly inspired by the work of The IEEE Global Initiative and the work of The MIT Media Lab around “Extended Intelligence”. CXI was launched jointly by the IEEE Standards Association and The MIT Media Lab. CXI’s mission is to proliferate the ideals of responsible participant design, data agency, and metrics of economic prosperity, prioritizing people and the planet over profit and productivity. Membership includes thought leaders from the EU Parliament and Commission, the UK House of Lords, the OECD, the United Nations, local and national administrations, and renowned experts in economics, data science, and multiple other disciplines from around the world.

More information can be found at:

globalcxi.org

- **The launch of EADUC**

The Ethically Aligned Design University Consortium (EADUC) is being established with the aim to reach every engineer at the beginning of their studies to help them prioritize values-driven, applied ethical principles at the core of their work. Working in conjunction with philosophers, designers, social scientists, academics, data scientists, and the corporate and policy communities, EADUC also has the goal that *Ethically Aligned Design* will be used in teaching at all levels of education globally as the new vision for design in the algorithmic age.

- **The launch of “AI Commons”**

The work of The IEEE Global Initiative has delivered key ideas and inspiration that are rapidly evolving toward establishing a global collaborative platform around A/IS. The mission of AI Commons is to gather a true ecosystem to democratize access to AI capabilities and thus to allow anyone, anywhere to benefit from the possibilities that AI can provide. In addition, the group will be working to connect problem owners with the community of solvers, to collectively create solutions with AI. The ultimate goal is to implement a framework for participation and cooperation to make using and benefiting from AI available to all.

More information can be found at:

www.aicommons.com

About *Ethically Aligned Design*

IEEE P7000™ Approved Standardization Projects

The IEEE P7000™ series of standards projects under development represents a unique addition to the collection of over 1,900 global IEEE standards and projects. Whereas more traditional standards have a focus on technology interoperability, functionality, safety, and trade facilitation, the IEEE P7000 series addresses specific issues at the intersection of technological and ethical considerations. Like its technical standards counterparts, the IEEE P7000 series empowers innovation across borders and enables societal benefit.

For more information or to join any working group, please see the links below. Committees that authored *Ethically Aligned Design*, as well as other committees within IEEE, that created specific working groups are listed below each project.

- **IEEE P7000™ - IEEE Standards Project Model Process for Addressing Ethical Concerns During System Design**
Inspired by Methodologies to Guide Ethical Research and Design Committee, and supported by IEEE Computer Society
standards.ieee.org/project/7000.html
- **IEEE P7001™ - IEEE Standards Project for Transparency of Autonomous Systems**
Inspired by the General Principles Committee, and supported by IEEE Vehicular Technology Society
standards.ieee.org/project/7001.html
- **IEEE P7002™ - IEEE Standards Project for Data Privacy Process**
Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
standards.ieee.org/project/7002.html
- **IEEE P7003™ - IEEE Standards Project for Algorithmic Bias Considerations**
Supported by IEEE Computer Society
standards.ieee.org/project/7003.html
- **IEEE P7004™ - IEEE Standards Project for Child and Student Data Governance**
Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
standards.ieee.org/project/7004.html

About *Ethically Aligned Design*

- **IEEE P7005™ - IEEE Standards Project for Employer Data Governance**
Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
standards.ieee.org/project/7005.html
- **IEEE P7006™ - IEEE Standards Project for Personal Data AI Agent Working Group**
Inspired by The Personal Data and Individual Agency Control Committee, and supported by IEEE Computer Society
standards.ieee.org/project/7006.html
- **IEEE P7007™ - IEEE Standards Project for Ontological Standard for Ethically Driven Robotics and Automation Systems**
Supported by IEEE Robotics and Automation Society
standards.ieee.org/project/7007.html
- **IEEE P7008™ - IEEE Standards Project for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems**
Inspired by the Affective Computing Committee, and supported by IEEE Robotics and Automation Society
standards.ieee.org/project/7008.html
- **IEEE P7009™ - IEEE Standards Project for Fail-Safe Design of Autonomous and Semi-Autonomous Systems**
Supported by IEEE Reliability Society
standards.ieee.org/project/7009.html
- **IEEE P7010™ - IEEE Standards Project for Well-being Metric for Autonomous and Intelligent Systems**
Inspired by the Well-being Committee, and supported by IEEE Systems, Man and Cybernetics Society
standards.ieee.org/project/7010.html
- **IEEE P7011™ - IEEE Standards Project for the Process of Identifying and Rating the Trustworthiness of News Sources**
Supported by IEEE Society on Social Implications of Technology
standards.ieee.org/project/7011.html
- **IEEE P7012™ - IEEE Standards Project for Machine Readable Personal Privacy Terms**
Supported by IEEE Society on Social Implications of Technology
standards.ieee.org/project/7012.html
- **IEEE P7013™ - IEEE Standards Project for Inclusion and Application Standards for Automated Facial Analysis Technology**
Supported by IEEE Society on Social Implications of Technology
standards.ieee.org/project/7013.html

About *Ethically Aligned Design*

Who We Are

About IEEE

IEEE is the largest technical professional organization dedicated to advancing technology for the benefit of humanity, with over 420,000 members in more than 160 countries. Through its highly cited publications, conferences, technology standards, and professional and educational activities, IEEE is the trusted voice in a wide variety of areas ranging from aerospace systems, computers, and telecommunications to biomedical engineering, electric power, and consumer electronics.

To learn more, visit the IEEE website:
www.ieee.org

About the IEEE Standards Association

The IEEE Standards Association (IEEE-SA), a globally recognized standards-setting body within IEEE, develops consensus standards through an open process that engages industry and brings together a broad stakeholder community. IEEE standards set specifications and best practices based on current scientific and technological knowledge. The IEEE-SA has a portfolio of over 1,900 active standards and over 650 standards under development.

For more information, visit the IEEE-SA website: standards.ieee.org

About The IEEE Global Initiative

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (The IEEE Global Initiative) is a program of the IEEE

Standards Association with the status of an Operating Unit of The Institute of Electrical and Electronics Engineers, Incorporated (IEEE), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity with over 420,000 members in more than 160 countries.

To learn more, visit The IEEE Global Initiative website:

standards.ieee.org/industry-connections/ec/autonomous-systems.html

The IEEE Global Initiative provides the opportunity to bring together multiple voices in the related technological and scientific communities to identify and find consensus on timely issues.

Names of experts involved in the various committees of The IEEE Global Initiative can be found at: standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ec_bios.pdf

IEEE makes all versions of *Ethically Aligned Design* available under the Creative Commons Attribution-Non-Commercial 4.0 United States License. Subject to the terms of that license, organizations or individuals can adopt aspects of this work at their discretion at any time. It is also expected that *Ethically Aligned Design* content and subject matter will be selected for submission into formal IEEE processes, including standards development and education purposes.

The IEEE Global Initiative and *Ethically Aligned Design* contribute, together with other efforts within IEEE, such as IEEE TechEthics™, (techethics.ieee.org), to a broader effort at IEEE to foster open, broad, and inclusive conversation about ethics in technology.

About *Ethically Aligned Design*

Our Process

To ensure the greatest cultural relevance and intellectual rigor possible in our work, The IEEE Global Initiative has sought for and received global feedback for versions 1 and 2 (after hundreds of experts created first drafts) to inform this *Ethically Aligned Design, First Edition (EAD1e)*.

We released [*Ethically Aligned Design, Version 1 \(EADv1\)*](#) as a Request for Input in December of 2016 and received [over two hundred pages](#) of in-depth feedback about the draft. We subsequently released [*Ethically Aligned Design, Version 2 \(EADv2\)*](#) in December 2017 and received [over three hundred pages](#) of in-depth feedback about the draft. This feedback included further insights about the eight original sections from EADv1, along with unique/new input for the five new sections included in EADv2.

Both versions included “candidate recommendations” instead of direct “recommendations”, because our communities had been engaged in debate and weighing various options.

This process was taken to the next level with *Ethically Aligned Design, First Edition (EAD1e)*, using EADv1 and EADv2 as its initial foundation. Although we expect future editions of *Ethically Aligned Design*, a vetting process has taken place within the global community that gave rise to this seminal work. Therefore, we can now speak of “recommendations” without any further restriction, and *EAD1e* also includes a set of policy recommendations.

This process included matters of “internal consistency” across the various chapters of *EAD1e* and also more specific or broader criteria, such as maturity of the specific chapters and consistency with respect to policy statements of IEEE. The review also considered the need for IEEE to maintain a neutral and thus credible position in areas and processes where it is likely that IEEE may become active in the future.

Beyond these formal procedures, the Board of Governors of IEEE Standards Association has endorsed the work of the IEEE Global Initiative and offers it for consideration by governments, businesses, and the public at large with the following resolution:

Whereas the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is an authorized activity within the IEEE Standards Association Industry Connections program created with the stated mission:

To ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity;

Whereas versions 1 and 2 of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)* were developed as calls for comment and candidate recommendations by several hundred professionals including engineers, scientists,

About *Ethically Aligned Design*

ethicists, sociologists, economists, and many others from six continents;

Whereas the recommendations contained in *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition* are the result of the consideration of hundreds of comments submitted by professionals and the public at large on versions 1 and 2;

Whereas through an extensive, global, and open collaborative process, more than a thousand experts of The IEEE Global Initiative have developed and are in the process of final editing and publishing *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition*; now, therefore, be it

Resolved, that the IEEE Standards Association Board of Governors:

1. expresses its appreciation to the leadership and members of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the creation of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition*; and
2. supports and commends the collaborative process used by The IEEE Global Initiative to achieve extraordinary consensus in such complex and vast matters in less than three years; and

3. endorses and offers *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), First Edition* to businesses, governments and the public at large for consideration and guidance in the ethical development of autonomous and intelligent systems.

Terminology Update

For *Ethically Aligned Design*, we prefer not to use—as far as possible—the vague term “AI” and use instead the term, *autonomous and intelligent systems* (or *A/IS*). Even so, it is inherently difficult to define “intelligence” and “autonomy”. One could, however, limit the scope for practical purposes to computational systems using algorithms and data to address complex problems and situations, including the capability of improving their performance based on evaluating previous decisions, and say that such systems could be considered as “intelligent”.

Such systems could be regarded also as “autonomous” in a given domain as long as they are capable of accomplishing their tasks despite environment changes within the given domain. This terminology is applied throughout *Ethically Aligned Design, First Edition* to ensure the broadest possible application of ethical considerations in the design of the addressed technologies and systems.

About *Ethically Aligned Design*

How the Document Was Prepared

This document was developed by The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which is an authorized Industry Connections activity within the IEEE Standards Association, a Major Organizational Unit of IEEE.

It was prepared using an open, collaborative, and consensus building approach, following the processes of the [Industry Connections framework program](https://standards.ieee.org/industry-connections) of the IEEE Standards Association (standards.ieee.org/industry-connections). This process does not necessarily incorporate all comments or reflect the views of every contributor listed in the Acknowledgements above or after each chapter of this work.

The views and opinions expressed in this collaborative work are those of the authors and do not necessarily reflect the official policy or position of their respective institutions or of the Institute of Electrical and Electronics Engineers (IEEE). This work is published under the auspices of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems for the purposes of furthering public understanding of the importance of addressing ethical considerations in the design of autonomous and intelligent systems.

In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors, omissions or damage, direct or otherwise, however caused, arising in any way out of the use of or application of any recommendation contained in this publication.

The Board of Governors of the IEEE Standards Association, its highest governing body, commends the consensus-building process used in developing *Ethically Aligned Design*, First Edition, and offers the work for consideration and guidance in the ethical development of autonomous and intelligent systems.

How to Cite *Ethically Aligned Design*

Please cite *Ethically Aligned Design*, First Edition in the following manner:

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, First Edition. IEEE, 2019. <https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html>

Key References

Key References

Key reference documents listed in *Ethically Aligned Design, First Edition*:

- **Appendix 1** - The State of Well-being Metrics (An Introduction)
bit.ly/ead1e-appendix1
(Referenced in *Well-being Section*)
- **Appendix 2** - The Happiness Screening Tool for Business Product Decisions
bit.ly/ead1e-appendix2
(Referenced in *Well-being Section*)
- **Appendix 3** - Additional Resources: Standards Development Models and Frameworks
bit.ly/ead1e-appendix3
(Referenced in *Well-being Section*)
- **Glossary**
bit.ly/ead1e-glossary



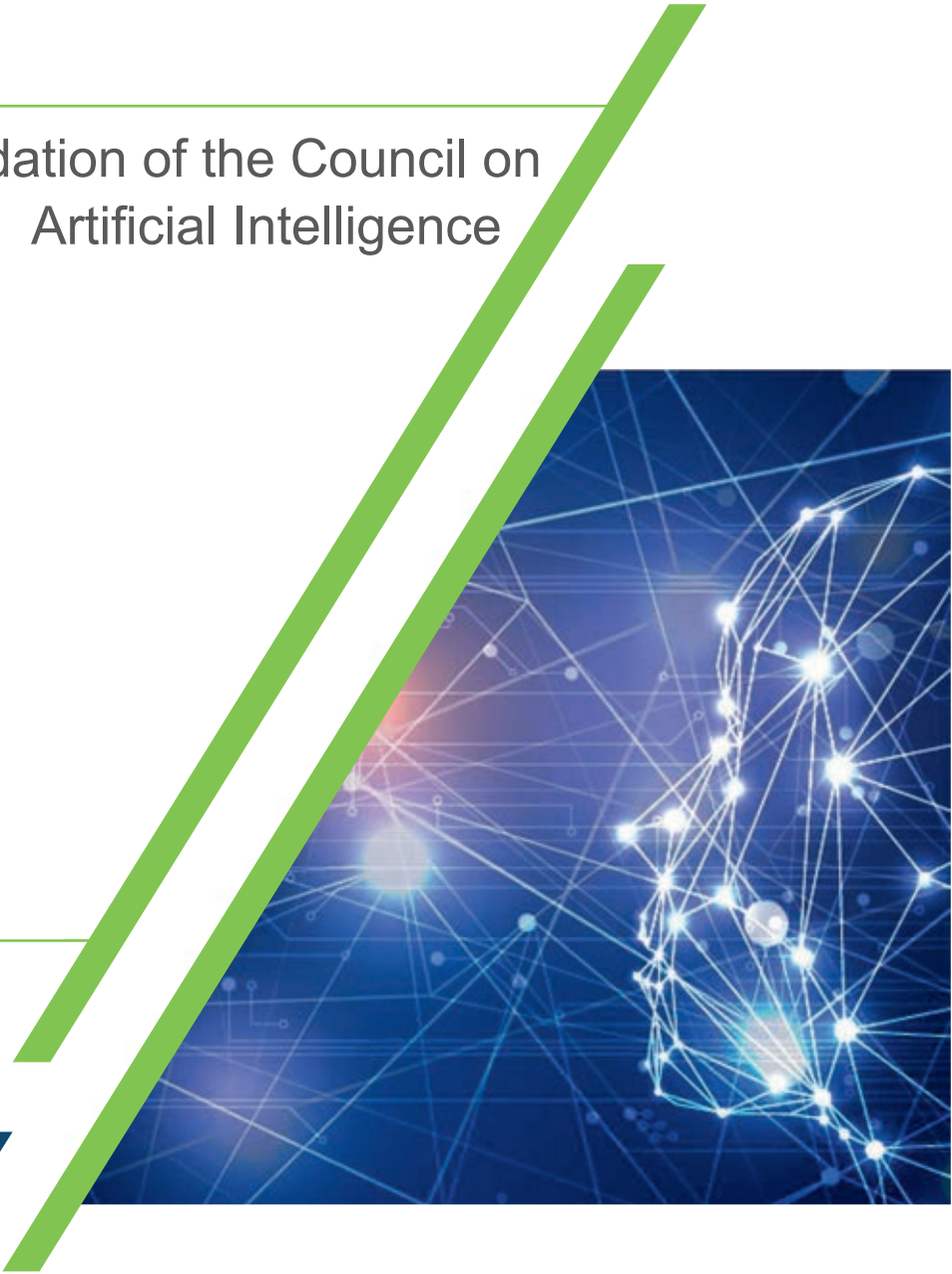
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") is a program of The Institute of Electrical and Electronics Engineers, Incorporated ("IEEE"), the world's largest technical professional organization dedicated to advancing technology for the benefit of humanity with over 420,000 members in more than 160 countries. The IEEE Global Initiative and Ethically Aligned Design contribute to broader efforts at IEEE about ethics in technology.





Recommendation of the Council on Artificial Intelligence

**OECD Legal
Instruments**



This document is published under the responsibility of the Secretary-General of the OECD. It reproduces an OECD Legal Instrument and may contain additional material. The opinions expressed and arguments employed in the additional material do not necessarily reflect the official views of OECD Member countries.

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

For access to the official and up-to-date texts of OECD Legal Instruments, as well as other related information, please consult the Compendium of OECD Legal Instruments at <http://legalinstruments.oecd.org>.

Please cite this document as:

OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449

Series: OECD Legal Instruments

Photo credit: © kras99/Shutterstock.com

© OECD 2019

This document is provided free of charge. It may be reproduced and distributed free of charge without requiring any further permissions, as long as it is not altered in any way. It may not be sold.

This document is available in the two OECD official languages (English and French). It may be translated into other languages, as long as the translation is labelled "unofficial translation" and includes the following disclaimer: *"This translation has been prepared by [NAME OF TRANSLATION AUTHOR] for informational purpose only and its accuracy cannot be guaranteed by the OECD. The only official versions are the English and French texts available on the OECD website <http://legalinstruments.oecd.org>"*

Date(s)

Adopted on 22/05/2019

Background Information

The Recommendation on Artificial Intelligence (AI) – the first intergovernmental standard on AI – was adopted by the OECD Council at Ministerial level on 22 May 2019 on the proposal of the Committee on Digital Economy Policy (CDEP). The Recommendation aims to foster innovation and trust in AI by promoting the responsible stewardship of trustworthy AI while ensuring respect for human rights and democratic values. Complementing existing OECD standards in areas such as privacy, digital security risk management, and responsible business conduct, the Recommendation focuses on AI-specific issues and sets a standard that is implementable and sufficiently flexible to stand the test of time in this rapidly evolving field. In June 2019, at the Osaka Summit, G20 Leaders welcomed G20 AI Principles, drawn from the OECD Recommendation.

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI and calls on AI actors to promote and implement them:

- inclusive growth, sustainable development and well-being;
- human-centred values and fairness;
- transparency and explainability;
- robustness, security and safety;
- and accountability.

In addition to and consistent with these value-based principles, the Recommendation also provides five recommendations to policy-makers pertaining to national policies and international co-operation for trustworthy AI, namely:

- investing in AI research and development;
- fostering a digital ecosystem for AI;
- shaping an enabling policy environment for AI;
- building human capacity and preparing for labour market transformation;
- and international co-operation for trustworthy AI.

The Recommendation also includes a provision for the development of metrics to measure AI research, development and deployment, and for building an evidence base to assess progress in its implementation.

The OECD's work on Artificial Intelligence and rationale for developing the OECD Recommendation on Artificial Intelligence

Artificial Intelligence (AI) is a general-purpose technology that has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges. It is deployed in many sectors ranging from production, finance and transport to healthcare and security.

Alongside benefits, AI also raises challenges for our societies and economies, notably regarding economic shifts and inequalities, competition, transitions in the labour market, and implications for democracy and human rights.

The OECD has undertaken empirical and policy activities on AI in support of the policy debate over the past two years, starting with a Technology Foresight Forum on AI in 2016 and an international conference on *AI: Intelligent Machines, Smart Policies* in 2017. The Organisation also conducted analytical and measurement work that provides an overview of the AI technical landscape, maps economic and social impacts of AI technologies and their applications, identifies major policy considerations, and describes AI initiatives from governments and other stakeholders at national and international levels.

This work has demonstrated the need to shape a stable policy environment at the international level to foster trust in and adoption of AI in society. Against this background, the OECD Committee on Digital Economy Policy (CDEP) agreed to develop a draft Council Recommendation to promote a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and applies to all stakeholders.

Complementing existing OECD standards already relevant to AI – such as those on privacy and data protection, digital security risk management, and responsible business conduct – the Recommendation focuses on policy issues that are specific to AI and strives to set a standard that is implementable and flexible enough to stand the test of time in a rapidly evolving field. The Recommendation contains five high-level values-based principles and five recommendations for national policies and international co-operation. It also proposes a common understanding of key terms, such as “AI system” and “AI actors”, for the purposes of the Recommendation.

More specifically, the Recommendation includes two substantive sections:

1. **Principles for responsible stewardship of trustworthy AI:** the first section sets out five complementary principles relevant to all stakeholders: *i)* inclusive growth, sustainable development and well-being; *ii)* human-centred values and fairness; *iii)* transparency and explainability; *iv)* robustness, security and safety; and *v)* accountability. This section further calls on AI actors to promote and implement these principles according to their roles.
2. **National policies and international co-operation for trustworthy AI:** consistent with the five aforementioned principles, this section provides five recommendations to Members and non-Members having adhered to the draft Recommendation (hereafter the “Adherents”) to implement in their national policies and international co-operation: *i)* investing in AI research and development; *ii)* fostering a digital ecosystem for AI; *iii)* shaping an enabling policy environment for AI; *iv)* building human capacity and preparing for labour market transformation; and *v)* international co-operation for trustworthy AI.

An inclusive and participatory process for developing the Recommendation

The development of the Recommendation was participatory in nature, incorporating input from a broad range of sources throughout the process. In May 2018, the CDEP agreed to form an expert group to scope principles to foster trust in and adoption of AI, with a view to developing a draft Council Recommendation in the course of 2019. The AI Group of experts at the OECD (AIGO) was subsequently established, comprising over 50 experts from different disciplines and different sectors (government, industry, civil society, trade unions, the technical community and academia) - see <http://www.oecd.org/going-digital/ai/oecd-aigo-membership-list.pdf> for the full list. Between September 2018 and February 2019 the group held four meetings: in Paris, France, in September and November 2018, in Cambridge, MA, United States, at the Massachusetts Institute of Technology (MIT) in January 2019, back to back with the MIT AI Policy Congress, and finally in Dubai, United Arab Emirates, at the World Government Summit in February 2019. The work benefited from the diligence, engagement and substantive contributions of the experts participating in AIGO, as well as from their multi-stakeholder and multidisciplinary backgrounds.

Drawing on the final output document of the AIGO, a draft Recommendation was developed in the CDEP and with the consultation of other relevant OECD bodies. The CDEP approved a final draft Recommendation and agreed to transmit it to the OECD Council for adoption in a special meeting on 14-15 March 2019. The OECD Council adopted the Recommendation at its meeting at Ministerial level on 22-23 May 2019.

Follow-up, monitoring of implementation and dissemination tools

The OECD Recommendation on AI provides the first intergovernmental standard for AI policies and a foundation on which to conduct further analysis and develop tools to support governments in their implementation efforts. In this regard, it instructs the CDEP to monitor the implementation of the Recommendation and report to the Council on its implementation and continued relevance five years after its adoption and regularly thereafter. The CDEP is also instructed to continue its work on AI, building on this Recommendation, and taking into account work in other international fora, such as UNESCO, the European Union, the Council of Europe and the initiative to build an International Panel on AI (see <https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence> and <https://www.gouvernement.fr/en/france-and-canada-create-new-expert-international-panel-on-artificial-intelligence>).

In order to support implementation of the Recommendation, the Council instructed the CDEP to develop practical guidance for implementation, to provide a forum for exchanging information on AI policy and activities, and to foster multi-stakeholder and interdisciplinary dialogue. This will be achieved largely through the OECD AI Policy Observatory, an inclusive hub for public policy on AI that aims to help countries encourage, nurture and monitor the responsible development of trustworthy artificial intelligence systems for the benefit of society. It will combine resources from across the OECD with those of partners from all stakeholder groups to provide multidisciplinary, evidence-based policy analysis on AI. The Observatory is planned to be launched late 2019 and will include a live database of AI strategies, policies and initiatives that countries and other stakeholders can share and update, enabling the comparison of their key elements in an interactive manner. It will also be continuously updated with AI metrics, measurements, policies and good practices that could lead to further updates in the practical guidance for implementation.

The Recommendation is open to non-OECD Member adherence, underscoring the global relevance of OECD AI policy work as well as the Recommendation's call for international co-operation.

Unofficial translation(s): [German](#), [Japanese](#).

For further information please consult: oecd.ai.

Contact information: ai@oecd.org.

THE COUNCIL,

HAVING REGARD to Article 5 b) of the Convention on the Organisation for Economic Co-operation and Development of 14 December 1960;

HAVING REGARD to the OECD Guidelines for Multinational Enterprises [[OECD/LEGAL/0144](#)]; Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data [[OECD/LEGAL/0188](#)]; Recommendation of the Council concerning Guidelines for Cryptography Policy [[OECD/LEGAL/0289](#)]; Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information [[OECD/LEGAL/0362](#)]; Recommendation of the Council on Digital Security Risk Management for Economic and Social Prosperity [[OECD/LEGAL/0415](#)]; Recommendation of the Council on Consumer Protection in E-commerce [[OECD/LEGAL/0422](#)]; Declaration on the Digital Economy: Innovation, Growth and Social Prosperity (Cancún Declaration) [[OECD/LEGAL/0426](#)]; Declaration on Strengthening SMEs and Entrepreneurship for Productivity and Inclusive Growth [[OECD/LEGAL/0439](#)]; as well as the 2016 Ministerial Statement on Building more Resilient and Inclusive Labour Markets, adopted at the OECD Labour and Employment Ministerial Meeting;

HAVING REGARD to the Sustainable Development Goals set out in the 2030 Agenda for Sustainable Development adopted by the United Nations General Assembly (A/RES/70/1) as well as the 1948 Universal Declaration of Human Rights;

HAVING REGARD to the important work being carried out on artificial intelligence (hereafter, “AI”) in other international governmental and non-governmental fora;

RECOGNISING that AI has pervasive, far-reaching and global implications that are transforming societies, economic sectors and the world of work, and are likely to increasingly do so in the future;

RECOGNISING that AI has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges;

RECOGNISING that, at the same time, these transformations may have disparate effects within, and between societies and economies, notably regarding economic shifts, competition, transitions in the labour market, inequalities, and implications for democracy and human rights, privacy and data protection, and digital security;

RECOGNISING that trust is a key enabler of digital transformation; that, although the nature of future AI applications and their implications may be hard to foresee, the trustworthiness of AI systems is a key factor for the diffusion and adoption of AI; and that a well-informed whole-of-society public debate is necessary for capturing the beneficial potential of the technology, while limiting the risks associated with it;

UNDERLINING that certain existing national and international legal, regulatory and policy frameworks already have relevance to AI, including those related to human rights, consumer and personal data protection, intellectual property rights, responsible business conduct, and competition, while noting that the appropriateness of some frameworks may need to be assessed and new approaches developed;

RECOGNISING that given the rapid development and implementation of AI, there is a need for a stable policy environment that promotes a human-centric approach to trustworthy AI, that fosters research, preserves economic incentives to innovate, and that applies to all stakeholders according to their role and the context;

CONSIDERING that embracing the opportunities offered, and addressing the challenges raised, by AI applications, and empowering stakeholders to engage is essential to fostering adoption of trustworthy AI in society, and to turning AI trustworthiness into a competitive parameter in the global marketplace;

On the proposal of the Committee on Digital Economy Policy:

I. **AGREES** that for the purpose of this Recommendation the following terms should be understood as follows:

- *AI system*: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.
- *AI system lifecycle*: AI system lifecycle phases involve: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase.
- *AI knowledge*: AI knowledge refers to the skills and resources, such as data, code, algorithms, models, research, know-how, training programmes, governance, processes and best practices, required to understand and participate in the AI system lifecycle.
- *AI actors*: AI actors are those who play an active role in the AI system lifecycle, including organisations and individuals that deploy or operate AI.
- *Stakeholders*: Stakeholders encompass all organisations and individuals involved in, or affected by, AI systems, directly or indirectly. AI actors are a subset of stakeholders.

Section 1: Principles for responsible stewardship of trustworthy AI

II. **RECOMMENDS** that Members and non-Members adhering to this Recommendation (hereafter the "Adherents") promote and implement the following principles for responsible stewardship of trustworthy AI, which are relevant to all stakeholders.

III. **CALLS ON** all AI actors to promote and implement, according to their respective roles, the following Principles for responsible stewardship of trustworthy AI.

IV. **UNDERLINES** that the following principles are complementary and should be considered as a whole.

1.1. Inclusive growth, sustainable development and well-being

Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, sustainable development and well-being.

1.2. Human-centred values and fairness

- a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights.
- b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art.

1.3. Transparency and explainability

AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:

- i. to foster a general understanding of AI systems,
- ii. to make stakeholders aware of their interactions with AI systems, including in the workplace,
- iii. to enable those affected by an AI system to understand the outcome, and,
- iv. to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision.

1.4. Robustness, security and safety

- a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.
- b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.
- c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.

1.5. Accountability

AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.

Section 2: National policies and international co-operation for trustworthy AI

V. RECOMMENDS that Adherents implement the following recommendations, consistent with the principles in section 1, in their national policies and international co-operation, with special attention to small and medium-sized enterprises (SMEs).

2.1. Investing in AI research and development

- a) Governments should consider long-term public investment, and encourage private investment, in research and development, including interdisciplinary efforts, to spur innovation in trustworthy AI that focus on challenging technical issues and on AI-related social, legal and ethical implications and policy issues.
- b) Governments should also consider public investment and encourage private investment in open datasets that are representative and respect privacy and data protection to support an environment for AI research and development that is free of inappropriate bias and to improve interoperability and use of standards.

2.2. Fostering a digital ecosystem for AI

Governments should foster the development of, and access to, a digital ecosystem for trustworthy AI. Such an ecosystem includes in particular digital technologies and infrastructure, and mechanisms for sharing AI

knowledge, as appropriate. In this regard, governments should consider promoting mechanisms, such as data trusts, to support the safe, fair, legal and ethical sharing of data.

2.3. Shaping an enabling policy environment for AI

- a) Governments should promote a policy environment that supports an agile transition from the research and development stage to the deployment and operation stage for trustworthy AI systems. To this effect, they should consider using experimentation to provide a controlled environment in which AI systems can be tested, and scaled-up, as appropriate.
- b) Governments should review and adapt, as appropriate, their policy and regulatory frameworks and assessment mechanisms as they apply to AI systems to encourage innovation and competition for trustworthy AI.

2.4. Building human capacity and preparing for labour market transformation

- a) Governments should work closely with stakeholders to prepare for the transformation of the world of work and of society. They should empower people to effectively use and interact with AI systems across the breadth of applications, including by equipping them with the necessary skills.
- b) Governments should take steps, including through social dialogue, to ensure a fair transition for workers as AI is deployed, such as through training programmes along the working life, support for those affected by displacement, and access to new opportunities in the labour market.
- c) Governments should also work closely with stakeholders to promote the responsible use of AI at work, to enhance the safety of workers and the quality of jobs, to foster entrepreneurship and productivity, and aim to ensure that the benefits from AI are broadly and fairly shared.

2.5. International co-operation for trustworthy AI

- a) Governments, including developing countries and with stakeholders, should actively co-operate to advance these principles and to progress on responsible stewardship of trustworthy AI.
- b) Governments should work together in the OECD and other global and regional fora to foster the sharing of AI knowledge, as appropriate. They should encourage international, cross-sectoral and open multi-stakeholder initiatives to garner long-term expertise on AI.
- c) Governments should promote the development of multi-stakeholder, consensus-driven global technical standards for interoperable and trustworthy AI.
- d) Governments should also encourage the development, and their own use, of internationally comparable metrics to measure AI research, development and deployment, and gather the evidence base to assess progress in the implementation of these principles.

VI. INVITES the Secretary-General and Adherents to disseminate this Recommendation.

VII. INVITES non-Adherents to take due account of, and adhere to, this Recommendation.

VIII. INSTRUCTS the Committee on Digital Economy Policy:

- a) to continue its important work on artificial intelligence building on this Recommendation and taking into account work in other international fora, and to further develop the measurement framework for evidence-based AI policies;
- b) to develop and iterate further practical guidance on the implementation of this Recommendation, and to report to the Council on progress made no later than end December 2019;
- c) to provide a forum for exchanging information on AI policy and activities including experience with the implementation of this Recommendation, and to foster multi-stakeholder and interdisciplinary dialogue to promote trust in and adoption of AI; and

- d) to monitor, in consultation with other relevant Committees, the implementation of this Recommendation and report thereon to the Council no later than five years following its adoption and regularly thereafter.

Adherents*

OECD Members		Non-Members	Other
Australia	United States	Argentina	
Austria		Brazil	
Belgium		Colombia	
Canada		Costa Rica	
Chile		Peru	
Czech Republic		Romania	
Denmark		Ukraine	
Estonia			
Finland			
France			
Germany			
Greece			
Hungary			
Iceland			
Ireland			
Israel			
Italy			
Japan			
Korea			
Latvia			
Lithuania			
Luxembourg			
Mexico			
Netherlands			
New Zealand			
Norway			
Poland			
Portugal			
Slovak Republic			
Slovenia			
Spain			
Sweden			
Switzerland			
Turkey			
United Kingdom			

* Additional information and statements are available in the Compendium of OECD Legal Instruments:
<http://legalinstruments.oecd.org>

About the OECD

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD Member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Lithuania, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Legal Instruments

Since the creation of the OECD in 1961, around 450 substantive legal instruments have been developed within its framework. These include OECD Acts (i.e. the Decisions and Recommendations adopted by the OECD Council in accordance with the OECD Convention) and other legal instruments developed within the OECD framework (e.g. Declarations, international agreements).

All substantive OECD legal instruments, whether in force or abrogated, are listed in the online Compendium of OECD Legal Instruments. They are presented in five categories:

- **Decisions:** OECD legal instruments which are legally binding on all Members except those which abstain at the time of adoption. While they are not international treaties, they entail the same kind of legal obligations. Adherents are obliged to implement Decisions and must take the measures necessary for such implementation.
- **Recommendations:** OECD legal instruments which are not legally binding but practice accords them great moral force as representing the political will of Adherents. There is an expectation that Adherents will do their utmost to fully implement a Recommendation. Thus, Members which do not intend to do so usually abstain when a Recommendation is adopted, although this is not required in legal terms.
- **Declarations:** OECD legal instruments which are prepared within the Organisation, generally within a subsidiary body. They usually set general principles or long-term goals, have a solemn character and are usually adopted at Ministerial meetings of the Council or of committees of the Organisation.
- **International Agreements:** OECD legal instruments negotiated and concluded within the framework of the Organisation. They are legally binding on the Parties.
- **Arrangement, Understanding and Others:** several ad hoc substantive legal instruments have been developed within the OECD framework over time, such as the Arrangement on Officially Supported Export Credits, the International Understanding on Maritime Transport Principles and the Development Assistance Committee (DAC) Recommendations.

Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015

New Artificial Intelligence Tools For Deep Conflict Resolution and Humanitarian Response

Daniel J. Olsher^a

^a*Integral Mind Technologies, Washington DC, United States
dan@intmind.com*

Abstract

Truly understanding what others need and want, how they see the world, and how they feel are core prerequisites for successful conflict resolution and humanitarian response. Today, however, human cognitive limitations, insufficient expertise in the right hands, and difficulty in managing complex social, conflict, and real-world knowledge conspire to prevent us from reaching our ultimate potential. This paper introduces *cogSolv*, a highly novel Artificial Intelligence system capable of understanding how people from other groups view the world, simulating their reactions, and combining this with knowledge of the real world in order to persuade, find negotiation win-wins and enhance outcomes, avoid offense, provide peacekeeping decision tools, and protect emergency responders' health. Ready to go today, portable, and requiring virtually no specialist expertise, *cogSolv* allows governments and local NGOs to use expert culture and conflict resolution knowledge to accurately perform a wide range of humanitarian simulations. *cogSolv* assists responders with training, managing complexity, centralizing and sharing knowledge, and, ultimately, maximizing the potential for equitable conflict resolution and maximally effective humanitarian response.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of the Organizing Committee of HumTech2015

Keywords: Conflict Resolution; Humanitarian Response; Peacekeeping; Culture; Knowledge; Artificial Intelligence

1. Introduction and Theoretical Motivation

Humans have proven themselves to be remarkable conflict resolvers, persuaders, and responders to humanitarian disasters of all kinds. Practically speaking, however, responders find themselves confronted by a myriad of cognitive and organizational limitations. Humanitarian contexts are characterized by complex, difficult-to-predict social systems grounded in psychology, culture, and deep knowledge bases. The information needed for response is often distributed across multiple experts, and is difficult to synthesize in ways sufficient to guide response. Countless fragments of information interact in unpredictable ways, making it exceedingly difficult to obtain the 'big picture' and truly understand what is going on. Moreover, NGOs, local groups, and government agencies alike often lack meaningful access to conflict resolution, cultural, and other key knowledge. Therefore, successful conflict resolution and humanitarian response often tend to require a certain amount of *luck* - having the right people come together with the right information.

One reason for this is that, often, critical knowledge is unconscious and not easily accessed or standardized, including cultural and other social knowledge as well as expert knowledge. Nowhere is this more true than when responders must work with those holding worldviews different than their own; the tendency to fall into ethnocentric traps and ignore key aspects of the other side's worldview is very difficult to avoid.

Yet, when seeking to work with and/or convince others who think differently from us, we will only achieve success if we design appeals with respect to the other side's *true* (and often unexpressed) point of view.

Furthermore, it is easy to overlook conflict solutions that *appear to be equitable* but in fact ignore key needs and values for the other side. In disaster response, perceived cultural insensitivity may cause survivors to ignore official communications such as evacuation orders [1], and the inability to manage complex chemical, equipment-related, and other practical knowledge often gives rise to critical health risks.

In the past, factors such as these have led to missed opportunities, renewed conflicts, sub-optimal outcomes, structural violence, and, ultimately, the loss of life. In the case of peacekeeping missions, characterized by the sending of signals that must be correctly understood by those with diverse worldviews, failure may mean the breaking of a ceasefire, rioting, or the resumption of war. Many knowledgeable commentators (see [2], for example) suggest that the failure of UNOSOM II (the mission upon which the movie *Black Hawk Down* was based) was due precisely to factors such as these.

When peacekeeping leaders 'get the call', there often isn't sufficient time to undertake deep study of the cultures they will be working within. (*personal communication, ex-SRSG*) As demonstrated by UNITAR training scenarios, it can be difficult indeed for peacekeeping battalion commanders to determine how to proceed in culturally-appropriate ways. Given the demonstrated need to devolve ever-increasing amounts of decision-making power to the field (cf. [3]), future commanders are likely to find themselves more and more dependent on incomplete information.

As an example, one such UNITAR training scenario, set in Africa, imagines an ex-soldier who has climbed a fence and broken into a UN MOVCON warehouse. Breaking his Rules Of Engagement (ROE), the fictitious peacekeeper shoots the ex-soldier. A crowd begins to gather outside the base, demanding the ex-soldier's body, and the commander must decide what to do. Using models developed in conjunction with a Ugandan informant, simulations have shown that, in such a situation, it would be essential for the UN to engage to some extent with local conflict resolution processes if further bloodshed were to be avoided. It is most probable, however, that under such a scenario the necessary knowledge would not be available to local decision-makers and they would not be aware of this.

Generally speaking, computers hold immense potential for helping humans overcome difficulties such as these. Unfortunately, however, in the past they have been unable to do so, as mainstream Artificial Intelligence (AI) has not had the ability to store and handle nuanced social data in a way that would allow it to in some sense 'understand' and productively model these types of complex systems.

With the recent advent of the *Atomic* approach to AI, however, this has now become possible. This school of thought represents a fundamentally new perspective on the discipline. Two core Atomic formalisms, *INTELNET* and *COGVIEW* [4], provide the foundation for *cogSolv*, the suite of technologies discussed in this paper. *INTELNET* allows highly nuanced knowledge about the world to be stored and conclusions drawn from it in exceedingly flexible, powerful ways. *COGVIEW* enables computers to conduct simulations grounded in complex psychological and cultural worldviews. Critically, *COGVIEW* models (known as Deep MindMaps) are human-readable *and* machine-processable at the same time, meaning that they can be created with only minimal training and used by personnel without significant specialist expertise. The exact same data that is entered into the computer can be easily used for teaching and discussion purposes.

As Spriet and Vansteenkiste (1982 in [5]) suggest, "Social systems are sometimes labeled in the literature as soft systems or ill-defined systems where the usefulness of traditional mathematical representations is questioned." *COGVIEW* models allow us to understand complex human situations while retaining their nuance, using flexible, brain-inspired algorithms to effect processing. *Ultimately, this enables us to generate remarkably human-like predictions across complex social systems.*

cogSolv and Atomic AI are optimized for the type of data found in humanitarian environments; in such contexts, the 'softer' aspects make all the difference.[4] *COGVIEW* is able to integrate disparate forms of information (such as emotional and practical/commonsense knowledge) quickly and effectively.

1.1. What Can *cogSolv* Do?

In conflict resolution, negotiation, advocacy, persuasion, peacekeeping, disaster response, and other key humanitarian processes, *cogSolv* simulations provide precise guidance as to how to respond, pointing out actions that should be undertaken or strenuously avoided. *cogResolv*, the conflict-focused component of *cogSolv*, can store and simulate expert conflict resolution techniques, automatically integrating these with situational/cultural models developed by field and HQ experts.

cogResolv acts as a *trusted advisor and ally* before, during, and after the mission, centralizing cultural and practical data. In protracted conflict or when stalemates arise, the computer helps find ways around blockages. *cogResolv* simulates the effects of actions and the perceptions that they will create for other parties, identifies hidden win-wins and potential problems, circumvents biases, and helps discover actions that can reinforce the resulting peace. It helps meet needs in creative ways, maximizing 'deep' (integrative) justice.

In line with GRIT (Gradual Reduction In Tensions) theory [6], cogResolv can suggest potential concessions that may reduce tensions while maximizing value for all sides. It makes the hidden explicit, models critical psychological factors such as pain and determination, helps increase decision quality, and models the ripple effects of small decisions across large, complex social systems.

cogResolv helps conflictants separate issues during negotiations, making all parties aware of the totality of the world in which they operate. Its *Integrative Justice Scores* provide a quick, concise metric of the extent to which the deep needs of all parties are being taken into account and hidden biases addressed.

Facilitating situational awareness, cogResolv allows practitioners to work together to manipulate a shared vision of a current situation and to visually indicate points of reference or areas of concern.

cogSolv and cogResolv also support *training and situational awareness*; officials sent to conflict sites on a moment's notice, peacekeepers, and students can all benefit from cogSolv's ability to quickly and easily facilitate understanding. cogSolv enables team members to quickly appreciate the existence, importance, and consequences of critical knowledge, helping to get everyone on the same page.

In summary, cogSolv's Artificial Intelligence capabilities provide decision-makers with critical tools for making socially-nuanced life-or-death decisions.

2. Core Humanitarian Focus Areas

Current cogSolv/cogResolv focus areas include:

- Conflict modeling/prediction, including protracted conflict,
- Persuasion (especially emotionally/subconsciously-driven: beliefs, values, religion),
- Social media analysis, including sentiment/topic detection and modeling,
- Knowledge/culture-based deep analysis of extremist messages,
- Nuanced conflict understanding and training,
- Peacekeeping,
- Disaster response, and
- Conflict early warning (grounded in analysis of prevailing social scenarios and social media inputs).

3. Potential Users

cogSolv and cogResolv find applicability to a wide range of humanitarian and conflict-sensitive domains, including the following:

- **Peacekeeping:** Interactions with local populations, calming tensions, mission design, gender sensitivity.
Who: Field battalion leaders, UN Department for Peacekeeping Operations (DPKO) personnel / HQ.
- **Development:** Locally-sensitive intervention design, anti-discrimination advocacy, empowerment of sex workers, gender sensitivity, calming of tensions.
Who: Field personnel, planners.
- **Early Warning/Data Mining/Machine Learning:** cogSolv capabilities for natural language and social media processing point the way to a capacity for early warning of conflict hotspots or likely social ruptures. cogSolv and the associated COGBASE[7] commonsense knowledge base support data mining, machine learning and deep learning, as well as other processes for discovering patterns in input data.
- **Diplomacy:** International negotiations, cooperation in international organizations (ASEAN, UNSC), human rights (especially elements oriented towards values, religions, cultures and other intangible variables). Notably included are resource-oriented conflicts, especially when multiple issues may be traded against one another.
Who: Those accredited to international fora, human rights personnel, cultural attachés.
- **DOS/DOD/Foreign Ministries/States:** Public information, de-escalation, cultural exchange, locally-sensitive project design, anti-extremism.
Who: Public Information Officers (PIO), liaison personnel.
- **NGOs, USAID:** Advocacy, anti-discrimination, gender/culture/religion-responsive planning, prediction of local areas of discontent with particular policies.
Who: Local field personnel, HQ planning personnel, USAID Innovation Lab.
- **FEMA, Emergency Responders:** Culture- and task-aware disaster response, bringing AI and deep knowledge management to local organizations.
Who: Any organization where having access to the right knowledge (lessons learned, chemical response models, etc.) at the right time can make a significant difference.
- **Oil companies:** Avoiding local conflict, planning project development in locally-sensitive ways.
Who: Those who negotiate with local communities, those at HQ responsible for overall peace and production continuation, project planners.

4. Current Status

All of the software described in this paper is currently functional and usable for real-world efforts. The underlying technology has been proven in various contexts ([7–10], and others). A significant military customer has expressed interest in using Atomic AI as its standard for commonsense knowledge, and Atomic principles are currently being employed in a US Government application which has received highly favorable feedback from evaluators.

Currently available Deep MindMaps include aspects of US, Chinese, and Iranian cultures, African conflict and peacekeeping models, Corporate Social Responsibility, HIV prevention, and more.

5. Packaging

Mindful of the real-world needs of field users, the entire cogSolv system can be run on a single laptop with no Internet connection, across the Web with minimal Internet speed requirements, or on a cell-phone/smartphone with or without Internet connection.

The system is easy-to-use and requires no technical background, making it useful to a wide range of responders, decisionmakers, planners, and conflict resolvers.

6. COGVIEW Deep MindMaps: Mapping Beliefs, Religions, Psychology

cogSolv relies on COGVIEW Deep MindMaps to understand people and the world in which they live. Deep MindMap diagrams describe important aspects of how others think and view the world. Simple to create and to understand, Deep MindMaps allow cogSolv to simulate the needs and selected aspects of the thought patterns of others. This in turn allows the computer to create counteroffers and persuasion strategies tailor-made for them, predict in useful part their likely reaction to certain actions, and assist users in ‘getting into the minds’ of others.

Deep MindMaps are able to represent nuanced information about local cultural and conflict resolution practices, including religious practices and viewpoints.

MindMaps provide a critical *knowledge multiplier* in that the information they contain is no longer locked inside the heads of experts - rather, it may be disseminated across the enterprise where it is able to influence the decision making processes of great numbers of personnel.

There are three general types of MindMaps:

- Cultural (Deep Worldview Models)
- Psychological (included with cogSolv)
- Conflict (deep goals and concerns)

Cultural/worldview models tell the computer how a specific group of people (as defined by the user) tends to see the world. Built by or in conjunction with informants, they help remove a significant source of inaccurate decision-making: *ethnocentrism*.

Psychological models provide cross-cultural insight into the human psyche, drawing on cognitive and social psychology. Users need not create such models, however, as a very complete set is provided as an integral part of the cogSolv suite.

In conflict contexts, conflict models provide a simple means of informing the system about the specific content of the conflict at hand.

6.1. How are MindMaps built?

Deep MindMaps are easy to create, and can be built in the field by those who have the best knowledge, written at HQ by experts, or created via some combination of the two. Once created, MindMaps are reusable and can be stored in common libraries.

Only minimal training is required in order to create MindMaps via a straightforward two-step process. In the first step, important concepts are identified in the domain of interest. In the second, those concepts are linked together in pairs. The structure of MindMaps makes it easy to test for correctness.

Because *humans can read and understand the exact same models that are presented to the computer*, there is no need to engage in time-consuming model translation between development and deployment stages.



Fig. 1. Sample Deep MindMap

6.2. How many MindMaps do I need?

For persuasion, cogSolv performs best with one Deep MindMap for each involved culture or subculture. For conflict resolution, one overall Conflict MindMap and at least one Cultural MindMap for each participant would be ideal. The system can work with less information, at the cost of reduced nuance.

MindMaps are meant to be reused across missions; it is envisioned that, for field use, prebuilt libraries of Deep MindMaps would be created at HQ in conjunction with informants and then made available for reuse in the field.

7. cogSolv/cogResolv: What Can They Understand?

cogSolv makes social factors such as religion, culture, values, and history much easier for outsiders to understand and take into account.

cogSolv's combined visualization, collaboration, and modeling capabilities allow interested parties to spatially comprehend the identities, psychological dynamics, and structural factors undergirding the complex relationships between disputants, stakeholders, and community and interest groupings, including:

- the in-depth nature of the relationships between parties, specifically focusing on psychological dimensions such as emotional connections, past history, past grievances, ethnic and clan concerns,
- social, economic, political, and power-related structure issues, including resource contestation, political access, and intergroup rivalries and power imbalances,
- general psychological principles, such as trauma that needs to be resolved, and community integration that may be required,
- the dynamical nature and potential relevance of community-based reconciliation methods (such as *mato-oput*), and
- general related historical circumstances and events.

Through clarity and nuanced simulation, cogSolv seeks to make the hidden explicit, increase decision quality, and model psychological factors such as pain and determination.

cogSolv can model the unobvious effects on complex systems of single changes, including the dynamic effects of changes and perturbations over time.

Essentially, cogSolv 'gets into the head' of participants, modeling subjective experience at a deep level. cogResolv *allows negotiators to discover which parts of the conflict 'space' are more fixed and thus less amenable to negotiation* and areas where there may be more room from the other parties' perspectives.

7.1. Peacekeeping

As alluded to above, in many ways peacekeeping is inherently constituted by signaling, especially so because peacekeepers often cannot resort to force to achieve their goals. This means that most actions troops take are calculated to send certain messages, using indirect methods calculated to have certain psychological effects. The system can model these.

Specifically, for local perspectives the system assists users in answering questions like those below:

1. 'Minimal understandings': Can we establish a minimal set of knowledge we must gain about local perspectives in order to properly design a peacekeeping mission? How should local culture modulate our peacekeeping actions?
2. Modulating emotions/fear/mistrust: how can we calibrate our messages to improve these factors?
3. How can we use local conditions to adjust the messages we send?
4. How can we maximize the legitimacy/correctness/appropriateness of our actions relative to cultural and local standards?
5. How do the 'peacekept' differentially perceive message form and content in different cultural/conflictual contexts?
6. What sorts of messages are sent through what actions?

For more detailed analysis on the use of Atomic AI in peacekeeping missions, see [11, 12].

8. Training and Situational Awareness

cogSolv and COGVIEW significantly enhance training and situational awareness capabilities.

Trainers can use cogSolv to quickly brief parties who have just entered the field of influence (consultants, military personnel, media, academics, and so on). Multiple-party access to a common picture enables new forms of teamwork and shared access to knowledge.

cogSolv allows trainers to include a greater totality of information not easily provided via other modalities, including relational and psychosocial factors, systems, structure, relationships and psychology. Deep MindMaps allow interested parties to visually arrange, drill-down and spatially understand the true nature of the situation at hand. Grievance details and possible 'angles' of resolution can be understood and simulated using spatial intelligences in addition to purely rationalistic or sequential methods.

8.1. Situational Awareness via Story Building

Varied research (cf. [13]) suggests that storytelling is an important part of how humans make sense of their world. Figure 2 demonstrates cogSolv's ability to automatically convert COGVIEW-based analysis into story form.

We are unhappy that you are engaging in Outsider Interference (-100),
 which is against our Religion ...
 One must not cause Fear (-100)
 One must not interfere with Honor (-100)
 Supporting Others supports Masculinity (1000), which is an important part of Tradition

Fig. 2. Output: Story Building

This functionality is useful when the story-based perspective is of interest and you wish to understand the other side via that lens, or when one wishes to understand the impact of particular goals on the other side from that side's perspective.

9. cogSolv Genies and Their Inputs: Easy To Create and Use

cogSolv is structured as a set of *Genies*, each of which solves a specific problem. Some Genies operate solely on COGVIEW Deep MindMaps, while others also accept simple inputs (of the form described in Section 10.1) describing a specific scenario for which they will be asked to perform a simulation.

For a current list of available *Genies*, please see <http://intmind.com/cogSolvGenies>.

10. Sample Humanitarian Applications and Genie Outputs

10.1. Brief Technical Introduction

In order to best understand the following demonstrations, it is important to understand two key cogSolv concepts: *energy/concept pairs* and *acceptance scores*.

Energy/concept pairs assign *energy values* to *concepts* (such as HAPPINESS or COMPUTER). Energy values are numbers and can be positive or negative. Positive energy values attached to a concept indicate that the attached concept is desirable, is present in some context, or is a goal that should be pursued. A negative energy value indicates concepts that are undesirable, not present, or should be avoided.

As an example, the energy/concept pair -150/FEAR could indicate that fear has been or *should be* lessened, or that fear creation should be avoided. Concepts are understood from the 'receiving perspective' - thus, the pair 100/DOMINANCE indicates that 100 units of dominance are being applied from the *outside* to the party whose perspective is being described.

When interpreting energy values, 100 is a 'typical amount', so -150/FEAR suggests that Fear has been or should be reduced 1.5 times 'a reasonably typical amount' that one might encounter in practical everyday life.

The second concept, *acceptance scores*, indicate how likely someone would be to accept or reject a particular proposition. Normally, scores range from -1 (absolute rejection) to 1 (absolute acceptance), but they can be much larger or smaller depending on simulation outcomes. As an example, one might assign the score +1 to the proposition OBTAIN FOOD AND SHELTER and -1 to the proposition EXPERIENCE STARVATION.

Understanding the examples In some of the examples given below, 'word clouds' are shown with concepts in **red** and **green** text. Red-colored text indicates concepts that have negative energy, and green-colored text the reverse. Words are sized in proportion to the energy they have received.

Depending on the Genie being used, green concepts often represent those that the user should attempt to augment. In the dissonance-induction context (Figure 3(b)), green concepts are those creating dissonances that are foreseeable but whose impact is likely to be misunderstood due to cultural factors. In this context, the color red denotes critical concepts that are *currently being ignored but should be more carefully considered* in order to create positive change.

10.2. Advocacy and Persuasion

cogSolv offers significant functionality for advocacy and persuasion. Related Genies help users employ deep knowledge about beliefs, cultures, and cognition during the persuasion process. cogSolv indicates exactly what to emphasize and how (and what to avoid) in order to maximize persuasive effectiveness from the other side's point of view. In line with *Social Justice Theory*[14], the system can also discover the specific 'anchor' concepts (see Figure 3(a) below) across which opinions are formed on specific issues.

cogSolv facilitates the use of Festinger's theory of *Cognitive Dissonance* to achieve belief change (see Directed Dissonance Reduction in [10]). As shown in Figure 10.2, cogResolv can calculate how and why dissonance is being created, supporting understanding, persuasion, and other related processes.

The running example below (starting with Figure 3) describes how Western governments could go about handling the recent wave of anti-LGBT sentiment in Africa. They could usefully include or exclude elements in their persuasive communications as suggested by the figures below, and could craft the core of their appeals taking these psychological simulation results into account.

As indicated in Figures 3, 3(a), and 3(b), cogSolv suggests an approach quite opposite to that currently in use, namely one focused on LOCAL DIGNITY, RELIGION, and TRADITION. cogSolv simulations suggest in part that differing versions of HAPPINESS, as well as concepts regarding POLITENESS, SOCIALITY, and SUFFERING (see Figure 3(a) for more in-depth analysis) are ultimately at issue.

Ultimately, indirect appeals are often the most powerful. During persuasion, *Potential Invoking Concepts* (PICs), shown in Figure 3, provide *alternate concepts* capable of evoking core concepts that the system recommends users include in their persuasive communications. PICs are drawn from the COGBASE commonsense knowledge database [7].

Recommended Persuasion Elements (<i>should use in persuasive communication</i>):	
RELIGION - Potential Invoking Concepts:	<i>pray, build cathedral, church, tell many person, god, temple, spirit, faith, islam, wicca, shamanism,...</i>
STRENGTH - Potential Invoking Concepts:	<i>go jog, fight war, stay healthy, rod, metal, bridge arch, power, hero, good part, might, vigor, sturdiness, brawn, sea power, concentration, invulnerability...</i>
LOCAL DIGNITY	
TRADITION - Potential Invoking Concepts:	<i>buy present others, propose woman, party, card, buy present, surprise, give gift, bake cake, convict suspect crime, ...</i>
SUPPORT OTHERS	
Disfavored Persuasion Elements (<i>should not include</i>):	
<i>Secularism, Linking of Development Assistance, Colonialism, Human Rights Discourse, General LGBT Perception, Outsider Interference, Sexuality</i>	

Fig. 3. How should we create persuasive appeals for the Africa LGBT scenario?



Sample text format data supporting the above (concept=energy, T denotes *target energy values*):
Happiness=10500/ T 1000, Core Emotions=-5900/ T 1000, Power=-5600, Local Cultures=3300, Respect=3300, Ideologies=3300, General LGBT Perception=-3300, Communitarianism=3300, Ego=3300, Tradition=3000, Morality=2600, Face=2500, Masculinity=2000, ... Honor=1000, Conflict=800, Offended=800/ T -1000, Local Dignity=700, Equality=-700, Christianity=-500, Religion=500, Christian Values=-500, anger=400, trauma=400...

10.3. Truly Just, Needs-Focused Conflict Resolution

As mentioned above, cogResolv focuses on resolving conflict in ways that are truly *just* in the sense that deep emotional and practical needs are met. cogResolv's access to the core needs of each party allows it to determine to what extent any particular resolution is actually just.

For conflict-driven contexts, cogResolv includes the following selected features:

10.3.1. Justice Score

In COGVIEW terms, a conflict may be considered to be *justly resolved* when 1) *target scores* (defined next) are maximized and 2) no significant *clashes* (see Figure 5) result. *Target scores*, defined as values attached to specific COGVIEW concepts (such as FAMILY, SAFETY, and BELONGING) indicate the core importance of certain concepts to a party's fundamental well-being. *Clashes*, in turn, indicate when a particular phenomenon violates fundamental, deeply-held values. The location of the clash within the Deep MindMap indicates the cause and nature of the incompatibility. cogSolv's Integrative Option Generator (Section 10.3.2 below) *inherently generates options leading to truly just results*.

Normal Justice Score values range from -1 to 1; values outside this range indicate particularly just or unjust resolutions.

10.3.2. Integrative Option Generator

When it is unclear how a conflict may be resolved in an integrative (highly equitable) manner, previous resolution attempts may have failed, and *new ideas are required*, this Genie is able to *find new ways of meeting old needs*. The system helps *separate issues* and *reframe conflicts*.

Each option in Figure 4 below can be interpreted as follows: a concept is given together with an associated energy. If the energy is positive, policy choices/actions that facilitate that concept should be chosen, and the reverse for negative. As suggested above, 100 units of energy is the 'normal' amount.

As an example, EQUALITY/700 suggests that strategists would do well to focus judiciously on that concept. LINKING OF DEVELOPMENT ASSISTANCE/-3000 suggests that strategies should not significantly invoke this concept, and may do well to explicitly disclaim it.

Options: Ameliorating 500 units/Colonialism (via relevant African perspective)
<p>Western-Country could undertake: Equality/700, Sociality/4300, Local Cultures/700, Linking of Development Assistance/-3000, Strength/1000, pleasure/1000, mad/-1000, anger/-1000, mean/-1000, trauma/-1000, hate/-1000, despise/-3400, scorn/-1000, embarrassment/-1000, Support Others/1000, empathy/1000, enjoy/1000, angry/-1000, Local Dignity/1000, unhappiness/-1000, joy/1000, like/1000, guilt/-400, regret/-400, remorse/-400, Outsider Interference/-3000, Religion/1000, Colonialism/-6000, happy/1000, Social Discomfort/-1000, Human Rights Discourse/-3000, care/1000, Love/1000, Dominance/-1000, Aggression/-1400, heartache/-1000, Support Others/1000, Psychological Drives/1000, Strength/1000, Religion/1000, Local Dignity/1000</p>

Fig. 4. Deep Win-Win (Integrative) Option Generator

10.4. Discover Concepts in Conflict (Find Conflict 'Essence')

This functionality, demonstrated in Figure 6, helps one understand the 'essence' of a particular conflict, explain the core of the conflict to others, and gain new perspectives on existing conflicts.

The Genie presents a list of core concepts that are most responsible for driving the conflict at hand. Red concepts are particularly problematic concepts (concepts that are not being properly addressed by the conflictants), and green concepts represent those that, if taken properly into account, could help push the conflict in the right direction.

10.4.1. Protracted Conflict

Untangling the complex issues leading to protracted conflict represents a very difficult task for humans. cogResolv can provide major support in that it is able to simultaneously 'compute all the angles' and point users towards the best solutions. cogResolv's Integrative Option Generator (Section 10.3.2) and Automated Negotiator Agent (next section) automatically generate non-obvious ways forward that simultaneously address all practical and psychological aspects of conflict and equitably maximize benefits for all sides.

10.5. Automated Negotiation

The ability to understand counterparts' worldviews, goals, needs, and so on, leads to the ability to automate and predict potential flows for entire negotiation processes.

cogResolv's *Automated Negotiator Agent* helps discover options that optimally maximize both sides' perceived value. The agent is able to automatically simulate opinions, needs, and goals on both sides of a conflict.

Sample Clashes:

Christian Values vs. Christianity, via:
Human Rights Discourse [-100],
Outsider Interference [-300],
Equality [700.0]

Communitarianism vs. Ideologies, via:
Colonialism [-100], Equality [700.0],
Christian Values [500.0],
Christianity [500.0],
Religion [1500.0], Local Cultures [700.0]

Empathy vs. Morality, via:
Colonialism [-100]

Face vs. Core Needs, via:
Equality [700.0], Christian Values [500.0],
Christianity [500.0], Religion [1500.0],
Local Cultures [700.0], Respect [700.0]

Fig. 5. Sample Clashes



Fig. 6. Concepts In Conflict (Conflict Essence)

At each round, the agent chooses options that have been determined to best meet the needs of the other side while avoiding overly negative costs for one's own side. Potential offers that would be insulting to or overly damaging to either side are automatically suppressed.

From Iran's perspective:

Proposal IRANIAN NUCLEAR WEAPONS/-300 receives desirability score -4.5658 (i.e. quite low)

Reasons: -3600/Security, -2700/Values, -2700/Power, -1880.0/Safety, 1600.0/Dominance, -1600/Country, -900/Control, -675/Equality, -600/Freedom, -600/Honor, -600/Respect ...

Agent chooses proposal Trade/132.3725, Diplomacy/65.0, Sanctions/100, score 1.3915.

Example 'Odious Proposal':

3000/Attack (i.e. allow other side to attack - even though this may offset other factors, US can't offer this as it too negatively affects its interests)

Fig. 7. Automated Negotiation Agent (Iranian Nuclear Weapons)

The system's ability to calculate the value of various offers allows it to *offer progressively more value* as negotiations continue.

As confirmed via human evaluation, the system's offers are remarkably human-like. In the case of cogResolv's simulation of the Iran/US conflict over nuclear weapons, cogResolv's recommendation was in fact nearly identical to a settlement which took place in 2013 (that is, some months after the initial simulation was run - see for example [15]).

11. Culture- and Knowledge-Aware Disaster Response

Experiences from the field clearly demonstrate the importance of cultural sensitivity to effective disaster response. [1, 16, 17]

It is critical to perform modeling in *both directions* - how responders should act in order to be viewed positively as well as the process by which viewpoints are generated on the survivor side. Bottom line: *if responders fail to cater to cultural needs, survivors won't trust you and may not evacuate or follow other directions.*

The benefits of Atomic AI for human decision making apply here as well; cogSolv's cogResponder component can manage detailed task and threat information and help responders triage and avoid emerging threats.

cogResponder simulates cultural perception both with respect to 1) responder actions and 2) Tweets and other social media data discussing the actions that responders take. Sentiment and task models are used to extract opinions being expressed. The latter capability allows cogResponder to automatically discover that messages about EXPLOSIONS affect human safety (including possibly EYES and HEARING).

cogResponder's deep culture and domain knowledge base allows it to provide scores for response activities across various *cultural and practical dimensions*, including Capability, Responsiveness, Correctness, Values Alignment, Solidarity, and Legitimacy.

Lastly, cogResponder enables responders to **master counterintuitive aspects of response**, including the

need to take specific actions for particular ethnic groups, which could include, for example, providing information through messages from friends and family instead of formal sources for Vietnamese communities. ***During response, intelligent actions build solidarity.***

11.1. Tweet/Social Media Processing for Disaster Response

cogResponder includes a powerful opinion mining engine capable of using deep semantics, Atomic AI, COGVIEW, and COGBASE [7], to *determine the real-world effects of events using commonsense knowledge* and, in turn, the pleasantness and emotional effects (including cultural and other perceptions) of raw social media textual content.

As an example, if an incoming tweet suggests that an explosion has taken place, the system understands that this is likely to cause pain and unhappiness, which will be viewed negatively *and will also reflect poorly on responders as they did not prevent this from occurring.*

In the sentence ‘I have no shoes’, the system’s knowledge enables it to understand that a shoe is an article of clothing, the lack of which affects the health of the individual, which in turn affects perception of response. The system contains significant knowledge about what *health* is and what affects it.

This knowledge also allows the system to determine that BOMB has semantics related to those of EXPLOSION, so *social media users can employ a wide range of vocabulary to describe the things they see.*

cogResponder can bring particular Tweets to responders’ attention based on the semantics described therein - ‘trapped’, family members in distress, and so on. **The cogSolv sentiment engine is the first to use deep semantics to this extent.**

Outputs include 1) trending topic and valence detection (i.e. ‘I love FEMA’ → positive sentiment towards FEMA; ‘Thankfully there was no explosion’ → negative energy into *explosion*, which provides positive sentiment for responders as well as the Tweet itself), and 2) semantic concept histories (*bomb* and *explosion* would trigger the same trending topics).

Finally, cogResponder can also discover trending locations so that hotspots may be quickly identified and resources diverted.

Input Sentence: ‘I got chemicals on me.’

Key Concept: Chemical

Computed Semantic Consequences and Dimensions Affected:

Explosion/600, High Temperature Explosion/400, Pain/200, Explosive Decomposition/200, Heat/200, Burn(Medical)/200, ... Fire/200, Oil/100, Combustibles/100, Eyes & Skin/-200.0

Cultural Dimensions:

Physical Effectiveness/-600, Personnel/-600, Physical Security/-600, Core Needs/-600, Responsiveness/-1200, Infrastructure/-1200, Health/-1700, Legitimacy/-1700, Correctness/-1800, Capability/-3500.0

Fig. 8. Deep-Knowledge Sentiment and Effect Mining

11.2. Task Models

In line with the AI functionalities put forth above, cogResolv is able to automatically comprehend response-related tasks, understand their implications, and prioritize subtasks. Commonsense knowledge acts here as a storehouse of lessons learned, providing detailed information about how to handle dangerous situations.

As an example, in a response where the chemical *chloropicrin* is involved, the system can use its knowledge of the profile and properties of this substance to indicate what tasks, in the current response context, workers should take in order to protect themselves. cogResponder can identify Personal Protective Equipment (PPE) that should be used, materials to be avoided, possible symptoms, and so on.

The goal is to use unobvious information and/or information that is likely to be overlooked in order to keep responders out of harm’s way, such as the fact that methyl bromide is often present where chloropicrin is and that sea ports often have both chemicals present, in order to help facilitate an ideal response. *The system provides real-time task prioritization based on the computed consequences of each choice and can adjust priorities automatically.*

12. Conclusion

It is all too easy to take the state of today’s world for granted - to assume that the limitations we have today as human beings can never be transcended and that we must therefore accept things as they are.

If, on the other hand, we could use the power of accurate, truly culturally-aware Artificial Intelligence to understand, reason about, and deeply dive into human data, it seems possible we could understand others more profoundly and in so doing find new solutions to problems that have eluded us in the past.

(a) General Task Modeling			(b) Consequence-based Reasoning		
Task Reasoning			Concept	Desire	Impact
Concept	Criticality	Import.	Concept	Desire	Impact
Search and Rescue	380	14.44	Vapors	-2400	141.94
Hospital Transport	300	9	Substance Chloropicrin	-4000	118.28
Deploy Food Aid	200	2	Vapor Cloud	-2400	70.97
Housing	100	1	Evacuation	1300	36.71
Health	100	1	Effluent	-1200	35.48
Trash	-100 (no trash)	1	Vapor Cloud Drift	-1200	35.48
Manage Donations	100	0	Disperse	1200	35.48
Damage Assessment	100	0	Effluent Disposal	1200	35.48
			Ventilate	1200	35.48
			Health Effects Methyl Bromide	-1400	28.89
			PPE	1300	28.25
			Substance Methyl Bromide	-1400	27.89
			Choose Incorrect PPE	-1200	25.15
			SAR	380	14.44
			Fire	-500	9.21
			Hospital Transport	300	9.00
			Burn(Medical)	-300	4.50
			Combustibles	-200	3.68
			Oil	-200	3.68
			Oxidizer	-200	3.68
			Paper	-200	3.68
			Wood	-200	3.68
			Explosion	-100	1
			High Temperature	-100	1
			Explosion	-100	1
			Lacrimation	-100	1
			Shock Explosion	-100	1
			Trash	-100	1
			Decontaminate	100	1
			Deploy Food Aid	100	1
			Health	100	1
			PPE Bags	100	1
			PPE Disposal	100	1
			Skin	100	1
			Housing	100	1

Fig. 9. Sample Disaster Response Output

Bringing the wisdom of experts to those who would normally not have access to it, avoiding humanitarian mission failure, improving disaster response, and reducing violent, intractable, and latent conflict alike - this is cogSolv's vision. Future work will involve enhanced field outreach and further development of *Genies* for specific users' needs and particular conflict theories.

As adoption grows, it is hoped that cogSolv's impact will as well.

Acknowledgements

The author gratefully acknowledges 1) the anonymous reviewer for the care and thought evidenced in the review comments, and 2) Craig Zelizer's invaluable contributions regarding interface and operation.

References

- [1] J. Seidenberg, Cultural competency in disaster recovery: Lessons learned from the Hurricane Katrina experience for better serving marginalized communities, Tech. rep., University of California, Berkeley Law School (2005).
- [2] T. Duffey, Cultural issues in contemporary peacekeeping, *International Peacekeeping* 7 (1) (2000) 142–168.
- [3] L. M. Howard, *UN Peacekeeping in Civil Wars*, Cambridge University Press, 2007.
- [4] D. Olsher, COGVUE & INTELNET: Nuanced energy-based knowledge representation and integrated cognitive-conceptual framework for realistic culture, values, and concept-affected systems simulation, in: *Proceedings, IEEE Symposium Series on Computational Intelligence*, 2013, pp. 82–91.
- [5] T. I. Ören, *Discrete Event Modeling and Simulation: A Tapestry of Systems and AI-based Theories and Methodologies*, Springer-Verlag, 2001, Ch. Towards a Modelling Formalism for Conflict Management and for Sociocybernetics", pp. 93–106.
- [6] C. D. Parks, *Graduated Reciprocation in Tension Reduction (GRIT)*, SAGE Publications, Inc., 2010, pp. 310–312.
- [7] D. Olsher, *Semantically-Based Priors and Nuanced Knowledge Core For Big Data, Social AI, and Language Understanding*, *Elsevier Neural Networks* 58 (2014) 131–147.
- [8] D. Rajagopal, E. Cambria, D. Olsher, K. Kwok, A graph-based approach to commonsense concept extraction and semantic similarity detection, in: *Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee*, 2013, pp. 565–570.
- [9] D. Olsher, H. G. Toh, Novel Methods for Energy-Based Cultural Modeling and Simulation: Why Eight Is Great in Chinese Culture, in: *Proceedings, IEEE Symposium Series on Computational Intelligence*, 2013, pp. 74 – 81.
- [10] D. Olsher, Changing Discriminatory Norms Using Models of Conceptually-Mediated Cognition and Cultural Worldviews, in: *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 2012, pp. 2138–2143.
- [11] D. Olsher, Cognitive/AI Peacekeeping Decision Support Models, in: *Proceedings, IEEE Global Humanitarian Technology Conference*, 2013, pp. 122 – 127.
- [12] D. J. Olsher, Cognitive-cultural simulation of local and host government perceptions in international emergencies, in: *2013 IEEE Global Humanitarian Technology Conference, GHTC 2013, San Jose, CA, USA, October 20-23, 2013*, 2013, pp. 112–117.
- [13] D. Herman, Storytelling and the sciences of mind: Cognitive narratology, discursive psychology, and narratives in face-to-face interaction, *Narrative* 15 (3).
- [14] M. Sherif, *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*, Yale University Press, 1981.
- [15] M. R. Gordon, Accord Reached With Iran to Halt Nuclear Program, *New York Times* (November 2013). URL <http://www.nytimes.com/2013/11/24/world/middleeast/talks-with-iran-on-nuclear-deal-hang-in-balance.html?pagewanted=all&r=0>
- [16] US Dept. of Health and Human Services, Office Of Minority Health, Cultural competency in disaster response: A review of current concepts, policies, and practices, Tech. rep., HHS (2008).
- [17] Federal Emergency Management Agency (FEMA), FEMA Think Tank May 2013, Tech. rep., FEMA (2013). URL <http://www.fema.gov/media-library-data/20130919-1822-27928-6806/transcripts20130919-27928-b21oy6.txt>



**BERKMAN
KLEIN CENTER**
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2018-6
September 25, 2018

**Artificial Intelligence & Human Rights:
Opportunities & Risks**

Filippo A. Raso
Hannah Hilligoss
Vivek Krishnamurthy
Christopher Bavitz
Levin Kim

This paper can be downloaded without charge at:

The Berkman Klein Center for Internet & Society Research Publication Series:
<https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights>

The Social Science Research Network Electronic Paper Collection:
<https://ssrn.com/abstract=3259344>

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • <http://cyber.law.harvard.edu/> •
cyber@law.harvard.edu



**BERKMAN
KLEIN CENTER**
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY



Artificial Intelligence & Human Rights:

OPPORTUNITIES & RISKS

September 25, 2018

Filippo Raso
Hannah Hilligoss
Vivek Krishnamurthy
Christopher Bavitz
Levin Kim



Table of Contents

Background & Context	3
Summary of Findings	4
1 Introduction	7
2 What is Artificial Intelligence?	10
3 What are Human Rights?	12
4 Identifying the Human Rights Consequences of AI	14
5 AI's Multifaceted Human Rights Impacts	17
5.1 Criminal Justice: Risk Assessments	20
5.2 Access to the Financial System: Credit Scores	26
5.3 Healthcare: Diagnostics	32
5.4 Online Content Moderation: Standards Enforcement	37
5.5 Human Resources: Recruitment and Hiring	42
5.6 Education: Essay Scoring	47
6 Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach	52
7 Conclusion	58
8 Further Reading	59

Background & Context¹

This report explores the human rights impacts of artificial intelligence (“AI”) technologies. It highlights the risks that AI, algorithms, machine learning, and related technologies may pose to human rights, while also recognizing the opportunities these technologies present to enhance the enjoyment of the rights enshrined in the Universal Declaration of Human Rights (“UDHR”). The report draws heavily on the United Nations Guiding Principles on Business and Human Rights (“Guiding Principles”) to propose a framework for identifying, mitigating, and remedying the human rights risks posed by AI.

Readers wishing to better understand the often-paradoxical human rights impacts of the six current AI applications that are detailed in this report are invited to explore a series of interactive visualizations that are available at ai-hr.cyber.harvard.edu.

¹ We would like to express our appreciation to our Berkman Klein colleagues Amar Ashar, Ryan Budish, Daniel Dennis Jones, Rob Faris, Jessica Fjeld, Sarah Newman and Casey Tilton for their helpful suggestions throughout the course of this project; to our interns and research assistants Sam Bookman, Daniel Chase, Christina Chen, Areeba Jibril, Adam Nagy, and Marianne Strassle for their assistance in finalizing this report; and to Urs Gasser and Jonathan Zittrain, respectively the Executive Director and Faculty Chair of the Berkman Klein Center, for their visionary leadership of our Center and of the Ethics and Governance of Artificial Intelligence Fund.

We are grateful to Kim Albrecht, Solon Barocas, Dinah PoKempner, Mark Latonero, An Xiao Mina, Brian Root, Maria Sapignoli, Seamus Tuohy, and Andrew Zick for consulting with us at various stages of this project and helping us refine our analytical methodology.

Finally, we wish to thank the Government of Canada for its sponsorship of this project, and in particular thank Tara Denham, Salahuddin Rafiqhuddin, Philippe-André Rodriguez, Marketa Geislerova, Maroussia Levesque, Asha Mohidin Siad and Jennifer Jeppsson of Global Affairs Canada for their consistent support of this project.

The views expressed in this report are those of the authors alone and do not reflect those of the Government of Canada or of the Berkman Klein Center for Internet & Society at Harvard University.

Summary of Findings

A Human Rights-based Approach to AI's Impacts

The ongoing dialogue regarding the ethics of artificial intelligence (AI) should expand to consider the human rights implications of these technologies.

International human rights law provides a universally accepted framework for considering, evaluating, and ultimately redressing the impacts of artificial intelligence on individuals and society.

Since businesses are at the forefront of developing and implementing AI, the United Nations Guiding Principles on Business and Human Rights are especially salient in ensuring that AI is deployed in a rights-respecting manner.

Determining Impacts

We propose that the best way to understand the impact of AI on human rights is by examining the difference, both positive and negative, that the introduction of AI into a given social institution makes to its human rights impacts. We take this view for two reasons:

1. Determining the human rights impacts of AI is no easy feat, for these technologies are being introduced and incorporated into existing social institutions, which are not rights-neutral.
2. Each application of AI impacts a multitude of rights in complicated and, occasionally, contradictory ways. Exploring these relationships within use cases allows for more nuanced analysis.

Measuring Impacts

Current implementations of AI impact the full range of human rights guaranteed by international human rights instruments, including civil and political rights, as well as economic, cultural, and social rights.

Privacy is the single right that is most impacted by current implementations of AI. Other rights that are also significantly impacted by current AI implementations include the rights to equality, free expression, association, assembly, and work. Regrettably, the impact of AI on these rights has been more negative than positive to date.

The positive and negative impacts of AI on human rights are not distributed equally throughout society. Some individuals and groups are affected more strongly than others, whether negatively or positively. And at times, certain AI implementations can positively impact the enjoyment of a human right by some while adversely impacting it for others.

Addressing Impacts

Addressing the human rights impacts of AI is challenging because these systems can be accurate and unfair at the same time. Accurate data can embed deep-seated injustices that, when fed into AI systems, produce unfair results. This problem can only be addressed through the conscious efforts of AI systems designers, end users, and ultimately of governments, too.

Many of the existing formal and informal institutions that govern various fields of social endeavor are ill-suited to addressing the challenges posed by AI. Institutional innovation is needed to ensure the appropriate governance of these technologies and to provide accountability for their inevitable adverse effects.

The Path Forward

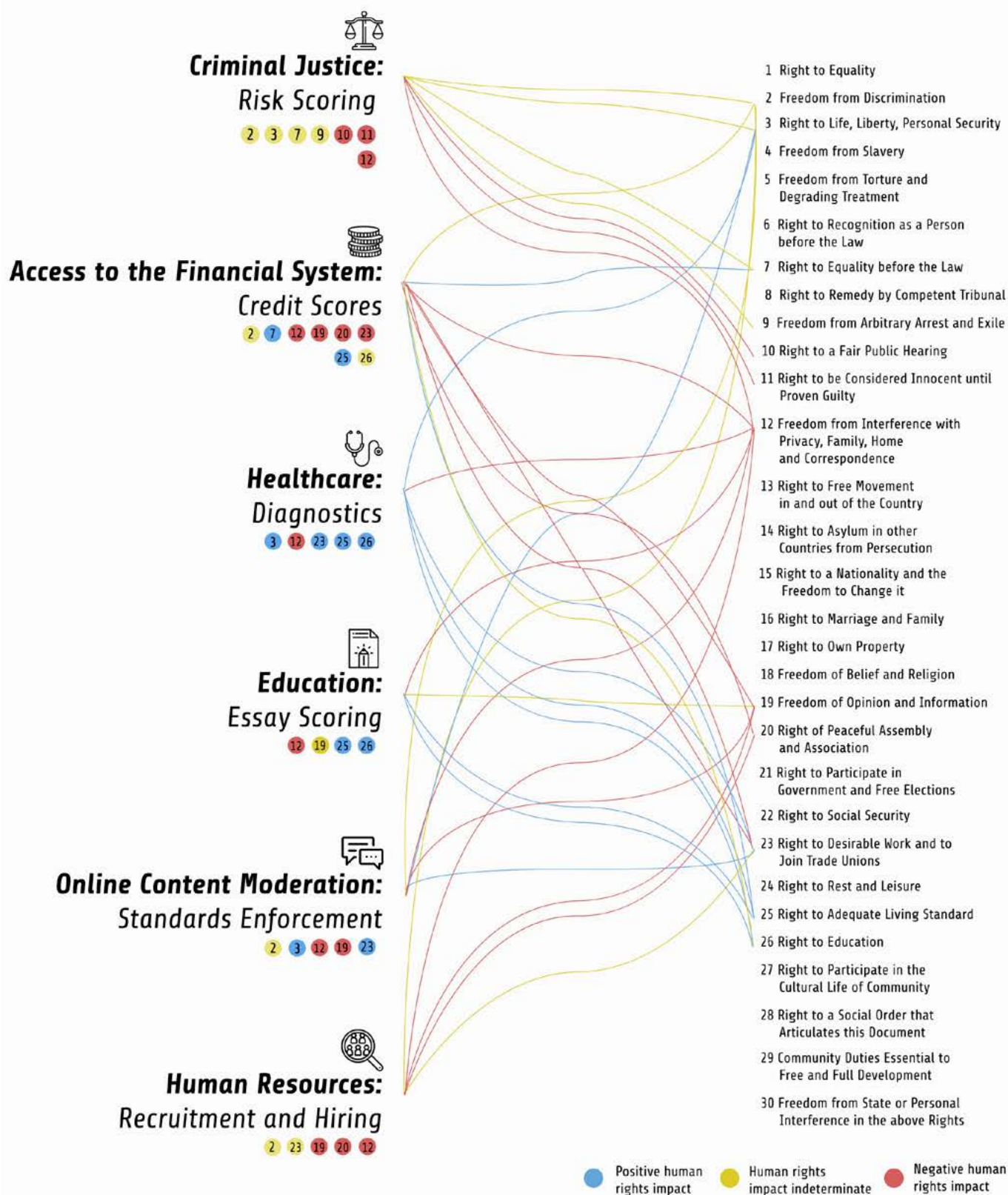
Human rights due diligence by businesses can help avoid many of the adverse human rights impacts of AI.

Non-state grievance and remedy mechanisms can provide effective redress for some, but by no means all, of the inevitable adverse impacts that AI will produce.

Governments have an important role to play in creating effective mechanisms to remedy the adverse human rights impacts of AI.

The role of government is essential to addressing the distributive consequences of AI by means of the democratic process.

ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS



1. Introduction

Artificial intelligence (“AI”) is changing the world before our eyes. Once the province of science fiction, we now carry systems powered by AI in our pockets and wear them on our wrists. Vehicles on the market can now drive themselves, diagnostic systems determine what is ailing us, and risk assessment algorithms increasingly decide whether we are jailed or set free after being charged with a crime.

The promise of AI to improve our lives is enormous. AI-based systems are already outperforming medical specialists in diagnosing certain diseases, while the use of AI in the financial system is expanding access to credit to borrowers that were once passed by. Automated hiring systems promise to evaluate job candidates on the basis of their bona fide qualifications, rather than on qualities such as age or appearance that often lead human decision-makers astray. AI promises to allow institutions to do more while spending less, with concomitant benefits for the availability and accessibility of all kinds of services.

Yet AI also has downsides that dampen its considerable promise. Foremost among these is that AI systems depend on the generation, collection, storage, analysis, and use of vast quantities of data—with corresponding impacts on the right to privacy. AI techniques can be used to discover some of our most intimate secrets by drawing profound correlations out of seemingly innocuous bits of data.

AI can easily perpetuate existing patterns of bias and discrimination, since the most common way to deploy these systems is to “train” them to replicate the outcomes achieved by human decision-makers. What is worse, the “veneer of objectivity” around high-tech systems in general can obscure the fact that they produce results that are no better, and sometimes much worse, than those hewn from the “crooked timber of humanity.”

These dystopian possibilities have given rise to a chorus of voices calling for the need for Fairness, Accountability, and Transparency in Machine Learning (“FAT” or “FAT/ML”). Advocates of this approach view the response to AI’s potential problems in terms of ethics. For example, the Institute of Electrical and Electronics Engineers—the world’s largest technical professional body that plays an important role in setting technology standards—has published an influential treatise on *Ethically Aligned Design* that suggests that “the full benefit of these technologies will be attained only if they are aligned with our defined values and ethical principles.”² In a similar vein, the governments of France and India have recently released discussion papers to frame their national strategies on AI that embrace an ethics-based approach to addressing the social impacts of these technologies.³

During the pendency of this project, however, several influential actors have come to recognize the

² IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, “Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.” Version 2. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

³ For France’s strategy, see: Cédric Villani, “For a Meaningful Artificial Intelligence: Towards a French and European Strategy” (AI For Humanity), accessed June 22, 2018, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. For India’s, see Amitabh Kant, “National Strategy for Artificial Intelligence” (NITI Aayog, June 2018), www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. Although both papers are substantial works that are each over 100 pages long, they barely mention the concept of human rights.

value of examining the challenges around AI from a human rights perspective.⁴ This incipient conversation on AI and human rights has already produced two significant documents. One is the Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems (“Toronto Declaration”), which was opened for signatures on May 16, 2018.⁵ As its full title suggests, the Toronto Declaration highlights the potential adverse effects of machine learning on rights to equality and non-discrimination and calls for the development of effective remedial mechanisms for all those who are adversely affected by these systems.⁶ The other is Global Affairs Canada’s Draft Strategy Paper on the Human Rights and Foreign Policy Implications of AI, which examines how AI can impact the rights to equality, privacy, free expression, association, and assembly, and suggests ways that these impacts can be redressed.⁷

This project is rooted in the belief that there is considerable value in adopting a human rights perspective to evaluating and addressing the complex impacts of AI on society. The value lies in the ability of human rights to provide an agreed set of norms for assessing and addressing the impacts of the many applications of this technology, while also providing

a shared language and global infrastructure around which different stakeholders can engage.⁸

While there are many different conceptions of human rights, from the philosophical to the moral, we in this project take a legal approach. We view human rights in terms of the binding legal commitments the international community has articulated in the three landmark instruments that make up the International Bill of Rights.⁹ This body of law has developed over time with the ratification of new treaties, the publication of General Comments that authoritatively interpret the provisions of these treaties, and through the work of international and domestic courts and tribunals, which have applied the provisions of these treaties to specific cases.

Our project seeks to advance the burgeoning conversation on AI and human rights by mapping the human rights impacts of the current deployment of AI systems in six different fields of endeavor. We strive to move beyond the predominant focus on AI’s impact on select civil and political rights, to consider how these technologies are impacting other rights guaranteed by international law—especially economic, social, and cultural rights.

4 For example, Amnesty International launched a structured initiative on Artificial Intelligence and Human Rights in 2017, while the New York-based Data & Society Research Institute hosted a workshop on Artificial Intelligence and Human Rights in April, 2018. See Sherif Elsayed-Ali, “Artificial Intelligence and the Future of Human Rights,” Oct. 19, 2017, <https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5>. Mark Latonero, “Artificial Intelligence & Human Rights: A Workshop at Data & Society,” May 11, 2018, <https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149>.

5 Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems, May 16, 2018. <https://www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-DoV2.pdf>.

6 Ibid.

7 Digital Inclusion Lab, Global Affairs Canada, “Artificial Intelligence: Human Rights & Foreign Policy Implications.” Accessed June 1, 2018. https://docs.google.com/document/d/1fhlJYznWSI7oD3TVJ5CgLGHJMj2HouEZiQ9a_qKbLGo/edit (“GAC Strategy Paper”).

8 Jason Pielemeier, “The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence,” Data & Society: Points, June 6, 2018, <https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a>.

9 The “International Bill of Rights” is a term to describe the three most important international human rights instruments, namely the Universal Declaration of Human Rights (“UDHR”), the International Covenant on Civil and Political Rights (“ICCPR”), and the International Covenant on Economic, Social, and Cultural Rights (“ICESCR”).

In so doing, we suggest what we believe to be the optimal method for identifying the human rights impacts of introducing a particular AI system into a given field of endeavor. Simply put, we believe it is important to recognize that AI systems are not being deployed against a blank slate, but rather against the backdrop of social conditions that have complex pre-existing human rights impacts of their own. This may well appear to be a self-evident truth, but in our view, the existing literature does not adequately consider the impact of these background conditions on the consequences of introducing AI. As a result, human rights impacts, both positive and negative, may be misattributed to AI, contributing to the extreme claims of optimists and pessimists alike about the extent to which AI is changing our lives.

Our report and the accompanying visualizations make clear that AI is already impacting the enjoyment of the full range of human rights—sometimes in paradoxical ways. In the final section, we examine and evaluate how international human rights law generally, and the growing field of business and human rights specifically, can help the developers, users, and regulators of AI systems to address many of these impacts.

2. What is Artificial Intelligence?

Despite its expanding presence across many aspects of our lives, there is no widely accepted definition of “artificial intelligence.”¹⁰ Instead, it is an umbrella term that includes a variety of computational techniques and associated processes dedicated to improving the ability of machines to do things requiring intelligence, such as pattern recognition, computer vision, and language processing.¹¹ With such a loose conceptualization and given the rapid growth of technology, it is no surprise that what is considered artificial intelligence changes over time. This is known as the “AI effect” or the “odd paradox”: formerly cutting-edge innovations become mundane and routine, losing the privilege of being categorized as AI, while new technologies with more impressive capabilities are labeled as AI instead.¹²

The impossibly large set of technologies, techniques, and applications that fall under the AI umbrella can be usefully classified into two buckets. The first is comprised of *knowledge-based systems*, which are “committed to the notion of generating

behavior by means of deduction from a set of axioms.”¹³ These include “expert systems” which use formal logic and coded rules to engage in reasoning. Such systems, which are sometimes also called “closed-rule algorithms,” include everything from commercial tax preparation software to the first generation of healthcare diagnostic decision support algorithms. These systems are good at taking concrete situations and reasoning optimal decisions based on defined rules within a specific domain. They cannot, however, learn or automatically leverage the information they have accumulated over time to improve the quality of their decision-making (unless they are paired up with some of the techniques described below).¹⁴

The second bucket of technologies uses statistical learning to continuously improve their decision-making performance. This new wave of technology, which encompasses the widely-discussed techniques known as “machine learning” and “deep learning,” has been made possible by the exponential growth of computer processing power, the

¹⁰ National Science and Technology Council: Committee on Technology, “Preparing for the Future of Artificial Intelligence,” Government Report (Washington, D.C.: Executive Office of the President, October 2016).

¹¹ One seminal textbook categorizes AI into (1) systems that think like humans (e.g., cognitive architectures and neural networks); (2) systems that act like humans (e.g., pass the Turing test, knowledge representation, automated reasoning, and learning); (3) systems that think rationally (e.g., logic solvers, inference, and optimization); and (4) systems that act rationally (e.g., intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting), Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence (Englewood Cliffs, NJ: Prentice Hall, 1995).

¹² Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).

¹³ Nello Cristianini, “On the Current Paradigm in Artificial Intelligence,” *AI Communications* 27, no. 1 (January 1, 2014): 37–43, <https://doi.org/10.3233/AIC-130582>.

¹⁴ Bruce G. Buchanan, “Can Machine Learning Offer Anything to Expert Systems?,” *Machine Learning* 4, no. 3–4 (December 1, 1989): 251–54, <https://doi.org/10.1023/A:1022646520981>.

massive decline in the cost of digital storage, and the resulting acceleration of data collection efforts.¹⁵ Systems in this category include self-driving vehicles, facial recognition systems used in policing, natural language processing techniques that are used to automate translation and content moderation, and even algorithms that tell you what to watch next on video streaming services. While these systems are impressive in their aggregate capacities, they are probabilistic and can thus be unreliable at the individual level. For example, deep learning computer vision systems can classify an image almost as accurately as a human; however, they will occasionally make mistakes that no human would make—such as mistaking a photo of a turtle for a gun.¹⁶ They are also susceptible to being misled by “adversarial examples,” which are inputs that are tampered with in a way that leads an algorithm to output an incorrect answer with high confidence.¹⁷

In this report, we focus on AI systems from both of these conceptual buckets that “perceive[] and act[]”¹⁸ upon the external environment by “tak[ing] the best possible action in a situation.”¹⁹ Simply put, the scope of our report is limited to analyzing those AI systems that automate the making of decisions that were formerly the exclusive province of human intelligence. This view of AI embraces everything

from medical diagnostic software that determine what is ailing a patient based on the available evidence, to self-driving vehicles that “decide” whether to steer, accelerate, or brake, millisecond by millisecond. The crucial factor for us is that the system must function and impact the external environment, rather than simply be a theoretical construct that remains under development, to be considered within the scope of our report. Furthermore, we limit our scope to AI technologies that are either currently in use or are far along in the development process; therefore, we do not delve into the realm of artificial general intelligence.²⁰ We restrict our consideration of AI in this report to those technologies that are being used to make decisions with real-world consequences for the simple reason that these are the technologies that are most likely to have discernible human rights impacts. By contrast, many other strains of AI research remain conceptual for now, and are thus yet to impact human rights.

¹⁵ Gheorghe Tecuci, “Artificial Intelligence,” *Wiley Interdisciplinary Reviews: Computational Statistics* 4, no. 2 (2012): 168–80, <https://doi.org/10.1002/wics.200>.

¹⁶ Adam Conner-Simons, “Fooling Neural Networks w/3D-Printed Objects,” MIT Computer Science & Artificial Intelligence Lab (blog), November 2, 2017, <https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects>.

¹⁷ Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and Harnessing Adversarial Examples,” *ArXiv:1412.6572 [Cs, Stat]*, December 19, 2014, <http://arxiv.org/abs/1412.6572>; A Nguyen, J Yosinski, and J Clune, “Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” *CVPR, IEEE*, 15 (2015)

¹⁸ Russell and Norvig, *Artificial Intelligence*, 7.

¹⁹ *Ibid.*, 27.

²⁰ Broadly speaking, an AGI system is one that can perform any task as well as a human can, or a “synthetic intelligence that has a general scope and is good at generalization across various goals and contexts,” Ben Goertzel, “Artificial General Intelligence: Concept, State of the Art, and Future Prospects,” *Journal of Artificial General Intelligence* 5, no. 1 (December 1, 2014): 1–48, <https://doi.org/10.2478/jagi-2014-0001>.

3. What are Human Rights?

As noted in the introduction, in this report we adopt a legal conception of human rights. We use the term human rights to refer to those individual and collective rights that have been enshrined first and foremost in the Universal Declaration of Human Rights (“UDHR”), and then further detailed in the International Covenant on Civil and Political Rights (“ICCPR”) and the International Covenant on Economic, Social and Cultural Rights (“ICESCR”).

The UDHR is the leading statement of the rights that every human being enjoys by virtue of their birth. Although the UDHR was adopted by means of a non-binding U.N. General Assembly resolution,²¹ Canada and many other states have long believed that there is an “obligation on states to observe the human rights and fundamental freedoms enunciated in the [UDHR] [that] derives from their adherence to the Charter of the United Nations,” which is binding international law.²²

The ICCPR and the ICESCR, meanwhile, are international treaties that are binding upon those states that have ratified them. These treaties elaborate upon the human rights that were first articulated by the UDHR at the international level, and clarify the duties of states in relation to two categories of rights. Whereas the ICCPR’s protections of civil and

political rights come into force immediately upon ratification,²³ the ICESCR instead requires states to take measures to progressively realize the economic, social, and cultural rights it protects, having due regard for the state’s economic condition and resources.²⁴

States shoulder a binding obligation under international law to protect human rights. This includes a duty to respect human rights in their own conduct, and to prevent natural and juridical persons subject to their jurisdiction (including corporations) from committing human rights abuses. These obligations persist even when privatizing the delivery of services that may impact human rights.²⁵

Especially since the end of the Cold War, businesses have come to be viewed as having their own responsibilities under international law to respect human rights.²⁶ The nature and scope of these responsibilities have been articulated most authoritatively in the United Nations Guiding Principles on Business and Human Rights (“UNGP” or “Guiding Principles”). Specifically, the responsibility to respect human rights requires enterprises to avoid causing or contributing to adverse human rights impacts through their own activities, and to seek to prevent or mitigate such impacts when the enterprise is

21 Universal Declaration of Human Rights (10 Dec. 1948), U.N.G.A. Res. 217 A (III) (1948) [hereinafter “UDHR”].

22 Letter from the Legal Bureau, Jan. 9, 1979, reprinted in *Canadian Practice in International Law*, 1980 Can. Y.B. Int’l L. 326.

23 International Covenant on Civil and Political Rights (New York, 16 Dec. 1966) 999 U.N.T.S. 171 and 1057 U.N.T.S. 407, entered into force 23 Mar. 1976, art. 2 [hereinafter “ICCPR”].

24 International Covenant on Economic, Social and Cultural Rights (New York, 16 Dec. 1966) 993 U.N.T.S. 3, entered into force 3 Jan. 1976, art. 2(1) [hereinafter “ICESCR”].

25 Human Rights Council Res. 17/4, Rep. of the Hum. Rts. Council, 17th Sess., June 16, 2011, U.N. Doc. A/HRC/RES/17/4 (July 6, 2011); Special Rep. of the Sec’y Gen., Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework, Hum. Rts. Council, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter “Guiding Principles”], Principle 5.

26 Guiding Principles, Principle 11.

“directly linked” to them via a business relationship.²⁷ This, in turn, requires enterprises to engage in ongoing due diligence processes to identify, prevent, and mitigate salient human rights risks.²⁸ To the extent that adverse human rights impacts do occur, businesses should provide remediation for those impacts through legitimate mechanisms²⁹—although it is emphatically the duty of the state to provide effective remedies through judicial and other mechanisms to those who have suffered business-related human rights abuses.³⁰

Although the Guiding Principles do not themselves have the force of law, they clarify how pre-existing international human rights standards apply to business activities, and provide useful guidance on how businesses can operate in a rights-respecting manner.³¹ In any event, since businesses are at the forefront of developing and deploying AI, the Guiding Principles are of immense importance to ensuring that the human rights impacts of these powerful new technologies are positive. Consequently, the Guiding Principles will feature prominently in the discussion that follows of the human rights impacts of AI systems that are currently in use, and in our suggestions regarding how they should be addressed.

²⁷ Ibid., Principle 13.

²⁸ Ibid., Principle 17.

²⁹ Ibid., Principle 22.

³⁰ Ibid., Principle 25.

³¹ Justine Nolan, “The Corporate Responsibility to Respect Human Rights: Soft Law or Not Law?,” in *Human Rights Obligations of Business*, ed. Surya Deva and David Bilchitz (Cambridge: Cambridge University Press, 2013), 138–61, <https://doi.org/10.1017/CBO9781139568333.010>.

4. Identifying the Human Rights Consequences of AI

AI is not being developed in a vacuum or deployed against a blank slate. Rather, specific actors in society are deploying AI to automate decision-making in particular fields of endeavor. They are doing so to achieve outcomes that they view as desirable, against the backdrop of social institutions that have their own, pre-existing human rights implications.

Consider, for example, the deployment of AI in the criminal justice system, which is discussed in more detail in the first case study below. Over the course of the last several hundred years, criminal defendants have been endowed with various rights to ensure the fairness of criminal proceedings. These include the presumption of innocence,³² the principle of legality,³³ the right to a fair trial,³⁴ and many others. Even so, no existing criminal justice system comes close to perfectly respecting the rights of defendants and other relevant rights-holders: every such system has at least some negative impacts on rights-holders that predate the introduction of AI.³⁵

It is only by embracing a comparative approach, that accounts for background conditions from the pre-AI world, that we can properly understand the human rights impacts of introducing AI into the criminal justice system or any other human institution. Unless the human rights implications, both positive and negative, of pre-existing institutional structures are identified and accounted for, the human rights impacts of introducing AI will be

conflated with the ongoing impacts of whatever was there before. Below, we propose a two-step methodology for avoiding such difficulties.

Step 1: Establish the Baseline

As noted, the first step is to simply consider the existing human rights implications, both positive and negative, of whatever field of endeavor AI is being introduced into. This evaluation properly involves consideration of the availability and effectiveness of institutional mechanisms that are currently in place to regulate and redress the negative human rights implications arising from that field. When human decision-making in the field in question has already been supplanted by a first-generation automated decision-making technology, such as a closed-rule diagnostic algorithm, the first step consists of evaluating the human rights implications of the pre-AI status quo.

Step 2: Identify the Impacts of AI

The second step involves identifying how the introduction of AI changes the human rights impacts of the field into which the technology is introduced. If the introduction of AI improves the human rights performance of the field, AI can be said to have a positive impact on human rights. That is true even if the field of endeavor continues to produce adverse human rights impacts after the introduction

³² UDHR art. 11.

³³ UDHR art. 11(2).

³⁴ ICCPR art. 14(1).

³⁵ For the last two years, the Berkman Klein Center has been conducting extensive research on the use of algorithms in the criminal justice system, in its capacity as one of the two anchor institutions for the Ethics and Governance of Artificial Intelligence initiative. The research outputs of this ongoing work can be found at <https://cyber.harvard.edu/research/ai>.

of AI. Conversely, if the human rights performance of the field of endeavor deteriorates with the introduction of AI, then it is clear that the technology has produced adverse human rights impacts. To a significant extent, the outcome of this evaluation will depend on whether the mechanisms currently in place to regulate and remedy the adverse human rights consequences of the field in question continue to be effective following the introduction of AI.

The human rights impacts of AI stem from at least three sources, two of which can be considered by conducting a human rights impact assessment before a particular system is deployed. The third source, meanwhile, can be hard to identify even after an AI system is in operation, due to the complexity of the technology:

1. *Quality of Training Data:* To the extent that the data used to “train” an AI system is biased, the resulting system will reflect, or perhaps even exacerbate, those biases.³⁶ This is a version of what is known as the “garbage in, garbage out” problem, and it can have profound consequences for a wide variety of human rights—depending on what the system is intended to do.
2. *System Design:* Decisions made by an AI system’s human designers can have significant human rights consequences. Human designers can, for example, prioritize the variables they would like the AI system to optimize and decide what variables the AI should take into consideration as it operates. Such design decisions can have both positive and negative human rights impacts, which will be informed by the individual life experiences and biases of the designers.

Some of these impacts will be foreseeable, while others will not be.

3. *Complex Interactions:* Once an AI system is introduced, it will interact with the environment in ways that produce outcomes that might not have been foreseen. These complex interactions can have significant human rights impacts. In some cases, the impacts of these interactions may be detectable through the use of certain analytical techniques, but the possibility exists that certain human rights impacts resulting from the deployment of an AI system will escape detection. This is not an issue that is unique to AI: pre-digital societies are staggeringly complex, and the human rights impacts of the actions of individuals and institutions are not always knowable at the time they are made or for some time thereafter.

Limitations of our approach

Our two-step methodology provides a useful, generalizable approach to identifying the positive and negative implications of introducing AI into an extant field of endeavor. This methodology, which we have validated in consultations with stakeholders from the technology and human rights communities, undergirds our assessment of the human rights impacts of AI across six different use cases below.

Our framework has its limitations, especially due to the scarcity of available information into the design and operation of any given AI system. This is due in part to the novelty of AI, but also because so much AI technology is proprietary, which results in information about the design, operation, and impact of

³⁶ Osonde Osoba & William Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. (Santa Monica: Rand Corporation, 2017). https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

the systems being treated by their creators as commercially sensitive information.³⁷

Consequently, the analysis we undertake in our six case studies, below, is at the level of detail that one would find in a sectoral human rights impact assessment. Based on our desktop research, we have drawn reasonable inferences as to the likely human rights impacts of introducing particular AI systems into the prevailing social and institutional context.

³⁷ Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," *Stanford Law Review* 70, no. 5 (2017): 1343–1429, <https://doi.org/10.2139/ssrn.2920883>.

5. AI's Multifaceted Human Rights Impacts

Applying the two-step framework from the previous section, we now explore the wide-ranging human rights consequences of introducing AI decision-making into six fields:

- Criminal Justice (risk assessments)
- Finance (credit scores)
- Healthcare (diagnostics)
- Content Moderation (standards enforcement)
- Human Resources (recruitment and hiring)
- Education (essay scoring)

We chose these six fields out of many possibilities because they illustrate the promise and the perils of this technology across a range of human rights. What is more, AI decision-making technologies are already in use in all of these fields, which allows our analysis to be grounded in the here and now, rather than speculating about future developments.

In choosing these six use cases, we consciously decided not to include two AI applications that have generated a great deal of debate and controversy: namely, self-driving vehicles and autonomous weapons systems. We excluded these applications from our analysis because both are much better studied than the use of AI in the other six fields that we have chosen. Furthermore, the issues surrounding autonomous weapons systems are more appro-

priately answered with reference to international humanitarian law rather than international human rights law, since such systems are meant to be used in times of conflict.

In undertaking this analysis, it quickly became apparent to us that each AI deployment had the potential to impact a large number of rights via their first- and second-order effects. In the interest of clarity and analytical efficiency, however, we have focused our analysis on those rights that we believe to be most impacted by the deployment in question. This is an exercise in line-drawing that is subjective by its very nature, but is part and parcel of the approach embraced by the Guiding Principles to identifying human rights impacts so that they may be appropriately addressed.³⁸

There are five main points that emerge from our analysis.

First and foremost, the six use cases we explore in detail reveal how AI-based decision-making technologies impact the full spectrum of political, civil, economic, social, and cultural rights secured by the UDHR and further expounded upon in the ICCPR and the ICESCR.

Second, the positive and negative human rights impacts caused by AI are not evenly distributed across society. Some individuals and groups experience positive impacts from the very same applications that adversely impact other rights-holders. In some

³⁸ Guiding Principles, Principles 17 and 24.

cases, a particular AI application can positively impact the enjoyment of a given human right for a particular class of individuals, while adversely affecting the enjoyment of the very same human right by others. For example, the use of automated risk scoring systems in the criminal justice system may reduce the number of individuals from the majority group who are needlessly incarcerated, at the very same time that flaws in the system serve to increase the rate of mistaken incarcerations for those belonging to marginalized groups.³⁹

Third, AI carries the serious risk of perpetuating, amplifying, and ultimately ossifying existing social biases and prejudices, with attendant consequences for the right to equality. This problem, which has been termed by one analyst as “counter-serendipity,” results from the fact that AI systems are trained to replicate patterns of decision-making they learn from training data that reflects the social status quo—existing human biases, entrenched power dynamics and all.⁴⁰ But therein lies the problem: to the extent that an AI accurately replicates past patterns of human decision-making, it will necessarily perpetuate existing social biases as well.⁴¹ What is worse, unlike human decision-makers, who have the agency and the free will to change their moral perspective over time, for the foreseeable future AI systems will not have any such capabilities of their own. Instead, they require constant attention by those who are responsible for the design and operation of such systems to ensure that their outputs are consistent with evolving notions of fairness.

To be sure, the automation of decision-making through AI offers the possibility of righting significant social wrongs by designing the systems to have ameliorative effects. Such effects could be achieved by seeking to correct for biases in human decision-making, or more controversially, through “algorithmic affirmative action”—that is, by designing algorithms to counter the historical disadvantages that marginalized groups have faced.⁴² The larger point, however, is that unless AI systems are consciously designed and consistently evaluated for their differential impacts on different populations, they have the very real potential to hinder rather than help progress towards greater equity.

Fourth, as is likely expected, most AI technologies have a deleterious impact on the right to privacy. AIs are data-hungry by their nature; they are fundamentally premised on algorithms automatically poring over vast datasets to generate answers, predictions, and insights. Accordingly, AI systems rely on the collection, storage, consolidation, and analysis of vast quantities of data. They also create powerful incentives to gather and store as much additional data as can be, in view of the possibility that new data streams will allow for AI systems to generate powerful new insights. Much of the data that fuels AI systems will either be personally identifiable, or rife with the possibility of being re-identified using an algorithm in the event that it was anonymized. Moreover, even if techniques such as differential privacy⁴³ are used to protect the privacy of particular individuals, AI technologies may gen-

39 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, “Machine Bias,” *ProPublica*, May 23, 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

40 Edward Tenner, *The Efficiency Paradox: What Big Data Can't Do* (New York: Alfred A. Knopf, 2018); Berkman Klein Center for Internet and Society, “Artificial Intelligence and Inclusion,” accessed June 22, 2018, <https://aiandinclusion.org/>.

41 Anupam Chander, “The Racist Algorithm?,” *Mich. Law Review* 115, no. 6 (2017): 1023, http://michiganlawreview.org/wp-content/uploads/2017/04/115MichLRev1023_Chander.pdf.

42 Ibid.

43 Cynthia Dwork et al., “Calibrating Noise to Sensitivity in Private Data Analysis,” in *Theory of Cryptography*, ed. Shai Halevi and Tal Rabin

erate insights from such data that are then used to make predictions about, and act upon, the intimate characteristics of a particular person—all while refraining from identifying the natural person. For example, a retailer might train an AI-based marketing system using sales data that has been de-identified and subjected to differential privacy techniques. But even assuming that the training data is discarded once the system is in operation, the insights generated by the system from the data it is tasked with analyzing can nevertheless have a significant impact on an individual's privacy.⁴⁴ Given that most extant AI applications have very significant privacy implications, we focus our analysis in the case studies below on the other rights that are impacted by these systems. This is a pragmatic choice made in the interests advancing the AI and human rights conversation beyond privacy.

Fifth, the rise of artificial intelligence poses a challenge for many of the existing mechanisms that

currently exist to right wrongs. In the United States, for example, individuals have a right to request a copy of their credit report and to require credit reporting agencies to investigate and correct any errors appearing on their report.⁴⁵ By contrast, there is currently no law in the United States that would provide an individual with recourse if a lender using an algorithm that crunches through thousands of variables from thousands of sources does so on the basis of erroneous data. Even in Canada⁴⁶ and the European Union,⁴⁷ where privacy laws currently in force allow individuals to demand the correction of errors in their data, the sheer volume of information that AI systems use as they make a decision makes it difficult to exercise this right effectively. Moreover, even if one aggrieved individual corrects errors in their own data, significant harms can occur due to the presence of systematic errors in a data set and ubiquitous data sharing, which can lead to unfair outcomes for potentially vast numbers of people.

(Springer Berlin Heidelberg, 2006), 265–84.

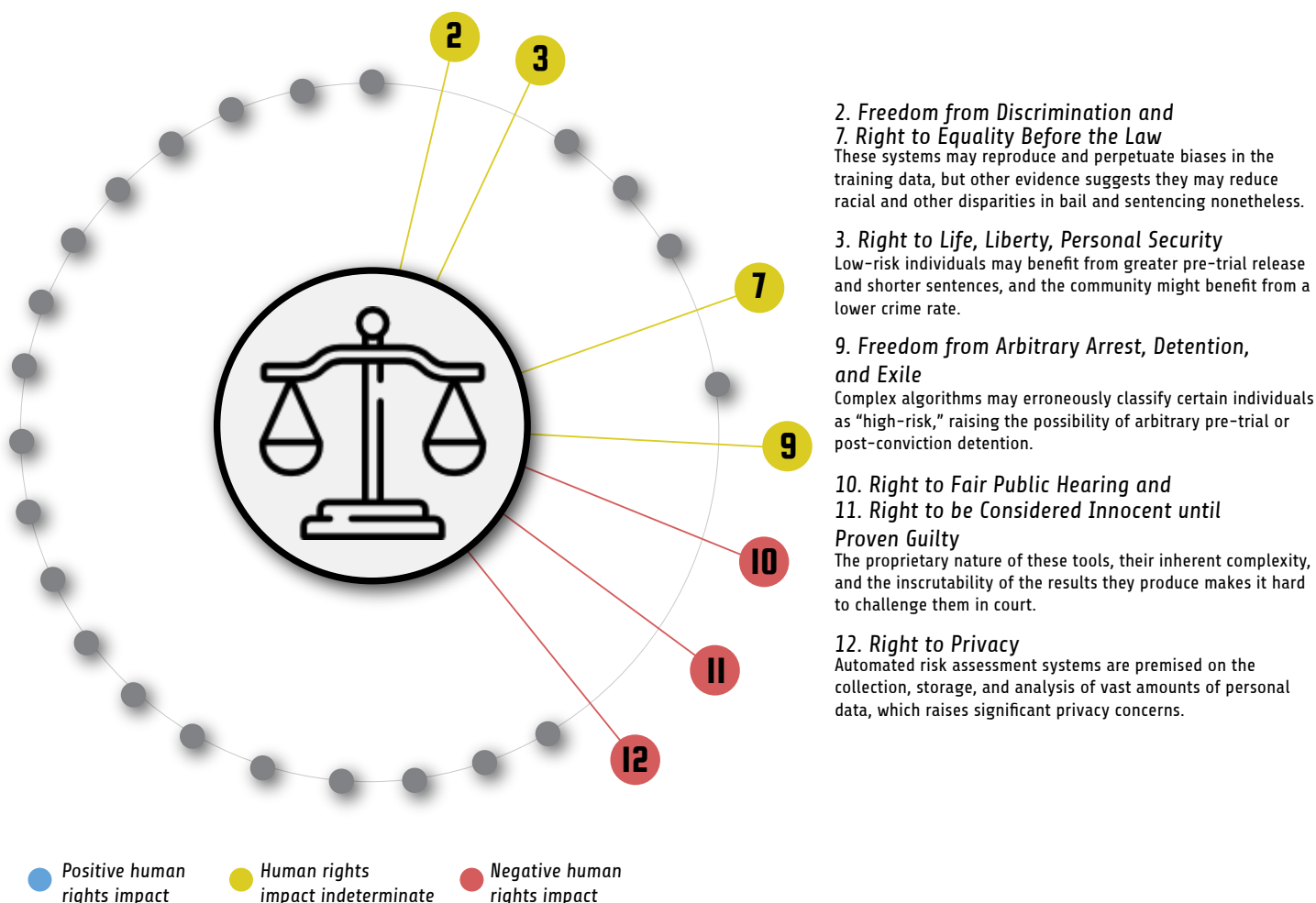
44 Charles Duhigg, “How Companies Learn Your Secrets,” *The New York Times*, February 16, 2012, sec. Magazine, <https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>.

45 *Fair Credit Reporting Act*, 15 U.S.C. § 1681i (2012) (“... if the completeness or accuracy of any item of information contained in a consumer’s file ... is disputed by the consumer ... the agency shall, free of charge, conduct a reasonable reinvestigation to determine whether the disputed information is inaccurate and record the current status of the disputed information, or delete the item from the file[.]”); *Fair Credit Reporting Act*, 15 U.S.C. § 1681j (2012) (free annual copy of one’s credit report).

46 *Personal Information Protection and Electronic Documents Act*, S.C. 2000, c. 5 (as amended June 23, 2015), Schedule 1 Principle 4.9 (“Upon request, an individual shall be informed of the existence, use, and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.”)

47 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. (L. 119) [henceforth “GDPR”], art. 19 (“The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.”).

5.1 Criminal Justice: Risk Assessments



The criminal justice system is the most potent and fearsome institution through which democratic societies may restrict an individual's enjoyment of their fundamental human rights. In view of the severity of its impacts on human rights, society has evolved a system of procedural rights to protect criminal defendants and convicts from the vagaries of human decision-making, from intentional abuse of power to unconscious influences ranging from racism to fatigue.⁴⁸

In search of both fairness and efficiency, justice systems are increasingly employing automated decision-making tools at every procedural stage. This is especially true of risk assessments, which are used to inform decisions about pretrial detention, sentencing, and parole. To the extent that they are fair and accurate, risk assessment tools can have a significant positive impact on the rights of individuals accused and convicted of crimes. The corollary, however, is that flaws or unknown limitations in the operation of such systems can have deleterious effects on a wide range of rights.

⁴⁸ Millicent H. Abel and Heather Watters, “Attributions of Guilt and Punishment as Functions of Physical Attractiveness and Smiling,” *The Journal of Social Psychology* 145, no. 6 (December 2005): 687–702, <https://doi.org/10.3200/SOCP.145.6.687-703>; Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, “Extraneous Factors in Judicial Decisions,” *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 17 (April 26, 2011): 6889–92, <https://doi.org/10.1073/pnas.1018033108>.

Traditional Approach to Risk Assessments

The first efforts to formalize the process of assessing an individual's risk of recidivism date back to the 1920s, when statisticians began to identify objective factors that are predictive of this risk for parolees.⁴⁹ As with AI now, the force driving the development of these earlier tools was the desire to avoid unnecessary deprivations of liberty and reduce the incidence of discrimination in the criminal justice system attributable to human bias. Statisticians developed these tools by collecting and analyzing information about defendants to identify factors that distinguish those that reoffend from those who do not.

As these assessments became more sophisticated, statisticians began to consider both static factors, such as a defendant's age and gender, as well as dynamic factors, such as a defendant's skill set or psychological profile.⁵⁰ Over time, these efforts led to the development of risk assessment inventories such as the Level of Service Inventory-Revised ("LSI-R")⁵¹ that, while developed and validated by statisticians, are deployed in the field by individuals

without much if any statistical expertise. Especially when such tools require their operators to make subjective determinations, such as whether an individual is engaging in antisocial behavior,⁵² these tools may suffer from low inter-rater reliability ("IRR"), calling into question the validity of the predictions generated by such tools for any given individual.⁵³ Furthermore, the data available to actuarial risk assessment systems to identify who is truly at a high risk of re-offending is systematically skewed by the fact that the pre-existing system has sentenced those it believes to pose the highest risk to long prison sentences, during which time those inmates cannot reoffend.⁵⁴

Risk assessment tools in the U.S. criminal justice system have been critiqued as inherently unfair due to the disproportionate targeting of minority individuals and communities by the police.⁵⁵ This, in turn, raises the risk that such tools will miscalculate the risk of recidivism for individuals from minority versus majority communities. Moreover, as the Supreme Court of Canada recently noted in *Ewert v. Canada*, risk assessment tools that are developed and validated based on data from majority groups

49 Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. (Chicago: University of Chicago Press, 2006) 48-60; James Bonta, "Risk-Needs Assessment and Treatment," in *Choosing Correctional Options That Work: Defining the Demand and Evaluating the Supply*, ed. Alan T. Harland (Thousand Oaks, Calif: Sage Publications, 1996); Thomas Mathiesen, "Selective Incapacitation Revisited," *Law and Human Behavior* 22, no. 4 (1998): 455-69.

50 James Bonta, "Risk-Needs Assessment and Treatment."

51 Ibid.

52 Thomas H. Cohen, "Automating Risk Assessment Instruments and Reliability: Examining an Important but Neglected Area in Risk Assessment Research," *Criminology & Public Policy* 16, no. 1 (February 2017): 271-79, <https://doi.org/10.1111/1745-9133.12272>.

53 In the risk assessment context, inter-rater reliability refers to the degree of agreement between distinct raters applying an assessment tool. A high IRR means raters apply the tool in the same manner as others; in other words, a high IRR means a particular defendant would receive the same score regardless of who conducted the assessment. A low IRR, in turn, would indicate raters may score the same defendant differently. Grant Duwe and Michael Rocque, "Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment (ROI): Recidivism Risk Assessment," *Criminology & Public Policy* 16, no. 1 (February 2017): 235-69, <https://doi.org/10.1111/1745-9133.12270>.

54 Shawn Bushway and Jeffrey Smith, "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us," *Journal of Quantitative Criminology* 23, no. 4 (December 1, 2007): 377-87, <https://doi.org/10.1007/s10940-007-9035-1>.

55 Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (2016): 671-732, <https://doi.org/10.15779/z38bg31>.

may lack validity in predicting the same traits in minority groups.⁵⁶ This may have deleterious effects on the rehabilitation of offenders from minority communities by impacting their access to cultural programming and their opportunities for parole, among other things.⁵⁷

The answer to the question of whether earlier generations of risk assessment tools have a positive or negative impact on the rights of criminal defendants and convicts to life, liberty, and security of person⁵⁸ is unclear. On one hand, they may represent an improvement over the situation where judges had essentially unfettered discretion regarding bail and sentencing decisions. On the other hand, the possibility of negative impacts exists due to the potential for the misclassification of some number of defendants as “high risk,” which results in their being sentenced more harshly than they otherwise would, or should, have been. Such tools also adversely impact criminal defendants’ rights to a fair public trial, to a defense, and to an appeal,⁵⁹ because their predictions are not subject to meaningful review by courts. Not only do courts lack the institutional capacity to review the operation of such tools, but the objective veneer that coats the outputs of these tools obscures the subjective determinations that are baked into them.

Furthermore, these tools raise fundamental questions as to whether it is fair to treat a particular individual more harshly simply because they share char-

acteristics with others who have reoffended. This is a particularly serious difficulty when it comes to individuals who are classified as “high risk” yet for whatever reason do not reoffend. While statistical techniques can determine with a high degree of accuracy the characteristics of individuals in a population who are likely to behave in a certain way, they cannot generate accurate predictions as to how any particular individual in that population will behave. This raises some truly vexing legal, moral, and philosophical questions that are common to all the case studies that follow.

AI-Generated Risk Assessments

In recent years, criminal justice systems in many different countries have begun to use algorithmic risk assessment tools. All such tools automate the analysis of whatever data has been inputted into the system. Most of these tools still rely on manually-inputted data from questionnaires similar to those that were part and parcel of the last generation of risk-assessment tools, while newer tools are fully automated and rely on information that already exists in various government databases.⁶⁰

Full automation improves the predictive accuracy and validity of risk assessment tools because the software interprets every piece of data consistently.⁶¹ Automation also obviates the need for manual data collection, entry, and scoring, which carries with it the possibility of improving the accuracy of these

⁵⁶ *Ewert v. Canada*, 2018 SCC 30. Note, however, that other tools—such as the Ontario Domestic Assault Risk Assessment Tool (“ODARA”)—are in widespread use in Canada and have been adopted by courts in several provinces and territories. For examples of courts relying on these tools, see *R v. Beharri*, 2015 ONSC 5900; *R v. Primmer* 2017 ONSC 2953; *R v. Sassie*, 2015 NWTCA 7; *R. v. Robertson*, 2006 ABPC 88.

⁵⁷ *Ewert v. Canada*, 2018 SCC 30.

⁵⁸ UDHR art. 3.

⁵⁹ UDHR arts. 10 and 11(1); ICCPR art. 14(5).

⁶⁰ The Minnesota Screening Tool Assessing Recidivism Risk 2.0 (“MnSTARR 2.0”) under development by the government of the U.S. State of Minnesota is a leading example of a fully-automated risk assessment tool in the criminal justice context. Kenneth C. Land, “Automating Recidivism Risk Assessment: Should We Stay or Should We Go?,” *Criminology & Public Policy* 16, no. 1 (February 2017): 231–33, <https://doi.org/10.1111/1745-9133.12271>.

⁶¹ *Ibid.*

systems by, for example, allowing additional variables to be considered.⁶²

Beyond full automation, the latest generation of risk assessment tools leverages machine learning techniques to continually rebalance risk factors in response to new inputs. In theory, the predictive power and accuracy of such systems should improve over time. This was the finding of a proof-of-concept study in New York City, where researchers used machine learning techniques to determine which criminal defendants should receive bail.⁶³ The study's results suggest that New York could reduce the number of people held in pretrial detention by 40% without any corresponding increase in the crime rate. Alternately, the city could reduce its crime rate by 25% by incarcerating the same number of people, but changing the criteria for who gets bail. In so doing, the number of African-Americans and Hispanics housed in the city's jails would be significantly reduced, with concomitant positive effects on the right to equality and non-discrimination.⁶⁴

For all of these potential positives, the single most widely-used algorithmic risk assessment system in the United States has been accused of perpetuating racial bias. An investigation by ProPublica found

that COMPAS, a proprietary risk-assessment system that certain U.S. state courts use in making bail and sentencing decisions, misclassified African-American offenders as "high-risk" at twice the rate of Caucasians, even though the system had nearly the same accuracy rate (63% vs. 59%) in predicting when individuals from both racial groups would reoffend.⁶⁵ In other words, COMPAS classified 45% of those African-American convicts who ultimately did not reoffend as "high risk," as compared to just 23% for similarly-situated Caucasians. Questions have been raised about the accuracy and the methodological validity of the ProPublica report,⁶⁶ but more fundamentally, an important paper published in the aftermath of the COMPAS controversy suggests that it may be well-nigh impossible to design algorithms that treat individuals belonging to different groups equally fairly across multiple different dimensions of fairness.⁶⁷ Assuming, however, that the issues ProPublica identified with COMPAS are well-founded and are true of other risk assessment algorithms, then there is a substantial risk that the rights of minority groups to equality and non-discrimination will be adversely affected by such tools.⁶⁸

Furthermore, there is a serious issue relating to the existence of systematic patterns of bias against

62 According to Barocas and Selbst, one source of bias is inaccuracies in the selected features. Additional features should, in theory, allow for more accurate generalizations to be developed. Barocas and Selbst, "Big Data's Disparate Impact."

63 Jon Kleinberg et al., "Human Decisions and Machine Predictions" (Cambridge, MA: National Bureau of Economic Research, February 2017), <https://doi.org/10.3386/w23180>.

64 UDHR art. 2.

65 Jeff Larson and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

66 For example, Anthony W Flores, Kristin Bechtel, and Christopher T. Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,'" *Federal Probation* 80, no. 2 (2016): 9.

67 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *ArXiv:1609.05807 [Cs, Stat]*, September 19, 2016, <http://arxiv.org/abs/1609.05807>.

68 UDHR art. 2.

minorities in the data being used to train these algorithmic risk-assessment tools, arising from the disproportionate police scrutiny that minority community members receive. Consequently, minority communities are over-represented in the training data, which results in variables that are close proxies for race being over-weighted by these algorithms in assessing the risk that any particular individual poses.⁶⁹ This, too, raises concerns about algorithmic risk assessment tools having negative impacts on the rights of minority groups to equality and non-discrimination.⁷⁰

There are also issues that arise from the development of these risk assessment tools by private companies who, for commercial reasons, guard their algorithms and the data that is used to train them as trade secrets.⁷¹ The secrecy that often surrounds the operation of these risk assessment tools can have adverse impacts on the rights of criminal defendants to defend themselves against criminal charges⁷² and to appeal a conviction.⁷³ The situation is further complicated when risk assessment algorithms rely upon machine learning techniques to adapt their performance over time, as the results generated by such techniques are oftentimes neither reproducible nor explainable in any meaningful way.

Summary of Impacts

The current generation of automated risk-assessment tools has the potential to positively impact the rights of “low-risk” criminal defendants and offenders to life, liberty, and security of the person.⁷⁴ If indeed such tools are more accurate than humans at predicting the risk of recidivism,⁷⁵ low-risk offenders will end up being incarcerated at a lower rate and for shorter periods of time than under the status quo. Members of society at large will also be more secure in the enjoyment of their right to security of the person should these tools result in a lower rate of crime.

It is hard to know, however, whether the current generation of automated risk assessment tools is having a negative or positive impact on the equality and non-discrimination rights of criminal defendants from groups that have historically been discriminated against, such as ethnic minorities and the mentally ill.⁷⁶ While the existence of systemic biases in the training data may result in the automation of existing social biases against individuals from these groups, the results of the New York City proof-of-concept study suggest that such systems may nevertheless ameliorate the over-representation of individuals from these groups in jail and prison populations.

⁶⁹ Barocas and Selbst, “Big Data’s Disparate Impact.”

⁷⁰ UDHR art. 2.

⁷¹ Rebecca Wexler, “Life, Liberty, and Trade Secrets,” *Stanford Law Review* 70, no. 5 (2018): 1343.

⁷² UDHR art. 11(1).

⁷³ ICCPR art. 14(5).

⁷⁴ UDHR art. 2.

⁷⁵ This assumption has been questioned. Julia Dressel and Hany Farid, “The Accuracy, Fairness, and Limits of Predicting Recidivism,” *Science Advances* 4, no. 1 (January 1, 2018), <https://doi.org/10.1126/sciadv.aao5580>.

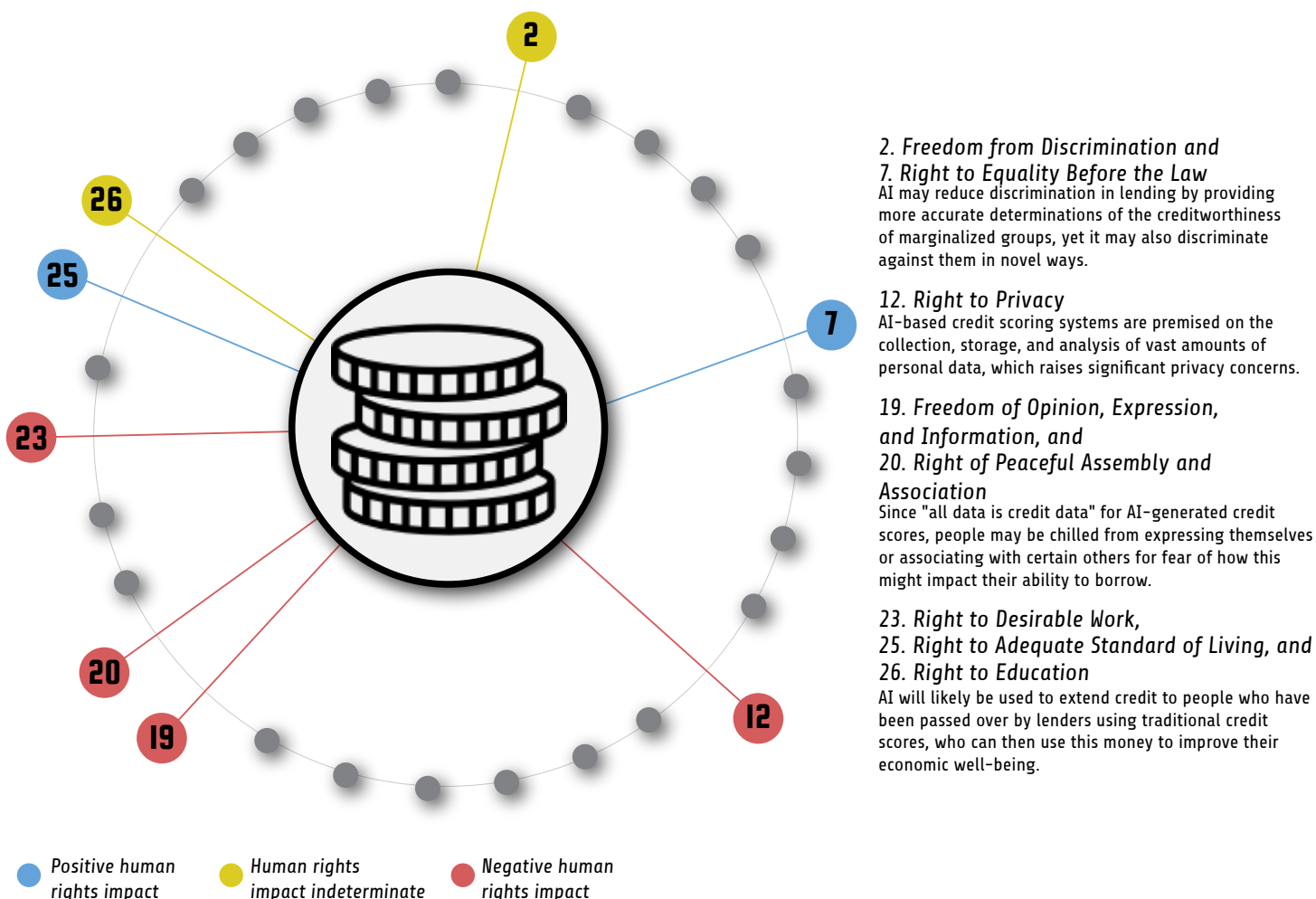
⁷⁶ UDHR art. 3.

Finally, in view of the inscrutability of the latest generation of automated risk assessment tools, and the secrecy surrounding these tools when they are developed by the private sector, we believe that these tools are likely to adversely impact the rights of criminal defendants to a fair and public hearing before an independent and impartial tribunal,⁷⁷ and to enjoy all of the guarantees needed for their defense.⁷⁸

⁷⁷ UDHR art. 10.

⁷⁸ UDHR art. 11(1). Relatedly, the right of criminal convicts under ICCPR art. 14(5) is similarly impacted.

5.2 Access to the Financial System: Credit Scores



Access to financial services such as banking and lending are an important means of promoting social and economic well-being. Access to credit in particular can help disadvantaged and marginalized individuals better enjoy their economic, social, and cultural rights by, for example, providing them with the means to pursue higher education,⁷⁹ access

health care,⁸⁰ purchase property,⁸¹ or start a business through which they can be gainfully employed.⁸² In view of the role that credit can play in advancing the achievement of a wide range of human rights, the Nobel laureate Muhammad Yunus has suggested that access to credit itself ought to be considered a human right.⁸³

⁷⁹ UDHR art. 26.

⁸⁰ UDHR art. 25.

⁸¹ UDHR art. 17.

⁸² UDHR art. 23.

⁸³ Matt Wade, "Access to Credit a 'Human Right,' Says the Father of Microfinance," *Sydney Morning Herald*, October 9, 2014, <https://www.smh.com.au/national/access-to-credit-a-human-right-says-the-father-of-microfinance-20141009-113j3x.html>. For more on the debate, see the following two resources: Marek Hudon, "Should Access to Credit Be a Right?," *Journal of Business Ethics* 84, no. 1 (2009): 17–28 and John Gershman and Jonathan Morduch, "Credit Is Not a Right," in *Microfinance, Rights and Global Justice*, ed. Tom Sorell and Luis Cabrera (Cambridge: Cambridge University Press, 2015), 14–26, <https://doi.org/10.1017/CBO9781316275634.002>.

Traditional Approach to Credit Scoring

In deciding whether to extend someone credit, lenders have long sought to ascertain the prospective borrower's risk of defaulting on the debt. Such determinations have historically been of dubious accuracy and rife with the possibility of discrimination, as lenders based them on their personal impressions of the borrower coupled with references from the community.⁸⁴ Nor were these determinations improved much by the development of the first credit reports around the turn of the 20th century, which consisted of compilations of information about an individual's personal affairs that were subject to the discretionary review of the lender.⁸⁵

In the United States, the legislative efforts of the 1970s to outlaw discrimination in lending based on race, religion, gender, age, and other similar traits roughly coincided with the development of the first credit scores, which attempted to reduce all of the information contained in an individual's credit score into a simple, numerical indication of that person's credit-worthiness.⁸⁶ Different companies use different approaches to calculate credit scores. FICO Scores, which are used by 90% of lenders in

the United States, are generated based on a combination of an individual's payment history, the amount that they owe, the age of their accounts, their sources of credit, and how much additional credit they have sought recently.⁸⁷

Despite their objective veneer, traditional credit scores suffer from several limitations that can adversely impact human rights. Since traditional credit scores rely on information gathered by credit bureaus about an individual's past financial history, oftentimes individuals with a "thin" credit file are given a credit score that is not indicative of their true risk of defaulting or are denied a credit score entirely.⁸⁸ Such "thin-file" borrowers tend to belong to marginalized groups such as minorities, young adults, immigrants, and recently-divorced women.⁸⁹ Since financial institutions are less likely to lend to individuals from these groups, even when in reality they are just as credit-worthy as "thick-file" applicants from other groups, the right to equality may be adversely impacted.⁹⁰

There are also issues relating to the fairness and accuracy of the data being fed into credit-scoring algorithms. In the United States at least, credit bu-

84 Matthew A. Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data," *Chicago-Kent Law Review* 93, no. 1 (March 9, 2018): 2–60.

85 Sean Trainor, "The Long, Twisted History of Your Credit Score," *Time*, accessed June 10, 2018, <http://time.com/3961676/history-credit-scores/>.

86 Those laws are the Fair Credit Reporting Act ("FCRA") of 1970, the Equal Credit Opportunity Act ("ECOA") of 1974 and the Community Reinvestment Act of 1977. Willy E. Rice, "Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950-1995," *San Diego Law Review* 33 (1996): 583–700.

87 Rob Kaufman, "5 Factors That Determine a FICO Score," *myFICO* (blog), September 23, 2016, <https://blog.myfico.com/5-factors-determine-fico-score/>.

88 Kenneth P. Brevoort, Philip Grimm, and Michelle Kambara, "Data Point: Credit Invisibles" (Consumer Financial Protection Bureau Office of Research, May 2015).

89 Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data." ("In other words, credit invisibles are generally either too young to have established a credit history, or have never been welcomed into the traditional banking system. As such, [algorithmic credit scores] would especially benefit the young, the low-income, and minorities." (internal citations omitted)).

90 UDHR art. 2. Brevoort, Grimm, and Kambara, "Data Point: Credit Invisibles."

reus rely on “furnishers”—banks, utilities, and other businesses—to voluntarily report relevant information, such as on-time payments, debt balances, and the like.⁹¹ In view of the legal obligations that attach to furnishers when they provide information to a credit bureau, such businesses are more likely to report adverse events (such as a missed payment or foreclosure) that negatively impact its own bottom line, as opposed to routine, unremarkable positive events (such as timely payments).⁹² Since individuals from minority communities suffer from adverse financial events (such as evictions) at a higher rate than would be predicted by their actual financial circumstances,⁹³ there is a significant risk that the information used to generate credit scores is systematically biased against minority communities.

Furthermore, even if all relevant information (both positive and negative) is reported to a credit agency, there is no guarantee that the credit scoring algorithm will consider it. For example, the FICO score in wide use in the United States considers only mortgage and credit-card payment history, but not rental or bill payment history.⁹⁴ In view of the long

legacy of discriminatory lending policies in the U.S. and of housing policies that make it much more likely that individuals from minority groups will rent rather than own a home,⁹⁵ these practices can have significant discriminatory impacts.⁹⁶

The growing use of credit scores beyond the lending context amplifies these effects. It is increasingly common for employers, landlords, and insurers to review an individual's credit score before offering them a job, renting them an apartment, or selling them insurance.⁹⁷ Employers may think that credit scores are a proxy for an applicant's integrity and responsibility, even though they have not been validated for that purpose.⁹⁸ Insurers may similarly view those with poor credit scores as posing a higher actuarial risk because “recklessness” in paying down one's debts shows the individual to be a reckless person in general.⁹⁹ Yet again, such practices pose a grave risk of perpetuating and amplifying age-old patterns of inequality and discrimination that bear little resemblance to reality.

91 “Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003” (Federal Trade Commission, January 2015).

92 Jocelyn Baird, “What Gets Reported to Your Credit Reports (and What Doesn't)?,” *NextAdvisor* (blog), accessed June 20, 2018, <https://www.nextadvisor.com/blog/what-gets-reported-to-your-credit-reports/>.

93 Deena Greenberg, Carl Gershenson, and Matthew Desmond, “Discrimination in Evictions: Empirical Evidence and Legal Challenges,” *Harvard Civil Rights* 51, no. 1 (2016): 44.

94 Preeti Vissa, “How Credit Scores Disproportionately Hurt Communities of Color,” *Huffington Post* (blog), December 15, 2010, https://www.huffingtonpost.com/preeti-vissa/credit-scores-and-the-for_b_797148.html; “How Credit History Impacts Your FICO® Score,” *myFICO* (blog), accessed June 20, 2018, <http://www.myfico.com/credit-education/credit-payment-history>.

95 Christopher E. Hebert et al., “Homeownership Gaps Among Low-Income and Minority Borrowers and Neighborhoods” (U.S. Department of Housing and Urban Development, March 2005); Sarah Ludwig, “Credit Scores in America Perpetuate Racial Injustice. Here's How,” *The Guardian*, October 13, 2015, sec. Opinion, <http://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why>.

96 UDHR art. 3.

97 “Past Imperfect: How Credit Scores and Other Analytics ‘Bake In’ and Perpetuate Past Discrimination” (National Consumer Law Center, May 2016) (“Credit history is used as a gatekeeper for many important necessities – employment, housing (both rental and homeownership), insurance, and of course, affordable credit.”).

98 Gary Rivlin, “Employers Pull Applicants’ Credit Reports,” *The New York Times*, May 11, 2013, sec. Business Day, <https://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html>.

99 “Past Imperfect: How Credit Scores and Other Analytics ‘Bake In’ and Perpetuate Past Discrimination.”

AI-Generated Credit Scores

In recent years, lenders have begun to use artificial intelligence to more accurately assess whether a potential borrower is a good credit risk. Unlike conventional credit scoring algorithms, the AI-based approach treats “all data as credit data” and analyzes vast amounts of data from many sources.¹⁰⁰ The resulting AI-generated credit scores are better than traditional scores at addressing some kinds of situations, at the same time as they create new challenges of their own.

The volume of data that AI-based credit scoring systems collect and analyze is so staggering as to be concerning. ZestFinance, one of the leading companies in this field in the US, considers over 3,000 variables in deciding whether to offer someone credit¹⁰¹—including whether the applicant tends to type in all-caps, which apparently is correlated with a higher risk of default.¹⁰² Lenddo, another American company in this space, examines an applicant’s entire digital footprint—including social media use, geolocation, website browsing habits, phone use history (including text and call logs), purchasing behavior, and more in deciding whether to extend them credit.¹⁰³

AI-generated credit scores have particularly significant applications in emerging markets, where almost everyone is a “thin-file” borrower. For example, the MyBucks Haraka app in use in India, the Philippines, and several Sub-Saharan countries uses data gleaned from an applicant’s mobile phone (call logs, geolocation information, and the like) and their social media accounts to generate an alternative credit score that partner banks can use to inform their lending decision.¹⁰⁴ This AI-based approach has the potential to help members of historically marginalized groups, such as women and ethnic minorities, gain access to credit in the developed and developing world alike,¹⁰⁵ thereby fostering financial inclusion and advancing the right to equality.¹⁰⁶

Early results suggest that these technologies are succeeding in fostering financial inclusion. ZestFinance claims that its AI-based technology allowed it to reduce its default rate to less than half of the prevailing industry average,¹⁰⁷ while Lenddo claims to have increased its approval rate by 15% while slashing defaults by 12%.¹⁰⁸ If these early results are accurate and generalizable, the positive impact on all of the economic, cultural, and social rights that access to credit enables would be very significant indeed.

100 James Rufus Koren, “Some Lenders Are Judging You on Much More than Finances,” *Los Angeles Times*, December 19, 2015, <http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html>.

101 “Zest Automated Machine Learning Data Sheet” (ZestFinance), accessed June 20, 2018, <https://www.zestfinance.com/hubfs/Underwriting/Zest-Automated-Machine-Learning-Data-Sheet.pdf?hsLang=en>.

102 Koren, “Some Lenders Are Judging You on Much More than Finances.”

103 “Credit Scoring: The LenddoScore Fact Sheet” (Lenddo), accessed June 20, 2018, https://www.lenddo.com/pdfs/Lenddo_FS_CreditScoring_201705.pdf.

104 Penny Crosman, “This Lender Is Using AI to Make Loans through Social Media,” *American Banker*, December 8, 2017, <https://www.americanbanker.com/news/this-lender-is-using-ai-to-make-loans-through-social-media>.

105 Brevoort, Grimm, and Kambara, “Data Point: Credit Invisibles.”; Geri Stengel, “How One Woman Is Changing Business Lending In Africa,” *Forbes*, January 14, 2015, <https://www.forbes.com/sites/geristengel/2015/01/14/how-one-woman-is-changing-business-lending-in-africa>.

106 UDHR art. 2.

107 John Lippert, “ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed out Deadbeats,” *The Washington Post*, October 11, 2014, sec. Business, https://www.washingtonpost.com/business/zestfinance-issues-small-high-rate-loans-uses-big-data-to-weed-out-deadbeats/2014/10/10/e34986b6-4d71-11e4-aa5e-7153e466a02d_story.html.

108 “Credit Scoring: The LenddoScore Fact Sheet.”

Yet there are considerable risks to this new approach as well. One arises from the quality and accuracy of the data used to train these systems, as well as the fairness and accuracy of the data these systems use to decide upon a particular individual's application for credit. The issues are similar in nature to those affecting traditional credit scoring algorithms, but they are different in degree due to the vast number of data sources that AI-based algorithms take into consideration.

Another arises from the subjective decisions that programmers make on how to code and categorize the data that they feed into their seemingly-objective algorithms.¹⁰⁹ For example, ZestFinance translates certain continuous variables (such as the length of time one spends reading their website's terms and conditions) into categorical values (like 0, 1, or 2).¹¹⁰ This is an inherently subjective process that can introduce explicit or implicit bias into the data and consequently into the results generated by the algorithm.

Furthermore, AI-generated scores may perpetuate existing patterns of discrimination through "network discrimination,"¹¹¹ whereby individuals are penalized (or rewarded) based on the characteristics of others who are in their personal network. For example, if there are two individuals in an identical financial position, yet the first individual's friends

live in "rich" neighborhoods while the second's friends live in "poor" neighborhoods, an algorithm may well determine the first to be a better credit risk than the second.¹¹² To the extent that such network factors correlate with invidious classifications such as those based on race and gender, the potential for discriminatory impacts is quite serious indeed.¹¹³

The use of AI in financial decision-making may even burden individuals' freedom of opinion, expression, and association by chilling individuals from engaging in activities that they believe will negatively affect their credit score. This is not a mere theoretical possibility. In 2009, American Express reduced the credit limit of an African-American businessman because "[o]ther customers who have used their card at establishments where [he] recently shopped . . . ha[d] a poor repayment history."¹¹⁴ Another lender in the U.S. reduced the credit limit of its customers who had incurred expenses at "marriage counselors, tire retreading and repair shops, bars and nightclubs, pool halls, pawn shops, massage parlors, and others."¹¹⁵

An even more extreme example of this phenomenon is China's incipient "social credit score" system, which generates a numerical index of an individual's "trustworthiness" based on a vast array of data points, including social media data, arrest

¹⁰⁹ Mikella Hurley and Julius Adebayo, "Credit Scoring in the Era of Big Data," *Yale Journal of Law & Technology* 18, no. 1 (2016): 148–216.

¹¹⁰ Ibid.

¹¹¹ danah boyd, Karen Levy, and Alice Marwick, "The Networked Nature of Algorithmic Discrimination," in *Data and Discrimination: Collected Essays*, ed. Seeta Peña Gangadharan (New America, 2014), 53–57.

¹¹² Kaveh Waddell, "How Algorithms Can Bring Down Minorities' Credit Scores," *The Atlantic*, December 2, 2016, <https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/>.

¹¹³ UDHR arts. 2 and 20.

¹¹⁴ Ron Lieber, "American Express Watched Where You Shopped," *The New York Times*, January 30, 2009, sec. Your Money, <https://www.nytimes.com/2009/01/31/your-money/credit-and-debit-cards/31money.html>.

¹¹⁵ Ibid.

and infraction records, volunteer activity, city and neighborhood records, and more.¹¹⁶ Those with high social credit scores enjoy benefits such as lower utility rates and more favorable borrowing conditions, while those with unfavorable scores might be unable to purchase airline or high speed rail tickets.¹¹⁷ These systems are being piloted in several communities, but a national roll-out of the “social credit score” system is expected by 2020. While anecdotal reports suggest that the “social credit scoring” system has curbed corruption and incentivized certain forms of good behavior, such as stopping for pedestrians at crosswalks,¹¹⁸ it is not hard to imagine how this system could chill a great deal of expressive and associative activity.¹¹⁹

Summary of Impacts

Compared to the status quo credit scoring algorithms, the introduction of AI into the lending process is likely to have an overall positive impact on the ability of objectively low-risk borrowers to access credit. This is likely to have positive impacts on the enjoyment by these individuals of the right to an adequate standard of living,¹²⁰ the right to work,¹²¹ and the right to education,¹²² as access to credit is a powerful enabler of these economic and social rights.

The introduction of AI into the lending process is also likely to have a positive impact on the right to equality and non-discrimination for some individuals, while adversely affecting it for others. On the positive side, the fact that AI-based algorithms consider a wide variety of data sources may improve the ability of well-qualified individuals from marginalized communities to access credit by overcoming the “thin-file” problem. On the other hand, the specter of “network discrimination” having a negative impact on the ability of members of these very same communities to borrow money cannot be discounted.

Finally, it is likely that AI-based decision-making algorithms in the financial sector will adversely impact the freedoms of opinion, expression, and association. In an era where “all data is credit data,” individuals may feel chilled from expressing certain points of view or associating with others, out of fear that an algorithm may use their behavior against them in the financial context.

¹¹⁶ Mara Hvistendahl, “In China, a Three-Digit Score Could Dictate Your Place in Society,” *WIRED*, Dec. 14, 2017, <https://www.wired.com/story/age-of-social-credit/>.

¹¹⁷ Simina Mistreanu, “Life Inside China’s Social Credit Laboratory,” *Foreign Policy*, April 3, 2018, <https://foreignpolicy.com/2018/04/03/life-in-side-chinas-social-credit-laboratory/>.

¹¹⁸ *Ibid.*

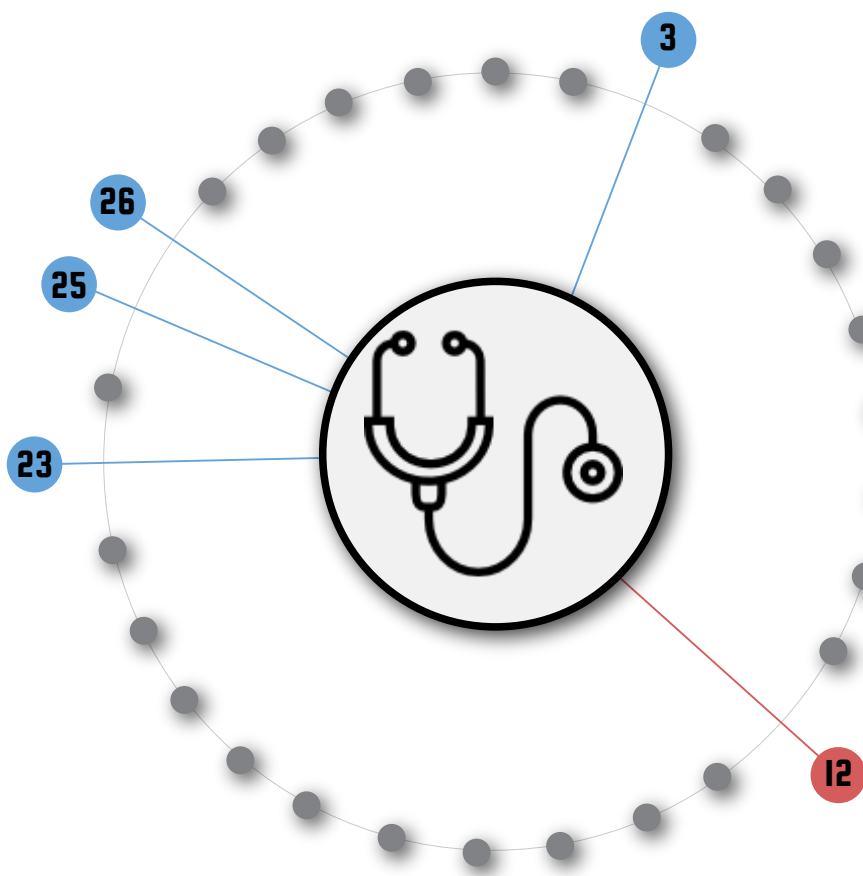
¹¹⁹ UDHR arts. 19 and 20. In this context, the U.N. Human Rights Committee has noted that it is “impermissible” for states to engage in conduct that create “chilling effects that may unduly restrict the exercise of freedom of expression...” United Nations Human Rights Committee, General Comment 34 (ICCPR Art. 19: Freedoms of opinion and expression) (2011), U.N. Doc. CCPR/C/GC/34, p. 12.

¹²⁰ UDHR art. 17.

¹²¹ UDHR art. 23.

¹²² UDHR art. 26.

5.3 Healthcare: Diagnostics



● Positive human rights impact ● Human rights impact indeterminate ● Negative human rights impact

3. Right to Life, Liberty, and Security of Person

AI-based diagnostic systems enhance the enjoyment of the right to life by making accurate, high-quality diagnostic services more widely available.

12. Right to Privacy

AI-based diagnostic systems require the collection of vast quantities of sensitive data relating to an individual's often-immutable health characteristics, raising serious privacy concerns.

23. Right to Desirable Work,

The improved health outcomes that AI-based diagnostic systems are likely to produce will reduce the number of people who are excluded from the dignity of work for medical reasons.

25. Right to Adequate Standard of Living, and

By detecting diseases earlier and more accurately, AI-based diagnostic systems will improve living standards and quality of life.

26. Right to Education

Should AI-based diagnostic systems deliver on their promise, fewer people will be excluded from the enjoyment of the right to the education for reasons of ill-health.

In the course of the last century, modern medicine has produced astonishing improvements in the length and the quality of the lives of all those who can access it. Not only does the ICESCR recognize “the right of everyone to the enjoyment of the highest attainable standard of physical and mental health,”¹²³ but good health is arguably a necessary condition for each and every one of us to enjoy the full range of human rights that we are guaranteed by law.

Recent advances in health outcomes are attributable to improvements in the three pillars of healthcare: prevention, diagnosis, and treatment. AI has applications across all three pillars, but its greatest impact to date has been on improving the accuracy of medical diagnosis.

¹²³ ICESCR art. 12.

Traditional Approach to Diagnostics

Physicians use a wide range of approaches to diagnose disease. Perhaps the simplest and most widespread is to identify the patient's symptoms and correlate them to conditions or diseases that are characterized by the same pattern of symptoms.¹²⁴ The same basic approach can be applied to interpreting the results of diagnostic tests: a radiologist reviewing MRI imagery or a pathologist analyzing a biopsy sample compare what they are seeing to what they have learned in order to make a diagnosis.

Needless to say, it takes years of training and years more of experience to develop the knowledge and mastery required to accurately diagnose the wide range of maladies that afflict our species. To simplify matters, physicians often rely on “diagnostic criteria” in determining what ails someone. These are essentially statistically-validated rules of thumb that can be used to rule in or rule out a particular condition.¹²⁵ By contrast, experts in particular diseases engage in *gestalt* pattern recognition to recognize the characteristic indicators of a particular disease in a sea of information.¹²⁶

Unfortunately, errors in diagnostics are extremely common, and they can have life-and-death consequences. One recent study found that 5% of patients in the U.S. are misdiagnosed every year,¹²⁷ while another found that misdiagnosis is the cause of 10% of patient deaths.¹²⁸ The challenge for physicians is growing as the number of diagnostic tests and procedures multiplies with the advance of medical science. Since each of these procedures has its own unique operating parameters and error rates,¹²⁹ it is becoming increasingly difficult for the average medical practitioner to choose the right test for their patient—or to even refer their patients to the right sub-specialists in view of their symptoms.¹³⁰

AI-Assisted Diagnostics

Medical diagnostics is one of the fields in which AI-based technologies went into widespread use. Efforts began in the 1970s to start codifying the knowledge of human diagnostic experts into automated “expert systems.”¹³¹ These systems, which are known as “diagnostic decision support systems” and are used in many healthcare settings today, require the human clinician to answer a series of questions

¹²⁴ Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*, ed. Erin P. Balogh, Bryan T. Miller, and John R. Ball (Washington, D.C.: National Academies Press, 2015), <https://doi.org/10.17226/21794>.

¹²⁵ The “Centor Criteria” for diagnosing strep throat in adults is an example of this. Since roughly 50% of patients who have all five of the criteria (cough, tonsillar exudates, swollen lymphatic nodes, fever, neither young nor old) will turn out to have strep throat, the Centor criteria presents a quick and easy way to rule in or rule out strep throat as a possibility in a patient presenting with these symptoms. Robert M. Centor et al., “The Diagnosis of Strep Throat in Adults in the Emergency Room,” *Medical Decision Making* 1, no. 3 (August 1981): 239–46, <https://doi.org/10.1177/0272989X8100100304>.

¹²⁶ John P. Langlois, “Making a Diagnosis,” in *Fundamentals of Clinical Practice*, ed. Mark B Mengel, Warren Lee Holleman, and Scott A Fields, 2nd ed. (New York: Kluwer Academic/Plenum Publishers, 2005).

¹²⁷ Hardeep Singh, Ashley N D Meyer, and Eric J Thomas, “The Frequency of Diagnostic Errors in Outpatient Care: Estimations from Three Large Observational Studies Involving US Adult Populations,” *BMJ Quality & Safety* 23, no. 9 (September 2014): 727–31, <https://doi.org/10.1136/bmjqs-2013-002627>.

¹²⁸ Ibid.

¹²⁹ Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*.

¹³⁰ J. Hickner et al., “Primary Care Physicians’ Challenges in Ordering Clinical Laboratory Tests and Interpreting Results,” *The Journal of the American Board of Family Medicine* 27, no. 2 (March 1, 2014): 268–74, <https://doi.org/10.3122/jabfm.2014.02.130104>.

¹³¹ MYCIN was one of the pioneer expert systems developed at Stanford starting in 1972. Nancy McCauley and Mohammad Ala, “The Use of Expert Systems in the Healthcare Industry,” *Information & Management* 22, no. 4 (April 1, 1992): 227–35, [https://doi.org/10.1016/0378-7206\(92\)90025-B](https://doi.org/10.1016/0378-7206(92)90025-B).

about the patient's condition that help to rule in or rule out certain specific diagnoses.

In the last several years, systems based on machine learning or deep learning have begun to be developed to facilitate and automate the diagnosis of illness across a range of medical specialties. Few of these technologies are currently in use, though early results suggest that they have great promise in improving the accuracy of medical diagnosis.

For example, an AI-powered image recognition system was able to detect cancerous skin lesions correctly 72% of the time, whereas human dermatologists correctly diagnosed the cancers 66% of the time.¹³² There are also anecdotes about AI-powered diagnostic systems quickly solving intractable mysteries. For example, a diagnostic system powered by IBM's Watson was able to diagnose a patient as possessing a rare form of leukemia within 10 minutes, even though her symptoms had stumped experts for several months.¹³³ The system did so by comparing information in the patient's medical records with over 20 million oncology records held by the University of Tokyo. To be sure, physicians currently outperform AI systems on a wide variety of diagnostic tasks—from microscopy to general diagnosis. Yet it is impressive that AI systems rival or outperform human experts in diagnosing conditions ranging from brain cancer to autism and Alzheimer's disease.¹³⁴

AI-based diagnostic systems also have the potential to provide greater access to specialist-level treatment than is currently possible. One of the few AI-based diagnostic systems to be approved for clinical use by the U.S. Food and Drug Administration is able to detect and diagnose diabetic retinopathy (a disorder affecting the vision of individuals suffering from diabetes) autonomously.¹³⁵ Whereas this condition was previously one that could only be diagnosed by a specialist, AI makes it possible for anyone trained in using the machinery to do so.

Summary of Impacts

AI-based diagnostic systems, especially the latest generation of systems that leverage artificial intelligence, are very likely to positively impact the right each of us enjoys to the highest attainable standard of health.¹³⁶ Not only do AI-based diagnostic systems appear to meet or exceed the performance of human experts in diagnosing disease, they have the potential to be much more accessible than specialized human experts, who require years of training and experience to rival the accuracy of an AI.

It is significant that in recognizing the right that each of us possesses “to a standard of living adequate for the health and well-being of himself and of his family,” Article 25 of the UDHR links access to medical care to the basic requisites of life, such as food, clothing, and housing. In view of this link

¹³² Siddhartha Mukherjee, “A.I. Versus M.D.,” *The New Yorker*, March 27, 2017, <https://www.newyorker.com/magazine/2017/04/03/ai-versus-md>.

¹³³ James Billington, “IBM's Watson Cracks Medical Mystery with Life-Saving Diagnosis for Patient Who Baffled Doctors,” *International Business Times UK*, August 8, 2016, <https://www.ibtimes.co.uk/ibms-watson-cracks-medical-mystery-life-saving-diagnosis-patient-who-baffled-doctors-1574963>.

¹³⁴ “AI vs Doctors,” *IEEE Spectrum: Technology, Engineering, and Science News*, September 26, 2017, <https://spectrum.ieee.org/static/ai-vs-doctors>.

¹³⁵ Angela Chen, “AI Software That Helps Doctors Diagnose like Specialists Is Approved by FDA,” *The Verge*, April 11, 2018, <https://www.theverge.com/2018/4/11/17224984/artificial-intelligence-idxdr-fda-eye-disease-diabetic-rethinopathy>.

¹³⁶ ICESCR art. 12.

between good health, access to health care, and the full range of economic, social, and cultural rights that each of us enjoys, the use of AI in medical diagnostics is likely to have positive impacts on the right of each of us to work and in so doing, ensure ourselves an existence worthy of human dignity.¹³⁷ Likewise, the better health outcomes that AI-based diagnostics are likely to produce will positively impact the enjoyment of the right to education by those who would otherwise be excluded by reasons of illness.¹³⁸

As with many other automated technologies, there is the possibility that AI-based diagnostic technologies will cause employment losses in the medical field. While the right to work does not entail the right to work in any particular position, occupation, or field, the state obligation to protect the right to work and progressively adopt measures to realize full employment could be burdened by the widespread adoption of AI-based technologies that displace workers.¹³⁹ Indeed, there is already evidence that the impressive performance of AI-based diagnostic systems is leading medical students to shy away from entering certain specialty fields, such as radiology, where AI systems routinely outperform humans.¹⁴⁰

Furthermore, the gathering of personal data necessary to create AI-powered tools creates particularly acute privacy risks in the healthcare context. In order to train the algorithms, healthcare providers must collect a vast range of intensely personal health and genetic data. The scope for the misuse of this data is vast—especially since an individual's genetic and health characteristics are often immutable—with potential implications for privacy,¹⁴¹ dignitary rights,¹⁴² freedom from discrimination,¹⁴³ and fair criminal procedure.¹⁴⁴ For example, such data could be used to deny a person health coverage on the basis of genetic factors that are beyond their control.¹⁴⁵ Or such data might be appropriated by the government for law enforcement purposes, as in the recent case from California of a 1970s-era serial killer who was identified based on the statistical analysis of DNA samples that his distant relatives submitted to a family ancestry website.¹⁴⁶

Going further, one can argue that the fundamental right to life may be positively impacted by the introduction of AI diagnostic systems, which hold the promise of not only reducing the rate of diagnostic errors, but making high quality diagnostic services cheaper or more widely available. Although the right to life is generally viewed as a protection against the arbitrary deprivation of life by the

¹³⁷ UDHR art. 23.

¹³⁸ UDHR art. 26.

¹³⁹ UN Committee on Economic, Social and Cultural Rights ("CESCR"), General Comment No. 18: The Right to Work (Art. 6 of the Covenant), 6 February 2006, UN Doc. E/C.12/GC/18.

¹⁴⁰ Thomas H. Davenport and D. O. Keith J. Dreyer, "AI Will Change Radiology, but It Won't Replace Radiologists," *Harvard Business Review*, March 27, 2018, <https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists>.

¹⁴¹ UDHR art. 12.

¹⁴² UDHR art. 1.

¹⁴³ UDHR art. 7.

¹⁴⁴ UDHR art. 10. In particular, it raises questions of self-incrimination, as protected by ICCPR art. 14(3)(g).

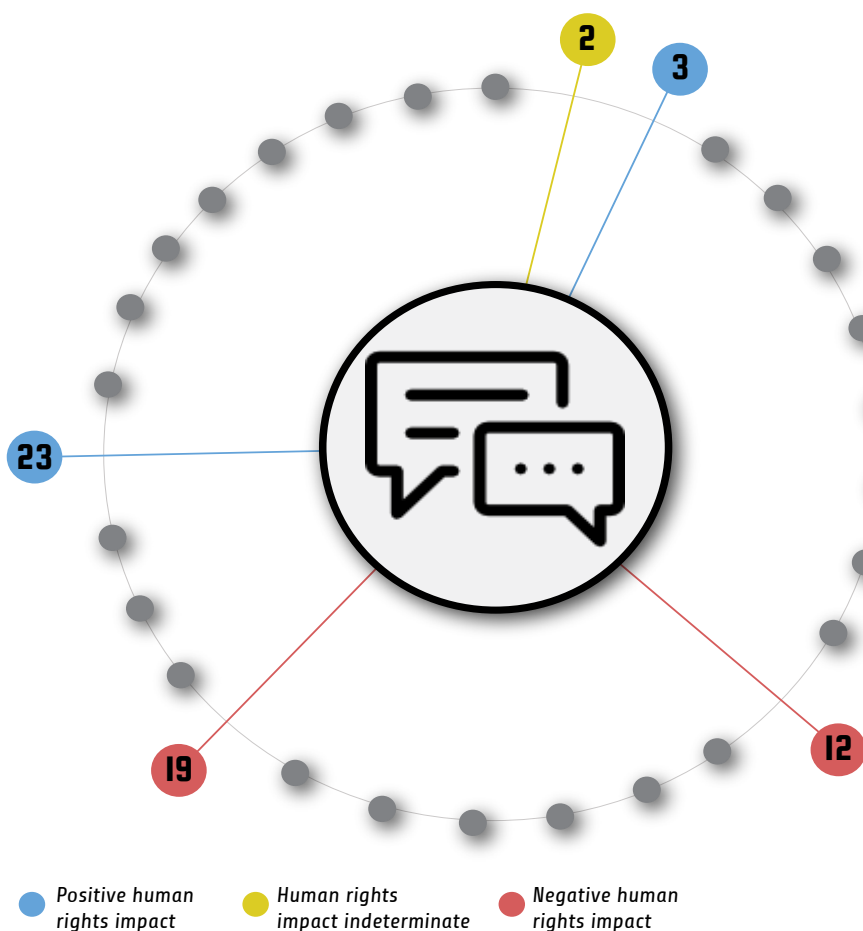
¹⁴⁵ For discussion on attempts to regulate genetic discrimination in the United States, see Louise Slaughter, "Genetic Information Non-Discrimination Act," *Harvard J. on Legislation* 50, no. 1 (2013): 41.

¹⁴⁶ Thomas Fuller, "How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect," *The New York Times*, April 26, 2018, sec. United States, <https://www.nytimes.com/2018/04/26/us/golden-state-killer.html>.

state, the Supreme Court of Canada has ruled that inadequate access to medical care can result in deprivations of the right to life.¹⁴⁷ Correspondingly, improvements in the availability of high-quality medical services can be viewed as enhancing the right to life.

¹⁴⁷ Chaoulli v. Quebec (Attorney General), 2005 SCC 35.

5.4 Online Content Moderation: Standards Enforcement



2. Freedom from Discrimination

AI systems may replicate the biases that human reviewers appear to show in reviewing content posted by marginalized groups, though they could be trained to avoid doing so.

3. Right to Life, Liberty, and Security of Person

AI-based content moderation systems are better than humans at finding content that is per se unlawful, such as child pornography, thereby enhancing community safety.

12. Right to Privacy

Privacy may be impacted by AI content moderation systems that automatically scan non-public communications for material that violates the law or the policies of a platform.

19. Freedom of Opinion, Expression, and Information

Current AI-based content moderation systems have higher error rates than humans, which may lead to large volumes of lawful content being erroneously removed.

23. Right to Desirable Work,

AI-based content moderation systems may free humans from the psychological toll that comes from policing online platforms for graphic, disturbing content.

The sheer amount of information that is available online has mostly been a blessing for humanity, though sometimes it can be a curse. On the one hand, never has so much information about so many different topics been available to most anyone, anywhere, who is fortunate enough to have an Internet connection. On the other hand, the dark side of humanity is also plain to view on the Internet. By virtue of the volume of what is available online, there is a substantial amount of content that is racist, sexist, gruesome, or harmful in other ways—such as by fomenting violence against identifiable groups or targeting individuals for bullying or harassment.

Some of the objectionable content online is subject to regulation by governments in conformity with international human rights law. Article 19(3) of the ICCPR recognizes that the right to free expression may be subject to certain exceptions provided by law that are necessary to protect the rights and reputations of others, or to protect national security, public order, public health, and morals. Moreover, Article 20 of the ICCPR expressly requires states to prohibit “propaganda for war” and the advocacy of “national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.”

Beyond expression that is unlawful and therefore subject to bona fide government regulation, there is also the problem of content that may be lawful

but is nonetheless undesirable—either because it is not being posted in an appropriate place online, or because it violates the standards of an established online community. This is precisely the context in which the private companies that operate the Internet platforms that house so much of the world's online content promulgate standards as to what is acceptable and what is not, and engage in the “virtue of moderating”¹⁴⁸ materials that are inconsistent with those standards.

Traditional Approach to Content Standards Enforcement

The setting and enforcement of content standards by private companies is a controversial topic. Some liken it to a form of censorship due to the burdens it places on the rights to free expression, thought, and association.¹⁴⁹ Indeed, some commentators have noted that the largest online platforms, such as Facebook and Google, exercise more power over our right to free expression than any court, king, or president ever has¹⁵⁰—in view of the very significant percentage of human discourse that occurs within the boundaries of these “walled gardens.”¹⁵¹

By the same token, however, the failure of companies to adequately deal with online content that is harmful to others places its own burden on human rights. Companies therefore face the unenviable challenge of balancing between different rights belonging to different rights-holders, all the while remaining mindful of their responsibility to respect human rights in so doing.¹⁵²

To discharge this difficult task with dispatch while avoiding discriminatory or speech-chilling outcomes, companies communicate their content guidelines to the public on their websites, at the same time as they have developed detailed internal guidance documents for their employees on when particular forms of content are subject to removal.¹⁵³ Ideally, both the external and internal guidelines will be informed by the principles of international human rights law, both with regards to their substantive content and the process that they outline.¹⁵⁴

Until recently, the primary means by which questionable content was brought to the attention of a company was through the efforts of individual users of the platform, who flagged content as unlawful or inappropriate. Human reviewers working for the company then assess the content against the guidelines and determine whether it should stay up

¹⁴⁸ James Grimmelman, “The Virtues of Moderation,” *Yale Journal of Law & Technology* 17, no. 1 (2015): 68.

¹⁴⁹ UDHR arts. 18, 19 and 20.

¹⁵⁰ Jeffrey Rosen, “The Deciders: Facebook, Google, and the Future of Privacy and Free Speech,” in *Constitution 3.0: Freedom and Technological Change*, ed. Jeffrey Rosen and Benjamin Wittes (Brookings Institution Press, 2013).

¹⁵¹ Jonathan Zittrain, *The Future of the Internet and How to Stop It* (New Haven, [Conn.]: Yale University Press, 2008).

¹⁵² Perhaps counter-intuitively, content moderation may be among the AI applications where the distributive effects of these technologies are most apparent. Absent government-established policies, companies will be tasked with choosing which rights and rights-holders are prioritized over others.

¹⁵³ Alexis C. Madrigal, “Inside Facebook’s Fast-Growing Content-Moderation Effort,” *The Atlantic*, February 7, 2018, <https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/>.

¹⁵⁴ Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) (by David Kaye), available at http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

or come down.¹⁵⁵ In view of the massive volume of content that the leading Internet platforms host, these companies now each employ thousands if not tens of thousands of individuals whose sole job is to determine the fate of content that has been flagged.

In recent years, companies have been coming under increased scrutiny both as to the substance of the standards they employ in adjudging content, and also for their decisions in particular cases. For example, Facebook's broader policy against the display of nudity on its platform drew controversy when it removed images of breast-feeding women and the infamous "napalm girl" photograph from the Vietnam War from its platform.¹⁵⁶ Facebook ultimately relented in the face of public pressure in both incidents, but that too raises further questions about the consistency of its application of policies that burden the right to free expression.

There are also growing concerns that company policies on acceptable content may discriminate against certain viewpoints or perspectives, usually in a manner that favors the powerful over the marginalized.¹⁵⁷ For example, Facebook permitted a U.S. Congressman to state his view that all radicalized Muslims should be "hunted" or "killed," whereas it banned activists associated with the Black Lives Matter movement from stating that "all white people are racist."¹⁵⁸ While these anecdotes are not

necessarily indicative of a larger pattern of bias or discrimination, they do raise troubling questions about how well these companies are meeting their responsibility to respect the full range of human rights in this important operational area.

Finally, there is also the issue of how companies are coming under growing pressure from governments to comply with their local laws on a global basis,¹⁵⁹ even when these laws are inconsistent with international guarantees of free expression, the right to association, and other human rights.¹⁶⁰ To date, companies have attempted to blunt these efforts by complying with local laws on a local basis, such as by prophylactically blocking content that is unlawful in a particular jurisdiction while leaving it available everywhere else. A string of recent court decisions has called into question the continuing viability of this technique, however.¹⁶¹

Meanwhile, other governments are even beginning to use company terms of service as a way to act against content that is lawful under domestic and international law, yet undesirable in the eyes of public policymakers. Some in the human rights community have expressed concerns that this activity of "referring" content for removal constitutes an end run around well-established judicial procedures for removing unlawful content.¹⁶²

¹⁵⁵ Brittan Heller, "What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech," *WIRED*, May 2, 2018, <https://www.wired.com/story/what-mark-zuckerberg-gets-wrong-and-right-about-hate-speech/>.

¹⁵⁶ Kate Klonik, "Facebook Erred by Taking down the 'Napalm Girl' Photo. What Happens Next?," *Slate*, September 12, 2016, http://www.slate.com/articles/technology/future_tense/2016/09/facebook_erred_by_taking_down_the_napalm_girl_photo_what_happens_next.html.

¹⁵⁷ Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica*, June 28, 2017, <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

¹⁵⁸ *Ibid.*

¹⁵⁹ Alicia Solow-Niederman et al., "Here, There, or Everywhere?," Berkman Klein Center Working Paper, March 2017, <http://blogs.harvard.edu/cyberlawclinic/files/2017/03/Here-There-or-Everywhere-2017-03-27.pdf>.

¹⁶⁰ For example, see Jens-Henrik Jeppesen & Laura Blanco, "European Policymakers Continue Problematic Crackdown on Undesirable Online Speech," Center for Democracy and Technology (blog), Jan. 18, 2018, <https://cdt.org/blog/european-policymakers-continue-problematic-crackdown-on-undesirable-online-speech/>.

¹⁶¹ For example, *Google Inc. v. Equustek Solutions Inc.* 2017 SCC 34.

¹⁶² Jens-Henrik Jeppesen, "First Report on the EU Hate Speech Code of Conduct shows need for transparency, judicial oversight, and appeals," Center for Democracy and Technology (blog), Dec. 12, 2016, <https://cdt.org/blog/first-report-eu-hate-speech-code-of-conduct-shows-need-trans>.

AI-Assisted Content Standards Enforcement

As the volume of online content in need of moderation grows inexorably and exponentially, the major online platforms are making significant investments in developing AI systems to automate this task. One major impetus for so doing are recently-enacted laws that require companies to promptly remove content that violates national laws, at the risk of facing substantial penalties for noncompliance.¹⁶³ These technologies are still in their infancy, and most simply work to identify potentially problematic content for a human reviewer to evaluate. That being said, fully automated content removal systems have been used against content that is suspected of violating copyright for a number of years,¹⁶⁴ and there are indications that some Internet platforms are employing fully automated content review and removal systems for at least some purposes.¹⁶⁵

The current generation of AI-based content review and removal systems is built on natural language processing (“NLP”) technology. As things stand right now, NLP technologies are domain-specific: that is to say, they were only built to identify the particular types of content on which they were trained and nothing else.¹⁶⁶ Hence an NLP system that is trained to detect say, racist speech, is incapable of detecting violent content. What is more, even

within a particular domain, NLP technologies are not sophisticated enough to understand all of the nuances of human speech. A system that is able to detect racist content in a blog post might not accurately identify such content in a tweet, which results in these technologies having a very substantial error rate. This has led skeptics, including Facebook CEO Mark Zuckerberg, to conclude that AI systems are not yet sophisticated enough to replace human reviewers.¹⁶⁷

This, however, does not render AI technologies useless. The speed at which they can sift through content makes them a powerful tool to assist, rather than to replace, human reviewers by identifying content that appears to be suspect.¹⁶⁸ AI systems can also be used to study the evolution of hate speech to spot emerging trends, as the Anti-Defamation League is currently doing with its Online Hate Index.¹⁶⁹

Summary of Impacts

Due to the higher error rates of existing AI-based content flagging systems as compared to human reviewers,¹⁷⁰ the use of these systems to automatically remove content that is suspected to violate the law or an online platform’s community standards is likely to have a negative impact on the rights to free

[parency-judicial-oversight-appeals/](#).

¹⁶³ Yascha Mounk, “Verboten,” *The New Republic*, April 3, 2018, <https://newrepublic.com/article/147364/verboten-germany-law-stop-ping-hate-speech-facebook-twitter>.

¹⁶⁴ “How Content ID Works - YouTube Help,” accessed June 21, 2018, <https://support.google.com/youtube/answer/2797370?hl=en>.

¹⁶⁵ Lizzie Dearden, “New Technology Can Detect ISIS Videos before They Are Uploaded,” *The Independent*, February 12, 2018, <http://www.independent.co.uk/news/uk/home-news/isis-videos-artificial-intelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html>.

¹⁶⁶ Natasha Duarte, Emma Llanos, and Anna Loup, “Mixed Messages? The Limits of Automated Social Media Content Analysis,” Center for Democracy & Technology (blog), November 2017, <https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/>.

¹⁶⁷ Heller, “What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech.”

¹⁶⁸ Susan Wojcicki, “Expanding our work against abuse of our platform,” YouTube Official Blog, Dec. 4, 2017, <https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html>.

¹⁶⁹ “The Online Hate Index,” Anti-Defamation League, accessed June 18, 2018, <https://www.adl.org/resources/reports/the-online-hate-index>.

¹⁷⁰ Duarte, Llanos, and Loup, “Mixed Messages?”

expression, opinion, and information.¹⁷¹ This is because such systems are likely to remove a significant volume of content that is lawful and consistent with the platform's own community standards. Such errors would deprive individuals of the opportunity to express themselves, and their audience from viewing the opinions those individuals have expressed, with few opportunities for recourse.¹⁷² That said, future developments in AI could well have a positive impact on these rights if they result in a lower error rate than the systems and procedures that are currently in use.

In their current state, these systems have the potential to positively impact the rights to life, liberty, and security of person¹⁷³ by improving the detection and removal of content that incites terrorism or hatred or violence against vulnerable populations. For example, automated techniques have been quite effective at detecting child pornography with a low error rate, and some forms of terrorist content exhibit consistent patterns that facilitate their detection by training a machine learning algorithm, by contrast to the fast-evolving and always-subjective nature of hate speech.¹⁷⁴

The net impact of these systems on the right to be free from discrimination¹⁷⁵ is indeterminate. As with free expression, AI systems could be trained to avoid the biases exhibited by human reviewers, but on the other hand there is a considerable risk that machine learning techniques will result in the replication and scaling of existing human patterns of bias into new automated content review systems.¹⁷⁶

One additional right that merits discussion in this context is that of the individuals employed in content review and moderation positions to just and favorable conditions of work.¹⁷⁷ The psychological toll that the frontline work of content review and moderation takes is considerable, as these individuals are exposed to the very worst of humanity day in and day out—from child pornography to gruesome acts of violence. Content reviewers are disproportionately female, but reviewers of all genders suffer from depression, burnout, anxiety, sleep difficulties, and even from post-traumatic stress disorder at extraordinary rates.¹⁷⁸ Using AI to lessen the psychological burden associated with this work could well have positive human rights impacts on a group of individuals who are often forgotten in conversations about how best to respond to problematic content online.

¹⁷¹ UDHR art. 19.

¹⁷² For example, it is only in April of this year that Facebook announced that it was creating a system by which its users could appeal content removal decisions that they believe to be in error. Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Facebook (official blog), April 24, 2018, <https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/>.

¹⁷³ UDHR art. 3.

¹⁷⁴ Hanna Kozłowska, "Facebook Is Revealing Data on How Good It Is at Moderating Content, but the Numbers Have Holes," *Quartz*, May 18, 2018, <https://qz.com/1277729/facebook-is-revealing-data-on-how-good-it-is-at-moderating-content-but-the-numbers-dont-say-much/>.

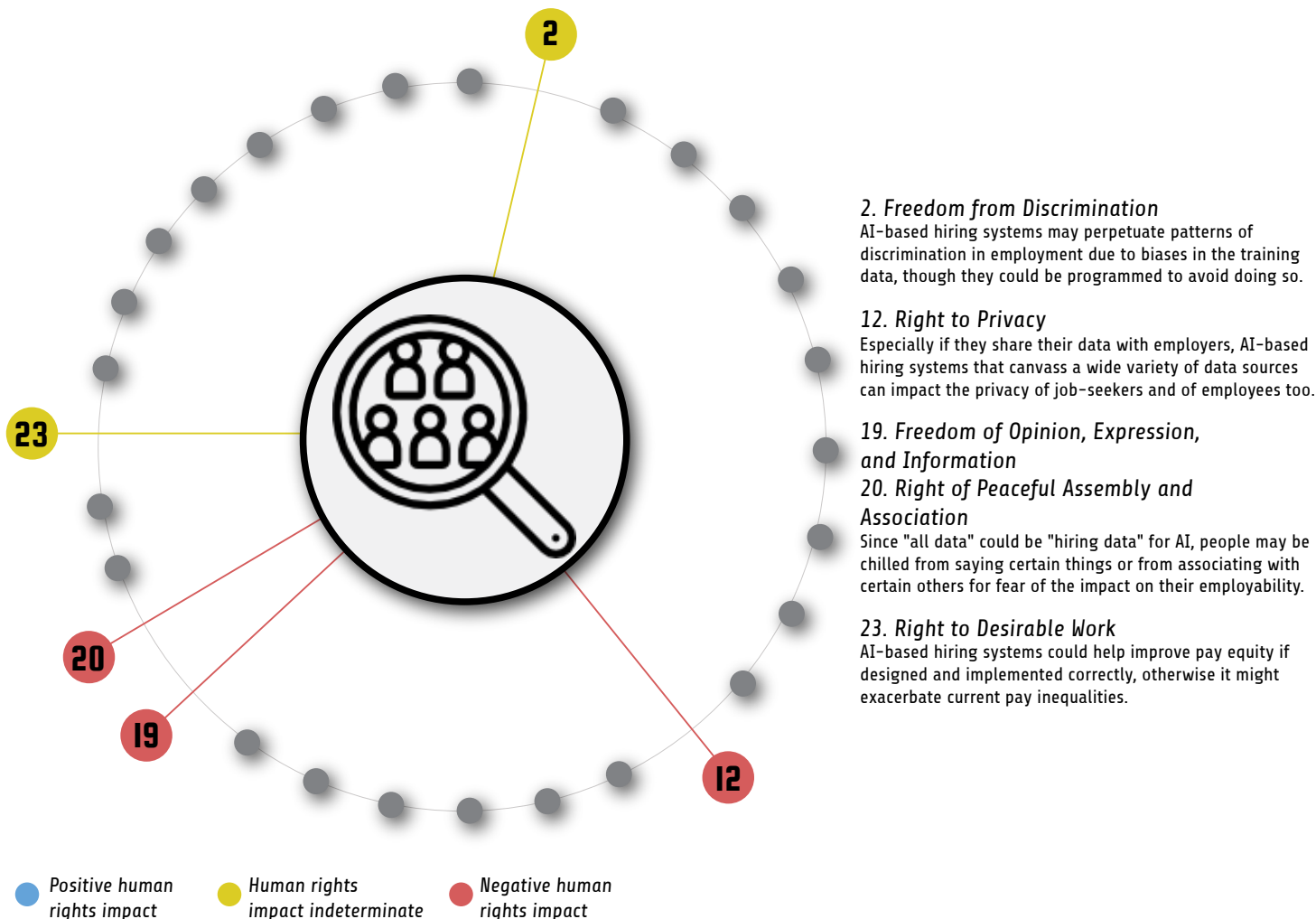
¹⁷⁵ UDHR art. 2.

¹⁷⁶ Reuben Binns et al., "Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation," *ArXiv:1707.01477 [Cs]* 10540 (2017): 405–15, <https://doi.org/10.1007/978-3-319-67256-4>.

¹⁷⁷ UDHR art. 23.

¹⁷⁸ Andrew Arshat and Daniel Etcovitch, "The Human Cost of Online Content Moderation," *Harvard Journal of Law & Technology Digest*, March 2, 2018, <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation>.

5.5 Human Resources: Recruitment and Hiring



Organizations cannot operate without employees, but the process of identifying, recruiting, and hiring new employees is hard work. Increasingly, public- and private-sector employers are turning to AI to help with the hiring process for at least two reasons.¹⁷⁹ The first is capacity: the number of applicants per position has multiplied in the last several years, while staffing levels at human resources ("HR") departments remain flat. The second is fairness: there is a growing awareness that hiring pro-

cesses are rife with implicit bias and discrimination, and that hiring decisions often boil down to "is this person like me?" Many organizations believe that AI may offer at least a partial solution to this challenge.

The responsibility of business to respect human rights applies not just to the services they provide and the products they sell, but also to their internal operations. Flawed hiring processes may have significant implications for the right to freedom from

¹⁷⁹ "A New Age of Opportunities: What Does Artificial Intelligence Mean for HR Professionals?" (Ontario: Human Resources Professionals Association, 2017).

discrimination,¹⁸⁰ the right to equal pay for equal work,¹⁸¹ and the rights to freedom of expression and association.¹⁸² Governments have recognized the need for mechanisms to provide remedy for individuals subjected to discriminatory hiring practices and have created institutions such as the U.S. Equal Employment Opportunity Commission (“EEOC”) and the Canadian Human Rights Commission. As AI-based hiring systems become commonplace, it will be important to evaluate whether these existing mechanisms are up to the task of ensuring that these new technologies are free from bias.

Traditional Approach to Recruitment and Hiring

Recruiters have long relied on technology to streamline the hiring process. Today, HR departments commonly use applicant tracking systems (“ATS”) to aggregate applicant information and filter them based on certain criteria, such as years of experience, education, or other keywords. Short-listed candidates are then interviewed and a decision is made on who to hire. Few quantitative data points are used in this process; instead, it relies on an often-flawed combination of pedigree and gut instinct.

Problems abound in this human-based system. There has been much debate about why we continue to see men from the majority group hired and

promoted over women and minorities. Some claim that it may reflect systemic inequities in society leading to men being more highly educated and better prepared for a particular job.¹⁸³ But research shows that much of the problem is discrimination resulting from implicit biases that manifest themselves in individual decision-making. In fact, our very definitions of success can exhibit bias. One experimental study shows that individuals are prone to shifting their definition of merit when evaluating applicants to advantage certain groups, which plays a role in gender and racial discrimination.¹⁸⁴ This distinction is important because, while systemic inequity is a serious problem, decisions marred by individual bias more directly implicate the right to be free from discrimination.¹⁸⁵

Research in the United States has shown, repeatedly, that “white-sounding” names receive 50% more callbacks than “African-American-sounding” names—despite otherwise identical resumes.¹⁸⁶ Moreover, males often receive more callbacks for traditionally “male” jobs. In one experiment designed to explore the dearth of women in STEM professions, researchers found that women were half as likely as men to be hired for a job, based on a flawed assumption that the female candidates performed worse on an arithmetic task. This was the case even when the women actually performed better than their male counterparts. The reason

¹⁸⁰ UDHR art. 7.

¹⁸¹ UDHR art. 23(2).

¹⁸² UDHR art. 20.

¹⁸³ Matthew Scherer, “AI in HR: Civil Rights Implications of Employers’ Use of Artificial Intelligence and Big Data,” *SciTech Lawyer* 13 (2017 2016): 12-16.

¹⁸⁴ Eric Luis Uhlmann and Geoffrey L. Cohen, “Constructed Criteria: Redefining Merit to Justify Discrimination,” *Psychological Science* 16, no. 6 (June 1, 2005): 474-80, <https://www.ncbi.nlm.nih.gov/pubmed/15943674>.

¹⁸⁵ UDHR art. 2.

¹⁸⁶ Marianne Bertrand and Sendhil Mullainathan, “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination” (Cambridge, MA: National Bureau of Economic Research, July 2003).

for this was that men were more likely to inflate their performance on the task in an interview and women were more likely to underestimate their performance.¹⁸⁷ Furthermore, the right to equal pay for equal work¹⁸⁸ is implicated by the fact that men receive higher starting salary offers than women for the same job.¹⁸⁹

Finally, the right to freedom of association¹⁹⁰ is implicated by human bias in the hiring process. Involvement in certain organizations, such as ethnic affinity groups or LGBTQ networks, can negatively impact job prospects. A résumé audit study found that women with a leadership role in a student LGBTQ organization received 30% fewer callbacks for a job posting than applicants with an identical resume but without the LGBTQ association.¹⁹¹ Individuals may not know that this information is limiting their job prospects, but if they did, they might feel pressure to disassociate themselves from controversial groups.

Ultimately, the impact of AI hiring systems on human rights will depend, in part, on whether the controls meant to mitigate or to remedy rights-related harms in the existing human-based system

can be applied to the new technology. It is to the technology now that we must turn in order to evaluate this question.

AI Assisted Hiring and Recruitment

Artificial intelligence is now being used to augment much of the hiring process. Job descriptions can be run through text analysis software to flag gendered language that might discourage highly qualified women from applying.¹⁹² Companies can enlist algorithms to advertise openings to eligible candidates through LinkedIn or Google's ad network.¹⁹³ AI is also being used to screen applicants using natural language processing to parse resumes.¹⁹⁴ Some technologies even draw on social media and other public data to supplement their analyses. After AI is used to narrow down the applicant pool, companies might invite candidates to conduct recorded interviews, where algorithms evaluate word choice, vocal inflection, and even emotions (using facial recognition).¹⁹⁵ These technologies purport to identify candidate "personalities" and help establish "fit" within the company. AI is clearly streamlining the hiring process, but the verdict on AI's ability to mitigate negative human rights impacts is unclear.

¹⁸⁷ Ernesto Reuben, Paola Sapienza, and Luigi Zingales, "How Stereotypes Impair Women's Careers in Science," *Proceedings of the National Academy of Sciences*, March 5, 2014, <http://www.pnas.org/content/111/12/4403>.

¹⁸⁸ UDHR art. 23(2)

¹⁸⁹ "2018 State of Wage Inequality in the Workplace Report," *Hired*, accessed June 21, 2018, <https://hired.com/wage-inequality-report>.

¹⁹⁰ UDHR art. 20.

¹⁹¹ Emma Mishel, "Discrimination against Queer Women in the U.S. Workforce: A Résumé Audit Study," *Socius* 2 (January 1, 2016). <https://doi.org/10.1177/2378023115621316>.

¹⁹² Software employed by Textio and Gender Decoder use NLP paired with research on language pattern and implicit word associations to make job descriptions more gender neutral. See Textio, <https://textio.com/>, and <http://gender-decoder.katmatfield.com/>.

¹⁹³ John Jersin, "How LinkedIn Uses Automation and AI to Power Recruiting Tools," *LinkedIn Talent Blog* (blog), October 10, 2017, <https://business.linkedin.com/talent-solutions/blog/product-updates/2017/how-linkedin-uses-automation-and-ai-to-power-recruiting-tools>.

¹⁹⁴ See examples "Sovren," accessed June 20, 2018, <https://www.sovren.com/>, and "Textkernel Launches the First Fully Deep Learning Powered CV Parser," *Textkernel* (blog), February 8, 2018, <https://www.textkernel.com/extract-4-o-textkernel-launches-the-first-fully-deep-learning-powered-cv-parsing-solution/>.

¹⁹⁵ HireVue.com, "HireVue: Video Interview Software for Recruiting & Hiring," accessed June 20, 2018, <https://www.hirevue.com>.

Previous sections have noted how the veneer of objectivity that technology provides can be dangerous, because it obscures how AI often replicates human biases at scale. This is particularly worrying when AI is used to devise predictors of success that will determine hiring and advancement opportunities for future applicants and employees. There is already evidence that gender stereotypes have seeped into the “word embedding frameworks”¹⁹⁶ used in many machine learning and natural language processing technologies. One of the more egregious cases revealed in a 2016 study found that an algorithm trained on Google News articles to understand word meanings would respond to the query “man is to computer programmer as woman is to x” with x = homemaker.¹⁹⁷ In view of this example, there is a very real danger that an AI-based hiring algorithm trained on performance reviews,¹⁹⁸ employee surveys, and other data points meant to uncover the attributes of successful employees will reproduce existing patterns of bias in future hiring decisions. The system may produce more consistent results across candidates than human hiring managers, but the outputs of such a system can hardly be described as fair.¹⁹⁹

Summary of Human Rights Impacts

With the foregoing in mind, the question of how machine learning technologies used in hiring will impact the right to be free from discrimination becomes more complicated.²⁰⁰ Without intentional intervention in the programming, it seems likely that AI will reproduce the existing systemic patterns of bias and prejudice exhibited in the training data. This may lead AI-based hiring systems to identify metrics for assessing candidates that reflect structural biases rather than the objective determinants of real-world employment performance.

Some technologists and researchers have identified this as a concern and are devising technical solutions. One solution proposes a decoupling technique²⁰¹ that, in the resume screening context, would allow an algorithm to identify top candidates using variables optimized based on other applicants of a certain category (e.g. race or gender) rather than against the entire applicant pool. In practice, this means the traits to select for a female or minority applicant would be identified based on the trends of other female or minority applicants, and these could differ from the identified successful traits of

196 Tolga Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” ArXiv:1607.06520 [Cs, Stat], July 21, 2016, <http://arxiv.org/abs/1607.06520>.

197 Word embedding is a set of language and feature modeling techniques used in NLP to map words or phrases onto vectors of real numbers. This allows computer programs to understand and use word meaning, see Tolga Bolukbasi et al., “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings,” ArXiv:1607.06520 [Cs, Stat], July 21, 2016, <http://arxiv.org/abs/1607.06520>.

198 Performance reviews and promotion structures often exhibit gender and racial bias, which contributes to the paucity of women and minorities as you move up the corporate ladder. See Kieran Snyder, “The Abrasiveness Trap: High-Achieving Men and Women Are Described Differently in Reviews,” *Fortune*, August 26, 2014, <http://fortune.com/2014/08/26/performance-review-gender-bias/>, and Buck Gee and Janet Wong, “Lost in Aggregation: The Asian Reflection in the Glass Ceiling” (The Ascend Foundation, September 2016), and Hannah Riley Bowles, Linda Babcock, and Lei Lai, “Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask,” *Organizational Behavior and Human Decision Processes* 103, no. 1 (May 2007): 84–103, <https://doi.org/10.1016/j.obhdp.2006.09.001>.

199 Companies like Koru are using this method to establish their client organizations’ “predictive hiring fingerprint” and are claiming that it reduces bias and the proclivity to hire based on pedigree. <https://www.joinkoru.com/>.

200 UDHR art. 2.

201 Cynthia Dwork et al., “Decoupled Classifiers for Fair and Efficient Machine Learning,” ArXiv:1707.06613 [Cs], July 20, 2017, <http://arxiv.org/abs/1707.06613>.

a male or majority applicant. The feasibility—and legality—of implementing a technical solution that optimizes fairness by distinguishing between individuals on sensitive personal characteristics attributes is highly dependent on jurisdiction.

More hope lies in companies intentionally designing algorithms to control for human biases and implementing auditing systems that can regularly test for bias and errors. Applied and Pymetrics, for example, have worked with academics to devise AI-based hiring systems that use anonymization, skills testing, work product analysis, neurological brain games, and other methods to remove bias based on gender, race, and pedigree.²⁰² While these efforts are promising, their outcomes still depend on the reliability of their algorithms and the level of bias in their data. Where AI is being used in the hiring process, it will be important to implement robust auditing structures to regularly examine biases in

the data and how these are influencing the outputs.

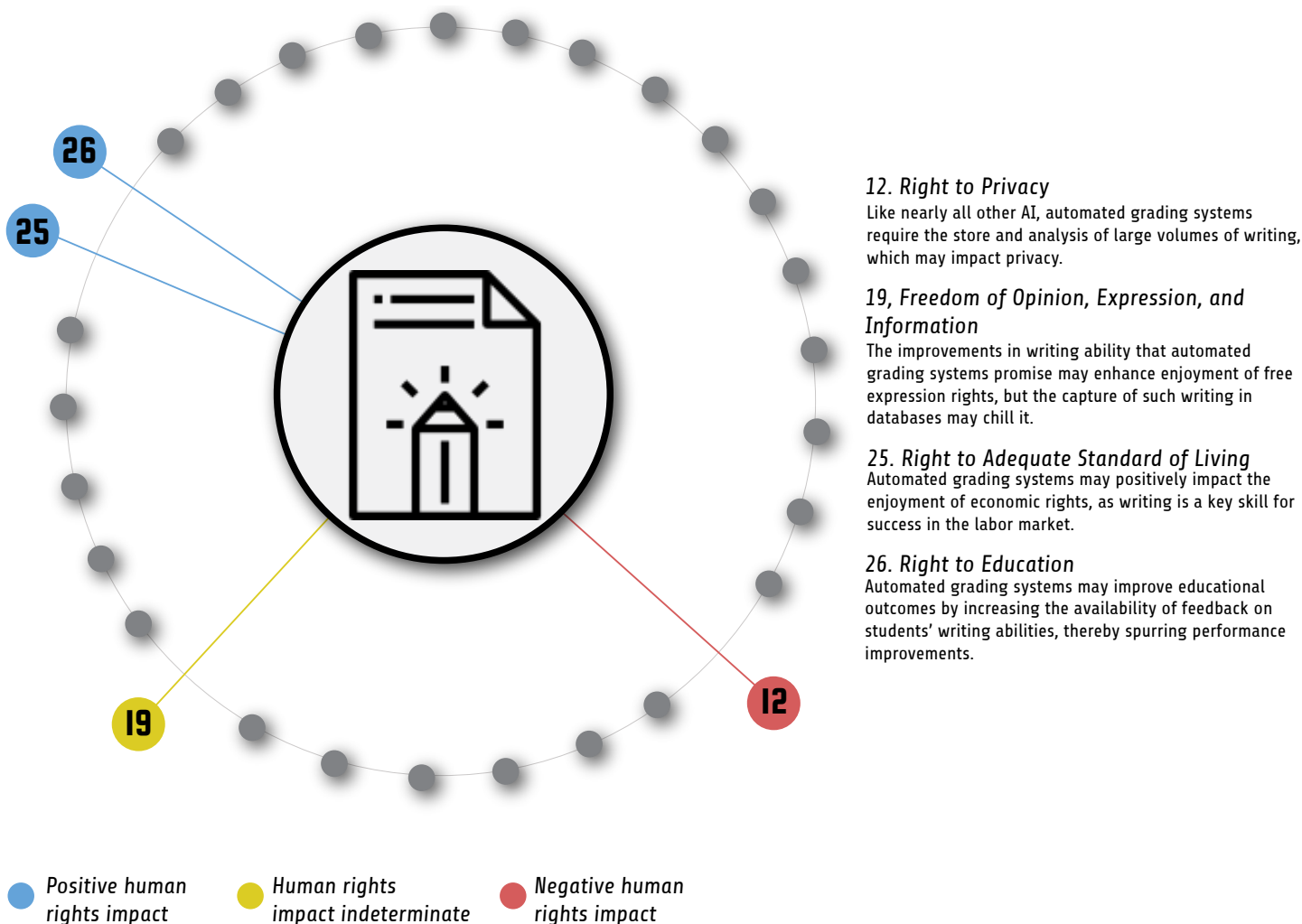
AI-based hiring systems may have a greater negative impact on the freedoms of association²⁰³ and expression²⁰⁴ than the current human-based system. Similar to the concern that people will carefully curate their associations to maximize their credit scores, there is a risk that applicants will feel compelled to disassociate themselves from organizations that might hurt their chances of securing employment, because AI-based systems are more likely to detect such associations than human recruiters. Likewise, AI-based hiring systems may chill individuals from engaging in certain forms of expressive activity out of fear that their words will be used against them in the employment context.

²⁰² Applied, “Can We Predict Applicant Performance without Requiring CVs? Putting Applied to the Test — Part 1,” *Medium* (blog), September 21, 2016, <https://medium.com/finding-needles-in-haystacks/putting-applied-to-the-test-part-1-9fad6379e9e>; Josh Constance, “Pymetrics Attacks Discrimination in Hiring with AI and Recruiting Games,” *TechCrunch* (blog), accessed February 22, 2018, <http://social.techcrunch.com/2017/09/20/unbiased-hiring/>.

²⁰³ UDHR art. 20.

²⁰⁴ UDHR art. 19.

5.6 Education: Essay Scoring



Access to education is a right in and of itself and a key enabler of a panoply of other human rights. Educational attainment is the primary engine of social mobility;²⁰⁵ those who are more educated are better able to participate in the economy, engage in civic and public life, and improve their personal and local circumstances.

One of the key skills education systems around the world seek to develop in their students is the ability to write. Not only is the quality of one's writing an

important factor in the job market, but it is also an important ability for achieving the "full development of the human personality" in the words of the Universal Declaration of Human Rights,²⁰⁶ as well as one of the primary means through which the right to free expression is exercised.

The importance of writing is perhaps why this skill is so frequently tested at every level of the education system. Most countries periodically administer standardized tests of their students' written abilities

²⁰⁵ Michael Greenstone et al., "Thirteen Economic Facts about Social Mobility and the Role of Education" (The Hamilton Project, June 26, 2013).

²⁰⁶ UDHR art. 26(2).

in the local *lingua franca* as they progress through the educational system. In many countries, students must pass writing tests to graduate from a particular level of schooling, or for admission to the next level. In North America, for example, both the Graduate Management Administrative Test (“GMAT”) and the Graduate Record Examinations (“GRE”) evaluate the ability of prospective management and law students to produce an analytical writing sample under time constraints.

How writing is marked matters. In the day-to-day school context, the quality of the feedback students receive on their writing will impact the prospects of student improvement over time. And in the context of gatekeeping exams such as the GMAT and the GRE, how test-takers’ writing is evaluated may have life-long impacts on their future opportunities. Consequently, as automated techniques are increasingly being used to evaluate student writing, the question arises of what their impacts will be.

Traditional Approach to Essay Grading

The traditional approach to grading writing, whether in the ordinary schooling or the standardized testing context, is for trained individuals to perform this task. When a large volume of writing needs to be evaluated against an established performance standard, a common approach is for the individuals responsible for the grading to each evaluate a

representative sample of what has been submitted, and to compare results in order to calibrate their approach for the rest of the grading.

In most contexts, evaluating the quality of writing requires not just considering the mechanics of grammar and syntax, but also evaluating the writing for the accuracy of the substance and on considerations such as style and rhetorical impact. Even so, in the context of one common standardized test, a study found that the score assigned to the writing component could be predicted accurately 90% of the time by considering just one variable: length.²⁰⁷

Although educators may have more time to engage with the style of the substance of a student’s writing than the evaluator of a standardized test, the resource constraints under which most school systems operate may lead teachers to turn, intentionally or not, to more mechanical assessments of their students’ writing. Such assessments may fail to provide the feedback students need for growth and improvement.²⁰⁸ Indeed, studies have found that educators often emphasize form over substance in teaching writing, which may well result in students prioritizing form over substance in their own writing.²⁰⁹

207 Arthur C. Graesser and Danielle S. McNamara, “Automated Analysis of Essays and Open-Ended Verbal Responses,” in *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics.*, ed. Harris Cooper et al. (Washington: American Psychological Association, 2012), 307–25, <https://doi.org/10.1037/13619-017>; Michael Winerip, “SAT Essay Test Rewards Length and Ignores Errors,” *The New York Times*, May 4, 2005, sec. Education, <https://www.nytimes.com/2005/05/04/education/sat-essay-test-rewards-length-and-ignores-errors.html>.

208 Deborah Reck and Deb Sabin, “A High-Tech Solution to the Writing Crisis,” *The Atlantic*, October 16, 2012, <https://www.theatlantic.com/national/archive/2012/10/a-high-tech-solution-to-the-writing-crisis/263675/>.

209 Gary A. Troia and Steve Graham, “Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know,” *Journal of Education and Psychological Consultation* 14, no. 1 (2003): 75–89.

AI-Graded Essays

Following decades of research and many fruitless attempts, automated essay scoring has recently become a reality. Indeed, these technologies are among the more mature current applications of artificial intelligence. Using machine learning, these systems are trained to grade written materials by ingesting a set of exemplars that have been graded by human experts. These systems identify features in the set of training materials that correlate with success, and they then assess any written materials that are fed into the system according to what they have learned.²¹⁰

Such automated systems are already in use in high-stakes standardized testing environments. For example, the written component of the GMAT exam is currently scored by an automated system and an expert human grader working separately. Should the AI and the human differ in the grade they assign by more than one point, a second human grader acts as a tiebreaker.²¹¹

The use of automated grading systems has the potential to positively impact the right to education in a number of different ways. Automated systems permit students to engage in more deliberate practice of their writing, receiving more feedback on at least

some elements of their writing than they would otherwise. In the context of communities without reliable access to quality writing instruction, whether due to poverty or other forms of marginalization, automated systems may be one of the only reasonably available means to obtain the feedback required to develop one's writing skill. Some studies have demonstrated that the instantaneous feedback of rudimentary grammar checkers have powerful effects on the quality of written expression, so it is reasonable to assume that the same is true of more nuanced AI-based approaches to the same basic endeavor.²¹²

Relatedly, automating certain aspects of the grading of writing might free educators to spend more time focusing on higher-order teaching tasks, such as engaging with students' ideas and arguments. To many, writing is more than sentence structure and word-choice: it is the expression of ideas and emotions, drawing on cognitive skills distinct from those underpinning grammar and syntax.²¹³ Moreover, automated systems have the potential to eliminate bias in grading by removing the opinions of the author and the grader and any relationship between them from the equation.²¹⁴

On the flipside, there are serious concerns relating to the fact that these systems cannot understand

210 Corey Palmero, "A Gentle Introduction to Automated Scoring.Pdf" (Measurement Incorporated, October 2017), <http://www.measurementinc.com/sites/default/files/2017-10/A%20Gentle%20Introduction%20to%20Automated%20Scoring.pdf>.

211 "How the Analytical Writing Assessment Is Scored," Economist GMAT Tutor, April 28, 2017, <https://gmat.economist.com/gmat-advice/gmat-overview/gmat-scoring/how-analytical-writing-assessment-scored/>.

212 Paul Morphy and Steve Graham, "Word Processing Programs and Weaker Writers/Readers: A Meta-Analysis of Research Findings," *Reading and Writing* 25, no. 3 (March 2012): 641–78, <https://doi.org/10.1007/s11455-010-9292-5>.

213 Troia and Graham, "Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know."

214 For examples of bias in writing grading, see John Malouff, "Bias in Grading," *College Teaching* 56, no. 3 (July 2008): 191–92, <https://doi.org/10.3200/CTCH.56.3.191-192>. Bias may also come in the form of pre-existing stereotypes influencing how critically one reviews the essay. For an example in the legal profession, see Arin N. Reeves, "Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills," *Yellow Paper* (Nextions, 2014).

what is written in the same way as human readers.²¹⁵ Some systems might well be able to detect offensive content, but at least for the foreseeable future, artificial intelligence systems will not realistically possess the general intelligence that humans do which enables them to evaluate the validity of written material.²¹⁶ Especially in our era where the truth and accuracy of written materials has become an issue of the utmost public importance, the likely inability of automated grading systems to assess factual validity is of concern.

Furthermore, there are also significant concerns about what incentives these systems create for those who are subject to being evaluated by them.²¹⁷ Consider, for example, that a famous essay by the renowned MIT linguist Noam Chomsky received a grade of “fair” when it was fed into an automated grading system.²¹⁸ If students respond to the growing prevalence of automated grading systems by focusing on form and length to the detriment of style and substance, these technologies may be doing them a disservice.

Finally, automated grading systems depend on the collection, storage, and analysis of vast quantities of written material. This raises not only the standard privacy-related concerns that accompany most AI systems,²¹⁹ but there is an additional risk that these systems might chill the full enjoyment of the right to free expression. Even in the educational context, a student might be more willing to share writing about something that is deeply personal or controversial with a trusted teacher as opposed to an AI system, if it is going to end up being catalogued in a vast database and subject to being used as training data for some other purpose.

Summary of Human Rights Impacts

On balance, the rise of automated grading systems is likely to have a positive impact on the right to education, as these systems can potentially increase global access to at least some feedback on people's writing. In much of the world today, educational systems are simply too overburdened to provide the kind of individualized evaluation and feedback on writing that is desirable, so the potential of these technologies to improve the situation over the sta-

215 Stephen P. Balfour, “Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review,” *Research & Practice in Assessment* 8, no. 1 (2013): 40–48. For an example of the concerns, see Kai Riemer, “On Rewarding ‘Bullshit’: Algorithms Should Not Be Grading Essays,” *Undark*, accessed April 21, 2018, <https://undark.org/article/rewarding-bullshit-algorithms-classroom/>.

216 For example, Dr. Les Perelman of MIT and his team developed a piece of software they call the “Basic Automatic B.S. Essay Language Generator” (BABEL) which, upon the input by the user of three words, auto-generates essays that routinely receive near-perfect scores from various automated grading systems that are currently in use, although their content is nothing but machine-generated nonsense. See Steve Kolowich, “Writing Instructor, Skeptical of Automated Grading, Pits Machine vs. Machine,” *Chronicle of Higher Education*, April 28, 2014, <https://www.chronicle.com/article/Writing-Instructor-Skeptical/146211>. Dr. Perelman's BABEL system is available to try out at <http://lesperelman.com/>. To be sure, humans are not immune to being fooled by machine-generated gibberish. See Danielle Wiener-Bronner, “More Computer-Generated Nonsense Papers Pulled From Science Journals,” *The Atlantic*, Mar. 3, 2014, <https://www.theatlantic.com/technology/archive/2014/03/more-computer-generated-nonsense-papers-pulled-science-journals/358735/>.

217 Jinhao Wang and Michelle Stallone Brown, “Automated Essay Scoring Versus Human Scoring: A Correlational Study,” *Contemporary Issues in Technology and Teacher Education* 8, no. 4 (2008): 16.

218 “Parts of Noam Chomsky's Essay ‘The Responsibility of Intellectuals’ Grammar Checked by ETS's Criterion and WhiteSmoke | Les Perelman, Ph.D.,” accessed June 21, 2018, <http://lesperelman.com/writing-assessment-robo-grading/parts-noam-chomskys-essay-grammar-checked/>.

219 Related privacy concerns have been raised about automated plagiarism-detection software, which also require the storage and retention of vast quantities of student-written material to be effective. See Bo Brinkman, “An Analysis of Student Privacy Rights in the Use of Plagiarism Detection Systems,” *Science and Engineering Ethics* 19, no. 3 (2013): 1255–1266.

tus quo is considerable. The feedback these systems provide might fall short of the Platonic ideal of individualized attention from expert human instructors, but they are a step in the right direction for the vast majority of students worldwide who lack affordable access to high-quality instruction.

Considering that the ability to write is a key enabler of a panoply of civil and political rights, from that of free expression to the right to take part in cultural and scientific life, the improvements that automated grading systems will make in individuals' ability to write when judged on a global basis is likely to improve their ability to enjoy these rights. This is especially true of the crucial right that everyone possesses to an adequate standard of living, as the

ability to write is so central to one's employment prospects and ability to participate effectively in various aspects of society.

The impact of these automated systems on the right to free expression, however, is more complex. On one hand, anything that improves the ability of people to express themselves in an effective manner would seem to positively impact the enjoyment of this right. On the other hand, as noted above, large-scale collection of written materials that automated systems necessarily entail may chill some individuals from setting down in writing things that they might have said otherwise.

6. Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach

The six use cases explored in the previous section demonstrate how artificial intelligence has the potential to positively impact the full range of human rights. Automating complex tasks that currently require the labor of highly trained professionals could well usher in greater access to specialized healthcare, education, and financial services. Such technologies also have the considerable potential to reduce and correct for various biases that plague human decision-making, from outright discrimination to our reliance on heuristics that sometimes lead us astray. Yet this promise comes at an almost inevitable cost to our privacy due to the data-intensive nature of these technologies which, in turn, may chill the exercise by many of their civil and political rights. Likewise, the possibility that these technologies will reproduce and ossify existing patterns of discrimination and bias, while also producing troubling distributive consequences, must be contended with.

This conundrum gives rise to the question of how we can enjoy the benefits of artificial intelligence—especially the vast potential for positive impacts on human rights—while minimizing its real negative risks.

This is not a new question, but rather one that has arisen with every major technological innovation throughout history. From the development of industrial machinery to the invention of the automobile, transformative technological changes have posed a profound challenge to the existing social

order. These technologies utterly transformed society in their time—oftentimes for the better, yet they were frequently accompanied by bad consequences, too. Industrialization, for example, democratized the availability of goods that were once luxuries, though at the cost of widespread economic displacement, Dickensian working conditions, suffocating air pollution, and colonial patterns of natural resource exploitation. Likewise, the automobile revolutionized human mobility and fundamentally transformed the economy, though with the negative consequences of air pollution, urban sprawl, and millions of traffic casualties every year.

The negative impacts of these and other transformative technologies were felt most acutely as they were first coming into widespread use. Over time, however, society responded by developing control mechanisms to attempt to enjoy the good while minimizing the bad. Some controls are regulatory in nature, such as laws and norms that specify how and when a technology might be used, while others are technological, such as design features that channel a technology towards certain uses and away from others.²²⁰ Oftentimes, controls are a mix of the two, such as regulatory standards that mandate specific design characteristics.

History shows that it can take quite some time to develop effective mechanisms to control new technologies. The Industrial Revolution began to transform Britain in the 18th century, but it was only in the mid-19th century that Parliament started enact-

²²⁰ In the U.S. context, Lawrence Lessig famously noted that the “East Coast Code” of laws and regulations promulgated in Washington D.C., and the “West Coast Code” produced by software engineers in Silicon Valley, are both fundamentally forms of regulation. Lawrence Lessig, *Code*, Version 2.0 (New York: Basic Books, 2006), <http://codev2.cc/download+remix/Lessig-Codev2.pdf>.

ing legislation to address its consequences.²²¹ Likewise, although automobiles became commonplace in the first decades of the 20th century, the first systematic studies of vehicular safety took place in the 1940s,²²² and the first comprehensive automobile safety laws weren't enacted until the 1960s.²²³

Of course, not all of these controls are effective or ideal. Most, if not all of them, are at least partially flawed. Some are too permissive to adequately address the negative consequences of the regulated technology, while others are too restrictive to permit the realization of its benefits. All the same, we must think about which controls are necessary, sufficient, and appropriate to reduce and redress the human rights impacts of artificial intelligence.

We are fortunate to be in a position to design the regulatory and technological controls required to maximize the human rights and other benefits of AI concurrently with the technology itself. AI is the first truly transformative technology to come of age following the articulation of the United Nations Guiding Principles on Business and Human Rights. It is emerging at a time that it is widely understood that businesses have a responsibility to respect human rights, and that due diligence is the key to do-

ing so. Whereas in the past private sector innovators could be ignorant or willfully blind to the human rights consequences of the technologies they are developing, that is no longer the case.

Due diligence, as the term is used in the Guiding Principles, is the essential first step toward identifying, mitigating, and redressing the adverse human rights impacts of AI. Therefore, as a minimum, public policy efforts should be directed toward ensuring that all who are involved in building these systems engage in the kinds of due diligence that will ensure that they respect human rights by design. Such efforts may be enhanced by mandating or incentivizing the developers and operators of AI systems to make available the training data and the outputs of their systems to external reviewers.²²⁴

It is heartening that many of the biggest players in developing AI have risk management systems in place that trigger human rights due diligence processes at all appropriate stages in the lifecycle of a technology.²²⁵ That being said, there are at least three challenges endemic to the AI space that may prevent human rights due diligence from being as effective as it might otherwise be.

²²¹ B.L. Hutchins and A. Harrison, *A History of Factory Legislation*, 2nd ed. (London: P. S. King & Son, 1911), <https://archive.org/details/history-offactory014402mbp>.

²²² For a popular and accessible history of automobile safety, listen to 99% Invisible, *The Nut Behind the Wheel*, accessed June 21, 2018, <https://99percentinvisible.org/episode/nut-behind-wheel/>.

²²³ "National Traffic and Motor Vehicle Safety Act of 1966," Pub. L. No. 89-563 (1966).

²²⁴ In this vein, New York University's AI Now Institute has developed a framework for public-sector entities in the United States to use in carrying out "algorithmic impact assessments" prior to purchasing or deploying automated decision systems. Dillon Reisman et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (New York University AI Now Institute, April 2018), <https://ainowinstitute.org/aiareport2018.pdf>.

²²⁵ For example, the eleven member-companies of the Global Network Initiative ("GNI"), which include some of the biggest players in the AI space, commit to "carry out human rights due diligence to identify, prevent, evaluate, mitigate and account for risks to the freedom of expression and privacy rights that are implicated by the company's products, services, activities and operations." GNI member-companies are independently assessed every two years to evaluate their compliance with this and other commitments. "Implementation Guidelines" (Global Network Initiative), accessed June 21, 2018, <https://globalnetworkinitiative.org/implementation-guidelines/>.

The first arises from the relatively low awareness among small, early-stage companies of the corporate responsibility to carry out human rights due diligence. This is by no means a challenge that is unique to AI, but given the potential of these technologies to scale up rapidly, it might be more problematic in this space than in other industry verticals.²²⁶ Moreover, given that certain AI systems are often empty vessels into which the end-user can feed whatever training data it wants to automate a formerly manual process, technology developers can be too remote from on-the-ground realities to assess the human rights impacts of the uses of their products.²²⁷ Correspondingly, there is an opportunity to significantly advance human rights by adopting measures to incentivize much wider due diligence efforts throughout the entire AI ecosystem.

The second arises from the difficulty in ascertaining the real-world impacts of any given AI application prospectively. That difficulty arises from the inscrutability of so many AI systems and from the complex interactions that these systems have once they begin to operate in the real world. It is hard enough to predict what human rights impacts a relatively anodyne product will have when it is released into the marketplace, hence the challenge of assessing the human rights impacts of AI systems before they

are deployed is all the more considerable. The problem is particularly acute in AI systems which utilize machine or deep learning, such that the AI developer herself may not be able to predict or understand the system's output.²²⁸

To be sure, the Guiding Principles make clear that human rights due diligence is an ongoing responsibility for precisely this reason: not all impacts can be predicted, even with reasonable diligence. Correspondingly, all entities that are involved in the development or use of these technologies must have measures in place to ensure that human rights due diligence is not a matter of "once and done." Especially given the complexity of AI systems and the fact that the results are often not explainable by conventional means, new analytical techniques and performance metrics may need to be developed to determine whether AI systems are helping or harming human rights. Developing these techniques and metrics is a challenge that the computer science community is working to tackle with alacrity. In Europe, this challenge has been framed in part by the provisions of the General Data Protection Regulation, which requires some human involvement in automated decision-making²²⁹ and encourages the development of "a right to an explanation."²³⁰

226 Dalia Ritvo, Vivek Krishnamurthy, and Sarah Altschuller, "Managing Users' Rights Responsibly—A Guide for Early-Stage Companies," 2016, http://www.csrandthelaw.com/wp-content/uploads/sites/2/2016/03/Managing-Users-Rights-Responsibly_A-Guide-For-Early-Stage-Companies-no-logos.pdf.

227 Consider, for example, the controversy that emerged around the time that this report was being finalized regarding the U.S. government's use of facial recognition technology supplied by Microsoft in implementing its (now-rescinded) policy of separating the children of unlawful migrants from their parents. One can speculate that Microsoft could not have foreseen how the U.S. government would use this technology at the time it contracted to provide this system, which highlights the need for companies in this space to conduct *ongoing* due diligence. Catherine Shu, "Microsoft Says It Is 'Dismayed' by the Forced Separation of Migrant Families at the Border," *TechCrunch*, June 19, 2018, <http://social.techcrunch.com/2018/06/18/microsoft-says-it-is-dismayed-by-the-forced-separation-of-migrant-families-at-the-border/>.

228 To be clear, many AI applications reason in ways beyond human comprehension. This is particularly true for applications based on machine learning and deep learning. Yet, that difference may be insufficient to justify holding AI back. In fact, it may even be a reason to delegate decisions to AI. David Weinberger, "Our Machines Now Have Knowledge We'll Never Understand," *WIRED*, April 18, 2017, <https://www.wired.com/story/our-machines-now-have-knowledge-we-ll-never-understand/>.

229 GDPR, art. 22.

230 *Ibid.*, art 22 and recital 71. For more information, see Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *AI Magazine* 38, no. 3 (October 2, 2017): 50, <https://doi.org/10.1609/aimag.v38i3.2741> and Sandra

A third complication arises from the uncertainty as to what constitutes an effective remedy to an adverse human rights impact generated by an AI system. Since a right without a remedy is no right at all, the right to remedy is the “third pillar” of the “protect, respect, remedy” framework on which the Guiding Principles rest. Specifically, Guiding Principle 25 recognizes the duty of the state to provide access to effective judicial and other remedies to all those who have been affected by business-related human rights abuses. Judicial remedies in particular may be better suited to addressing the adverse consequences of AI on some human rights over others. For example, judicial remedies may well be more effective in detecting and redressing adverse human rights impacts caused by the use of AI in the criminal justice system, as compared to some other fields of endeavor. Especially since criminal procedural rights are articulated in much more detail than most other human rights in domestic and international law, there is simply more material to work with in terms of identifying when an AI system is adversely impacting rights in this category.²³¹

Another major challenge facing both judicial and non-judicial remedies is the nature of the harm that AI makes possible. That is, all remedial sys-

tems, whether public or private, are much better at remediating substantial harms suffered by the few, as opposed to less significant harms suffered by the many. Consider, for example, some of the most widely-known operational level grievance mechanisms that have been established in the last decade to remedy the adverse human rights impacts of businesses. These include mechanisms established by mining companies to compensate the victims of sexual violence,²³² or by global technology companies to vindicate the right that some courts have recognized of individuals to be “forgotten” online.²³³ These systems provide remedies to individuals or to small groups of people that have suffered a particularized human rights harm, but they are simply not designed to cope with much more diffuse and oftentimes covert harms that might be every bit as pernicious.²³⁴

These difficulties are magnified in the AI space by the challenge of detecting the harm and determining and proving causation. Consider, for example, the difficulties that a loan applicant would face in proving that a lending algorithm has discriminated against them, in a situation where seven prospective lenders turned them down, but three others offered them credit. Assuming that the objective truth of

Wachter, Brent Mittelstadt, and Luciano Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation,” *International Data Privacy Law* 7, no. 2 (May 2017): 76–99, <https://doi.org/10.1093/idpl/ixx005>.

231 The fact that the judiciary might be better able to grapple with the adverse impact of AI in its own backyard does not mean that it will reach the right answer in every case, or that the remedies it provides when a violation is found will be sufficient. For example, the U.S. Supreme Court has been roundly criticized for denying review of *Wisconsin v. Loomis*, 881 N.W.2d 749 (2016). See Ellora Israni, “Algorithmic Due Process: Mistaken Accountability and Attribution in *State v. Loomis*,” *Harvard Journal of Law & Technology Digest*, 2017, <https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1> and Taylor R Moore, “Trade Secrets and Algorithms as Barriers to Social Justice” (Center for Democracy & Technology, August 2017). The point is simply that some alleged rights violations attributed to AI systems are easier to redress judicially in some fields as opposed to others.

232 Yousuf Aftab, “Pillar III on the Ground: An Independent Assessment of the Porgera Remedy Framework” (Enodo Rights, January 2016), <http://www.enodorigths.com/assets/pdf/pillar-III-on-the-ground-assessment.pdf>.

233 Jacques de Werra, “ADR in Cyberspace: The Need to Adopt Global Alternative Dispute Resolution Mechanisms for Addressing the Challenges of Massive Online Micro-Justice,” *Swiss Review of International & European Law* 26, no. 2 (2016): 289, <https://doi.org/10.2139/ssrn.2783213>.

234 To be sure, the same criticism can be leveled against traditional courts which, despite such innovations as class-action lawsuits and contingency fee arrangements, remain an unaffordable and inaccessible option for many victims of rights violations.

the matter is that some of the seven decliners engaged in discrimination, would this person even suspect that they have been the victim of discrimination? What might be the required elements of proof to establish a discrimination claim? How costly would it be to bring such a claim, as against the anticipated value of the remedies available? What expert evidence and analysis would be required to open the “black box” of the algorithm, especially when it is protected by trade secrets and intellectual property law? Now assume that the stakes of what the algorithm is deciding are much lower than a loan, and yet there are adverse consequences distributed over a large population. Who would go through the trouble of seeking a remedy for that harm, and how and where might they do so?

Then there are the additional complications around remedying AI’s impacts on economic, social, and cultural rights. International and domestic legal systems alike have much more developed doctrines and procedures with regard to civil and political rights than they do for economic, social, and cultural rights. This is due in part to the fact that the international human rights community has prioritized civil and political rights over economic, social, and cultural rights for the last 70 years.²³⁵ But it is also because the state duty in relation to economic, social, and cultural rights is to progressively realize them over time, in view of the available resources, which makes it much harder to identify when these rights are adversely impacted—especially by businesses.

The recent General Comment on “State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities” makes clear the difficulties.²³⁶ The General Comment notes that the state obligation to protect human rights requires them to “prevent effectively infringements of economic, social and cultural rights in the context of business activities” by adopting “legislative, administrative, educational and other appropriate measures, to ensure effective protection against Covenant rights violations linked to business activities.”²³⁷ Yet all of the examples the General Comment provides of “rights violations linked to business activities” are attributable to state failures to regulate the marketplace.²³⁸

The underdevelopment of the regime of economic, social, and cultural rights makes it difficult for businesses engaged in human rights due diligence to know what they should do when their systems adversely impact one of these rights. Consider, for example, the impact that at least some AI systems are sure to have on employment. While it is the duty of the state to protect the right to work, large-scale workforce displacement caused by the deployment of AI obviously burdens this right. As things stand right now, however, it is difficult for a business to determine what if anything it should do to mitigate this impact on the right to work.

The example of the right to work points to the profound challenges related to addressing the distributive consequences of AI—especially with regard to

²³⁵ Samuel Moyn, *Not Enough: Human Rights in an Unequal World* (Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 2018); Samuel Moyn, “How the Human Rights Movement Failed,” *The New York Times*, April 24, 2018, sec. Opinion, <https://www.nytimes.com/2018/04/23/opinion/human-rights-movement-failed.html>.

²³⁶ United Nations Economic and Social Council, Committee on Economic, Social and Cultural Rights, General Comment 24 on “State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities” (2007), U.N. Doc. E/C.12/GC/24.

²³⁷ *Ibid.*, para. 14.

²³⁸ *Ibid.*, paras. 18–22.

economic rights, in view of the fears that AI might trigger widespread unemployment. As several of the case studies find, a particular AI system being used in a particular field of endeavor can positively impact the enjoyment of a particular human right by some individuals, at the same time as it adversely affects that right for others.

It may well be that over the course of time, the international human rights system can develop guidance that is more responsive to such distributive issues in the context of AI. Until that happens, however, there is a strong case to be made that national governments should weigh in on these questions, whether by soft suasion or hard regulation, in a manner that is rights-respecting, yet also reflects the country's particular values and public policy priorities. In other words, when due diligence reveals human rights impacts that have complex distributive consequences, the need for government public policy leadership is at its greatest.²³⁹

There is also an important role for governments to play in creating conditions that encourage businesses to take their human rights responsibilities seriously. In his influential study of private initiatives to improve labor rights in global supply chains, Richard Locke found that “the effectiveness of private regulatory programs is very much tied to the strength of public authoritative rule-making institutions.”²⁴⁰ In the AI space, governments could consider creating incentives to ensure that effective due diligence is undertaken, and to build capacity

among earlier-stage companies to develop their technologies in a rights-respecting manner.

Finally, there is a crucial role for governments to play in creating accountability and redress systems for their own use of algorithmic tools, and for those adverse impacts that cannot easily be addressed by private grievance mechanisms. The General Data Protection Regulation (“GDPR”), which recently came into force in the European Union, is noteworthy in this regard for its provisions requiring “data subjects” to be provided with “meaningful information about the logic involved, as well as the significance and the envisaged consequences” of the automated processing of their personal data.²⁴¹

It is too soon to tell whether the GDPR's embrace of algorithmic “explainability” will prove to be successful in creating greater accountability for such systems, or whether this approach will instead chill promising developments in AI that produce useful results even if their logic defies human comprehension. What is certain, however, is that collaboration between technology companies, governments, and representatives of the diverse community of stakeholders that AI will impact is required to develop new ways of ensuring that this technology delivers on its promise in a rights-respecting manner.

²³⁹ As Richard Locke notes in a related context, “[t]he inherent problem with private voluntary initiatives... is their inability to reconcile diverse and conflicting interests and thus promote solutions that require collective action among [] myriad actors...” Richard M. Locke, *The Promise and Limits of Private Power: Promoting Labor Standards in a Global Economy*, Cambridge Studies in Comparative Politics (Cambridge; New York: Cambridge University Press, 2013): 178.

²⁴⁰ *Ibid.*, 68.

²⁴¹ GDPR art. 14(2)(g).

7. Conclusion

As should now be clear, the relationship between artificial intelligence and human rights is complex. A single AI application can impact a panoply of civil, political, economic, social, and cultural rights, with simultaneous positive and negative impacts on the same right for different people. Multiply these impacts across the full range of cases where AI is already in use or will soon become commonplace, and the magnitude of this technology's impact on society begin to become clear.

Society has dealt with revolutionary technological change in the past, and we have always arrived at a new equilibrium. But we today are better placed than our forebears for the change that is upon us because of the adoption, 70 years ago this December, of the Universal Declaration of Human Rights.

The UDHR gives us a powerful and universally-accepted framework not just for identifying and overcoming past and present wrongs, but also for building a future that respects and honors the rights of every person. This depends, however, on our remaining vigilant to the impacts of our actions on the rights of others. Hence the importance the Guiding Principles place on due diligence both before we deploy these powerful new technologies, and throughout their lifecycle, too.

We are heartened by the growing attention that human rights-based approaches to assessing and addressing the social impacts of AI have begun to receive. We view it as a promising sign that so many of the private enterprises at the forefront of the AI revolution are recognizing their responsibility to act in a rights-respecting manner. But the private sector cannot do it alone, nor should it: governments have a crucial role to play, both in their capacities as developers and deployers of this technology, but also as the guarantors of human rights under international law.

The fundamental role of government in defining and making available remedies for human rights violations cannot be overstated. Of equal or greater importance, however, is the governmental responsibility to evaluate and address the distributive consequences of AI. The institutions and processes of democratic government are the only ones with the legitimacy to determine what distribution of benefits and burdens across society is fair, so now is the time for them to embrace their role in shepherding society through the changes that lie ahead.

8. Further Reading

Understanding AI

David Weinberger, “Our Machines Now Have Knowledge We’ll Never Understand,” *Wired*, April 18, 2017, <https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/>.

This article explores how artificial intelligence is changing the way we think about knowledge by delving into the “explainability” debate. It explains machine learning, artificial neural networks and their implications in an accessible manner. Many machine learning models, like Google’s AlphaGo algorithm, are “ineffably complex and conditional”: they make decisions based on opaque and unexplainable patterns, not transparent principles. These systems are becoming more accurate as data becomes more abundant, yet there is a tradeoff between being able to understand why an algorithm makes a decision (a function of its complexity) and its accuracy. Because reality is complex, artificial intelligence may be able to account for an abundance of factors that are beyond human comprehension. Yet the lack of explainability can make it difficult to identify bias in algorithms. (*Category: Concise overview*)

Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society* 3, no. 1 (January 5, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674.

This article describes three different kinds of “opacity” in algorithms: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) opacity arising from the characteristics of machine learning algorithms that make them useful. Burrell argues that recognizing these distinct forms of opacity is important to determining what technical and non-technical solutions can prevent algorithms from causing harm. On (1), secrecy may be essential to the proper function of an algorithm (such as to prevent it from being gamed), but such algorithms are easily reviewable by trusted and independent auditors. Regarding (2), the solution to technical illiteracy is simply greater public education. Finally, (3) is difficult because there may be a trade-off between fairness, accuracy, and interpretability. Certain AI techniques could be avoided in fields where transparency is crucial, or new benchmarks could be developed to assess such algorithms for discrimination and other issues. (*Category: In-depth*)

Defining the Problem: What's at Stake?

Navneet Alang, “Turns Out Algorithms are Racist,” *New Republic*, August 31, 2017. <https://newrepublic.com/article/144644/turns-algorithms-racist/>.

This article explains that AI is only as good as the data it is fed. Computers and software, even at their most sophisticated, are essentially input-output systems that are “taught” by feeding them enormous amounts of data. Hence if the input data reflect gender, racial, or other biases, so too will the output. (*Category: Concise overview*)

Robert Hart, “If you’re not a white male, artificial intelligence’s use in healthcare could be dangerous,” *Quartz*, July 10, 2017, <https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/>.

While acknowledging the potential of AI to revolutionize healthcare, this article points to the danger that this technology will perpetuate healthcare inequalities due to its reliance on existing stores of medical data. Groups including women (especially pregnant women), the young, and the elderly are excluded from many medical research studies, which may result in errors when individuals from these groups are treated by AI systems trained on this data. (*Category: Case study – healthcare*)

J. Kleinberg et al., “Human Decisions and Machine Predictions,” National Bureau of Economic Research, February 2017, <https://www.cs.cornell.edu/home/kleinber/w23180.pdf>.

This article highlights how machine learning can help humans make better decisions, using the example of judges making bail decisions. While acknowledging the difficulties associated in fully automating bail determinations—both due to biases in the training data that would be fed into such a system and the complex mix of factors that judges weigh—AI-based analyses show that the number of people jailed before trial can be substantially reduced without impacting the crime rate. Such insights can then be used by judges to improve their own decision-making. (*Category: Case study – criminal justice*)

danah boyd, Karen Levy, and Alice Marwick, *The Networked Nature of Algorithmic Discrimination*, Washington, DC: New America Foundation, 2014, <https://www.danah.org/papers/2014/DataDiscrimination.pdf>.

This study points to the dangers surrounding algorithms making predictions about you based on those you associate with—such as your friends and neighbors. While existing laws prohibit racial and gender discrimination, among other things, an individual’s position in a social network is deeply affected by these and other variables—which algorithms might then use to make predictions about us, leading to unfair results. Consequently, the authors argue that we need to rethink our models of discrimination to consider not just an individual’s immutable characteristics, but also the impacts of how algorithms position us within a network and society, too. (*Category: Novel issues – networked discrimination*)

Approaches to Regulating AI

S. Wachter et al., “Transparent, explainable, and accountable AI for robotics,” *Science Robotics* 2, no. 6 (May 31, 2017), <http://robotics.sciencemag.org/content/2/6/eaan6o8o.full>.

This article provides a brief overview of the challenges facing governments seeking to regulate AI. After briefly grounding the regulation of automated systems in its historical context, it raises the critical questions facing regulators seeking to enact optimal laws in the space. (Category: *Concise overview*)

IEEE Global Initiative on Ethics of and Autonomous and Intelligent Systems, “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems” (IEEE, 2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

This document, prepared by the world’s largest professional organization in the technology space, calls for an ethics- and values-based approach to dealing with the impacts of intelligent and autonomous systems that prioritizes human well-being within a given cultural context. (Category: *In-depth analysis*)

Corrine Cath et al., “AI and the ‘Good Society’: the US, EU, and UK approach,” *SSRN Electronic Journal* (2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906249

In October 2016, the White House, the European Parliament, and the UK House of Commons each issued reports outlining their vision on how to prepare society for the widespread use of AI. This article provides a comparative assessment of these three reports to facilitate the design of policies favorable to the development of a ‘good AI society’. (Category: *Comparative overview*)

National Science and Technology Council: Committee on Technology, *Preparing for the Future of Artificial Intelligence*. Washington, D.C.: Executive Office of the President, October 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

This report surveys the current state of AI research, its existing and potential applications, and the questions that AI raises for society as a whole and for public policy in particular. The report recommends certain specific governmental and non-governmental actions within the U.S. context and sets forth a strategic plan for U.S. government funding of AI. (Category: *Regulatory template*)

Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. New York University AI Now Institute, April 2018. <https://ainowinstitute.org/aiareport2018.pdf>.

This report proposes an “Algorithmic Impact Assessment” framework to be used by public entities in the U.S. in procuring and deploying AI systems. The framework is designed to support affected communities and stakeholders as they seek to assess the claims made about automated decision systems, and to determine where their use is acceptable. It offers key elements for the construction of a public agency algorithmic impact assessment, and a practical accountability framework combining agency review and public input. (Category: *Regulatory template*)

Business and AI

Sherif Elsayed-Ali, “Why Embracing Human Rights Will Ensure AI Works for All,” April 2018, <https://www.weforum.org/agenda/2018/04/why-embracing-human-rights-will-ensure-ai-works-for-all/>.

This article recommends four actions to prevent discriminatory outcomes in machine learning based on the UN Guiding Principles on Business and Human Rights. These are: (1) active inclusion of underrepresented populations in datasets and AI development, (2) fairness in the interpretation of biased data, (3) a right to understand how algorithmic decisions are made, and (4) access to redress when wrong decisions are made. (*Category: Concise overview*)

R. Jorgenson, “What Platforms Mean When They Talk About Human Rights,” *Policy & Internet* 9, no. 3 (May 29, 2017) <https://doi.org/10.1002/poi3.152>.

This article examines how two major Internet platforms—Google and Facebook—make sense of human rights. Based on primary research, the authors find that the companies frame themselves as strongly committed to human rights. Yet this framing focuses primarily on government rights violations, rather than the companies’ own adverse impacts on its users’ rights. (*Category: Case study—Internet platforms*)

Shift, Oxfam and Global Compact Network Netherlands, *Doing Business with Respect for Human Rights: A Guidance Tool for Companies*, 2016, https://www.businessrespecthumanrights.org/image/2016/10/24/business_respect_human_rights_full.pdf.

This report provides comprehensive guidance to businesses on what they should do to operationalize their responsibility to respect human rights, as recognized in the UN Guiding Principles on Business and Human Rights. (*Category: In-depth*)

Location-Based Data in Crisis Situations

Principles and Guidelines

March 2019



Association for the Advancement of Science (AAAS). As a program of AAAS – the world's largest multidisciplinary scientific membership organization – SRHRL fosters and facilitates the responsible practice and application of science in the service of society. The Program is committed to promoting high standards for the practice of science and engineering; advancing the human right to enjoy the benefits of scientific progress and its applications; engaging scientists, engineers and their professional associations in human rights efforts; monitoring and enhancing assessment of emerging ethical, legal, and human rights issues related to science and technology; furthering the use of science and technology in support of human rights; and initiating activities to address the impact of developments at the intersection of science, technology, and law.

This report was prepared by:

Jessica Wyndham, Program Director, AAAS SRHRL

Ellen Platts, former Staff Assistant, AAAS SRHRL

Jonathan Drake, Senior Program Associate, AAAS SRHRL

Acknowledgement

The authors wish to acknowledge Dr. Susan Wolfinbarger who, together with Dr. Mark Frankel, conceived of and initiated this project. They also acknowledge the participants in the three workshops who came from academia, civil society, government, industry, international non-governmental and multilateral organizations, and professional scientific societies. These participants, together with multiple reviewers from academia and civil society offered their experience and intellectual rigor to the Principles and Guidelines.

Primary support for this project was provided by the U.S. National Science Foundation through award number 1560948.

Disclaimer

The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the AAAS Board of Directors, its Council, or membership, or the National Science Foundation.

Contact

AAAS welcomes comments and questions regarding its work. Please send information, suggestions, and any comments to SRHRL at srhrl@aaas.org.

© Copyright 2019

American Association for the Advancement of Science

Scientific Responsibility, Human Rights, and Law Program

1200 New York Avenue, NW

Washington, DC 20005 USA

Cite as: AAAS Scientific Responsibility, Human Rights and Law Program, "Location-Based Data in Crisis Situations: Principles and Guidelines" (Report prepared by Jessica Wyndham, Ellen Platts, and Jonathan Drake), March 2019. DOI: 10.1126/srhrl.aax3877

Location-Based Data in Crisis Situations: Principles and Guidelines

Background

This document is the product of three workshops hosted by the American Association for the Advancement of Science (AAAS) in Washington, D.C. in 2016 and 2017. Participants in the workshops came from academia, civil society, government, industry, international non-governmental and multilateral organizations, and professional scientific societies. Their aim was to address the ethical issues associated with rapid growth in the use of location-based data such as remotely-sensed imagery, geotagged social media posts, and electronic communications records by crisis response actors.

Purpose and Scope

Data relating to the location of infrastructure, resources, and people can have positive and negative applications. In all circumstances, there are potential risks and benefits associated with collecting, aggregating, representing, using, and storing such data. In the context of crises, however, the nature and significance of the risks and benefits will differ from a non-crisis context. The following principles and guidelines aim to fill a gap in ethical guidance for the generation, collection, analysis, dissemination, and use of location-based data in crisis situations.

“Location-based data” is information that contains or is associated with position. In addition to spatial information, it typically contains a temporal component and is often, but not always, generated by electronic devices that are “location-aware” through the integration of satellite or terrestrially-based positioning systems. Location-based data are frequently shared via the internet and social media, and range from highly granular datasets, (e.g., individual geotagged photographs) to highly aggregated ones, (such as total reports of an earthquake over an entire region). Data that are location-based include data associated with specific coordinates (at varying levels of resolution), street addresses, or other information that can be geolocated, and named locations that may or may not be ambiguous (e.g., a city name in a Twitter profile, or a unique building mentioned in the text of a document). These data can apply either to individuals or groups, and in crisis situations such data are increasingly ubiquitous and used by multiple actors.

Per this definition, the term “location-based data” encompasses, but is not limited to: (a) data specifically and deliberately collected in preparation for or in response to an event by volunteers, local actors, or institutions, which contains a location or geographic component; (b) “Volunteered Geographic Information” (VGI) – user-generated content, created by individuals or groups for a purpose other than a deliberate data collection effort and which contains spatial information in a way that can be used by others (e.g., social media feeds); (c) ambient data and other data collected, frequently as a secondary process and without the specific knowledge and/or consent of the individual or groups (e.g., CCTV feeds, mobile phone records); and (d) remotely-sensed imagery, including geolocated photographs.

“Crisis situations” are situations of conflict and other situations of violence and natural disaster, including those that result in violations of human rights and international humanitarian law, and/or the destruction of cultural heritage. Because these situations are inherently dynamic and complex, affecting people and involving location-based data in a multiplicity of ways, the judgement of whether a given situation constitutes a crisis is often challenging, and will always contain an element of subjectivity. For

the purposes of planning and executing a response, however, defining the crisis in time and space is a key component of planning an effective response. The principles and guidelines that follow are designed to operate within this framework, and should be applied accordingly.

Primary Audiences for This Document:

- Newcomers to the use of location-based data, particularly in crisis situations;
- Volunteer networks of technical experts as well as other non-traditional data collection teams;
- Academic researchers such as geographers, including those working alone and those working at the behest of another actor;
- Local partners who may have their own codes of ethics, but need to find shared points of understanding with collaborators coming into their communities; and
- The broader community of non-experts involved in aid and relief work, as well as participatory science efforts.

Secondary Audiences:

- Existing organizations that are already familiar with and involved in the use of location-based data in crisis situations, including humanitarian and human rights organizations, government agencies and other entities with dedicated individuals or teams conducting research using location-based data and professional societies and associations with a focus on such research.

Principles:

1. Do No Harm: Identify and minimize potential risks of location disclosure, particularly as they may affect the vulnerability of individuals and populations
2. Define Your Purpose: Ensure action is mission-driven and goal-oriented
3. Do Good Science: Employ scientifically rigorous and responsible methods
4. Collaborate and Consult: Engage with local partners
5. Give Access to Your Data: Share data openly, when safe and practicable

Principle #1: Do No Harm - Identify and minimize potential risks of location disclosure, particularly as they may affect the vulnerability of individuals and populations

- A. *Know your specific context:* When locally contextualized, location-based data will inherently implicate and identify potentially vulnerable individuals, groups, organizations, and resources which will likely change over time. Risks specific to a particular crisis may include a context of violence, of discrimination against and/or targeting of specific gender, ethnic, religious, cultural or other groups, and/or economic and technological marginalization. These risks pertain to volunteers collecting data as well as to individuals and groups whose data are collected. Depending on the context, these risks might be significantly increased in the context of hostile state, military or paramilitary actors. The kinds of risks that can arise from failing to address the context specific to a particular location can include data generated in good faith being used maliciously. Be prepared to walk away from projects if it is clear that victims, bystanders, and response personnel cannot be adequately protected.
- See Decision Tree: “Should I collect”, Step 2.
- B. *Assess the risks in a given crisis as it evolves:* Because most location-based data have an inherent temporal dimension, changing contexts under conditions of crisis have the potential to enable new forms of identification that can result in new risks and opportunities. When using location-based data in a crisis situation, each stage of the data management cycle - from data collection to analysis, communication, storage, archiving, and deletion - engenders different risks and ethical obligations. Risk assessments should be conducted periodically throughout the crisis as it evolves. They should consider harmful outcomes related to the (mis)use of location-based data that have occurred in similar situations in the past, as well as potential harms that are foreseeable in light of the current crisis. Consulting with local communities and leaders is a key component of this process, as they may be able to envision risks that outsiders cannot.
- See Decision Tree: “Should I collect”, Step 4.

One possible way to organize these parameters is by incorporating them into a framework known as a “risk matrix” that plots the probability of the harm taking place against the potential outcome’s severity, with the risk represented by the product of these two factors. Such a matrix, shown in the appendix, can be helpful in identifying which types of events must be planned for in advance. While assigning numeric values to the severity and probability of possible harms is a somewhat subjective exercise, comparisons with similar response efforts in different locations or at different times may provide helpful guidance for this purpose. It is likewise important to recognize which combinations of likelihood and severity constitute “red lines” that must not be

crossed, and to consider that the vulnerability of various groups is not static; it may emerge or recede over the course of a crisis. Likewise, be aware that the risks to individuals may differ from the risks to groups. In the event that a large degree of uncertainty exists concerning the probability and/or severity of potential adverse outcomes, caution is warranted; responders should consider whether a particular activity is vital to achieving the desired outcome. When conducting such an assessment, however, one must not lose sight of the potential benefits of pursuing a course of action, nor of the fact that parameters may influence decision-making that defy categorization as “risk” or “benefit”.

C. *Proceed according to the risks associated with each stage of the data cycle in the context of the crisis:*

- i. *Collection:* Select volunteers carefully, and take steps to protect their privacy. Ensure that data collected are consistent with the requirements and context of the situation, consider potential sources of error in the data, and whether those errors may add to risk. Protect the privacy and identity of subjects, and ensure volunteers are trained to do the same. Engage with local partners and, at minimum, ensure they are aware of the data collection effort and why it is being undertaken. Be aware that some location-based data –particularly user-generated data– may have been geotagged inadvertently. If volunteers submit data directly to the response effort, ensure that they are aware of the full nature of the data they are sharing, and allow them to opt out of submitting certain metadata if desired. If metadata is included, special care must be taken to determine whether the risks associated with the data may be enhanced when combined with other freely available information (e.g., identifying populations that prefer to remain hidden). If so, thought should be given to whether the data or its collection strategy might be modified in such a way that mitigates the risk while preserving the usefulness of the data to the crisis response (e.g., masking the data). The collection effort should be continuously monitored and evaluated to ensure that the data being acquired are still necessary. If not, the collection should be suspended.

See Decision Tree: “Should I collect”, Step 3.
- ii. *Sourcing:* When relying on location-based data collected by others, identify its provenance, assess its quality and determine its use based on the origin of the data. Evaluate the potential for false or spoofed location data, and evaluate the motivations that might be behind a release. Even if accurate, some data may have been selectively curated or released in order to advance a purpose unrelated to, or potentially at odds with, the response effort. When enlisting sources weigh the risks to the source against the risk of not having the data.

See Decision Tree: “Should I collect”, Step 4.
- iii. *Analysis:* Recognize that combining multiple datasets often enables insights that are more actionable than the sum of their parts. In light of this, assess the risks and benefits of using analytical methods that can generate new location-based information where none was provided (e.g., methods that can determine the home location of a Twitter user based on analysis of the full set of their tweets, profile information, and other internet based information)

See Decision Tree: “Should I share”, Step 2.

about them). Recognize that the context of different crises (e.g., natural disaster vs. conflict) can affect the level of antagonism present in a situation, and that this may affect decision-making.

- iv. *Communication*: Whether data are shared or visualized and, if so, which data, should be decided based on a broad assessment of the associated risks. Consider not only the risks to individuals and groups, but also to ecosystems and infrastructure. Ensure that communication and/or visualization of data does not reveal the locations of vulnerable populations, sites, and artifacts, and take steps to avoid putting any at additional risk. Even when the underlying data are obscured, visualizations may be sufficiently detailed to enable mis-use. If the use of such information is essential, ensure that informed consent has been obtained from individuals who generated or are most affected by the data. The type and level of consent given when the data were collected should be taken into account, as should the potential that the data's availability may change from being helpful to being harmful in the future. When communicating location-based data findings with other actors, decision-makers, or the public, consider the ways the location information may be used and perceived by various audiences, and aim to disseminate and clearly communicate the data in a way and at a locational precision that will take these factors into account while minimizing personal risk, retraumatization, misinterpretation, or misuse.
- v. *Storage*: Ensure that data are stored securely, and if practical consider storing datasets separately so that they cannot be combined in the event of a data breach. Use encryption for the storage of any data that contain, or could be extrapolated into containing, personally identifying, demographically identifying, or other sensitive information such as cultural data. Ensure that encryption keys are entrusted to a limited number of trustworthy individuals, but not so few as to hamper operations. Make regular backups of data and store them in multiple locations, such that a mishap at one site will not result in catastrophic data loss, while remaining cognizant of the increased risk associated with maintaining multiple copies of the data.

See Decision Tree: "Should I collect", Step 4.
- vi. *Archiving and deletion*: Location-based data need not be kept indefinitely. Once the crisis has passed, consider whether there is any benefit to maintaining the data in archival form, and whether those benefits outweigh the potential risks. Input from the affected community can be particularly helpful in making this determination (see principle #4).

Principle #2: Define Your Purpose - Ensure action is goal-oriented

- A. *Ensure that collection of location-based data is necessary*: Given the emerging forms of both benefits and unintended consequences that can arise from location-based data, collection of location-based data (both generally, and specific types of location-based data) should only occur when it is necessary to the articulated goal of the project, taking into account the needs and interests of the local community as well as of the overall crisis response effort. Groups that are new to data collection in crisis might want to consider partnering

See Decision Tree: "Should I collect", Step 1.

See Decision Tree: "Should I share", Step 1.

with recognized and experienced humanitarian groups for the design of collection and analyses protocols.

- B. *Avoid overcollection of data:* Data should be collected only if a clear plan exists for its use; “it might be useful later” is not a sufficient justification for collecting. Recognize that, depending on the specific nature of the crisis situation, the threshold of what constitutes overcollection may change. Conducting a review of existing data already collected by others may help in making this determination. If increased data collection is deemed necessary, however, it still must always further the project’s stated goal, and take place subject to the guidance articulated in principle #1, above. If possible, however, avoid having to re-collect data. When possible, consider both the data’s immediate purpose and its likely purpose in any further stages of the crisis.
- C. *Consider the boundaries of the data collection effort:* As early as possible in the crisis response effort, identify the spatial and temporal horizons that will be applied to data collection. Establish a clear set of criteria that can be used to define the end of the crisis and associated data collection efforts, taking into account factors such as funding, resources, and timeline. Although the time at which the crisis began is often obvious, this is not always the case, in which case these same or similar criteria can often be used to define the time before which the data to be collected should not cover. Similarly, the extent of the geographic area affected by a crisis is often difficult to define precisely, however data collection efforts will often have to be bounded in order to be useful – the boundaries of the data collection effort and the broader crisis may not overlap. When defining the boundaries of both the crisis and the response, consider both the availability of data and the ways in which it is going to be applied in the response effort.

Principle #3: Do Good Science - Employ scientifically rigorous and responsible methods

- A. *Verify Data Sources:* Collection and analysis of VGI and location-based data should lead to accurate and verifiable results that are relevant and actionable in the crisis context. Ideally, data that are used in the response effort should come from reputable sources which are transparent about their own methods of collecting and aggregating the resulting data. Wherever possible, verify the provenance and accuracy of third-party data sources. If these cannot be established, consider whether alternative sources of information may be available, and keep in mind that no information at all is often superior to inaccurate information.
- B. *Use caution when employing experimental methods:* Conducting response efforts with scientific rigor does not preclude either innovative approaches or non-traditional methods, however these should be identified as such, and should only be employed following a thorough risk-benefit analysis involving both the response’s scientific experts and the affected community. If the data or their analysis have the potential to be used in a legal context, consider the impact that using experimental methods may have on such a proceeding, versus more established methods that have already been subject to rigorous scientific scrutiny. Consider that the requirement for scientific rigor does not imply that practices must be complicated; simple methods can still be rigorous.
- C. *Train volunteers engaged in the collection of data:* Methods and technologies involving location-

based data are frequently distinct from those associated with other data, and must be understood by those employing them. Before individuals begin collecting data, provide training to ensure that they are aware of the risks and responsibilities (both to themselves and the research subjects) associated with the task, making them aware of any and all standards for data, including chain of custody standards. Consider the ethical obligations of individuals playing different roles in the data management lifecycle.

- D. *Act in accordance with recognized standards of ethical conduct:* Early on in the response, recognize existing humanitarian, human rights, and ethical frameworks that may be applicable to the investigation, and come to an agreed understanding about common standards. Ethical research practices should be adopted in order best to serve the public good, including privacy and confidentiality, and the protection of human subjects. Consider that ethics and legality may not always align, and be prepared to address such discrepancies. Consider the legal environment of your volunteers. Understand the frameworks and standards guiding the activities of other individuals and organizations operating in similar situations, and determine the standards guiding the collection, analysis, communication, and sharing of location-based data in that context.
- E. *Consider potential sources of bias:* Relying on location-based data, particularly when gathered in a way that relies on the technological capacities and access of local communities and individuals will inevitably give rise to inherent bias in the data and risk empowering some segments of society while perpetuating or exacerbating the marginalization of others. Such biases are often amplified in crisis situations. Assessing the extent of this bias and, where possible, correcting for it will be critical to achieving successful outcomes in a crisis situation. For example, if connectivity is known to have been degraded in certain neighborhoods of a city, reports coming out of that zone might be given more statistical weight than those originating in areas where communications infrastructure remains intact. Recognize that these biases may be temporal as well as spatial, and evaluate the relevance of information accordingly. Identify, be respectful of, and assess how the cultural context may impact the investigation and its data. When sharing data or publishing or otherwise communicating information derived from the data, clearly communicate the level, source, and kind of bias in the data, along with any associated uncertainty. In addition to bias, make every possible effort to perform accuracy assessments which will enable the quantification and communication of error rates associated with the data.
- F. *Submit to peer review:* The methods that a given investigation relies upon should be clearly identified and open to peer and community review, even if that is not immediately possible in a crisis situation. Provided that data collection takes place in a responsible manner, rigorous scientific analysis and peer review can take place at any time, including after the crisis has abated. Such review might, for example, accompany reports to funding organizations.

Principle #4: Collaborate and Consult - Engage with local partners

- A. *Engage community actors*: Recognize that the people best qualified to understand the needs of a community in the wake of a disaster are the people living there. Contact community leaders, explain who you are, and the methods you are contemplating applying – they may already have ideas that you have not considered. Manage community expectations regarding data collection and be realistic about what you are attempting to achieve. Whenever possible without compromising safety, consult and collaborate with these individuals and their communities to best understand their context and needs, and incorporate their input into the design of your work and methods. Recognize that trusted community networks may be able to solve dilemmas –ethical and otherwise– that you are unable to. It may not always be necessary to collaborate, however the process of disclosure regarding methods, data, funders, et cetera is highly effective in establishing and maintaining trust. Strongly consider putting such disclosures in writing.
- See Decision Tree: “Should I collect”, Step 2.
- B. *Build local capacity*: Work with subject matter experts to identify the local populations and partners relevant to the data collection effort. When safe and practicable, engage with local partners to develop any new data collection tools, define data requirements, gather baseline data in advance of potential disasters (known as “data preparedness”), and perform gap identification, data validation, and verification. In particular, engage with local partners to understand the opportunities as well as limitations, risks, and biases (including those that may be introduced by the partners themselves), presented by using location-based data in the context of the crisis. These partners can be extremely helpful in identifying contextually appropriate solutions to difficult issues. If local capacity exists, it should not be duplicated or undermined.
- C. *Obtain Informed Consent*: Particular attention should be paid to the ethical issues associated with obtaining consent in the midst of an emergency or ongoing conflict, recognizing that the nature and form of consent may differ depending on the role and relationship of the person or group of individuals with regard to the data. When initial consent is obtained, the data collected under that consent should be subject to ethical prescriptions regarding the use of those data for a new purpose. Existing ethical codes such as those described in principle #3 may be useful in providing specific guidance in this regard.
- See Decision Tree: “Should I collect”, Step 3.
- D. *Promote community resilience and responsible stewardship of data through the communication and repatriation of knowledge*: Local partners engaged in and/or affected by the collection of location-based data have a right to know how that data is being used, and should be granted access to any data collected by and about them, as well to as any insights and analysis derived from these data, taking the community’s level of data literacy into account. They also have the right to rectify false, inaccurate, or incomplete data collected about them, to remove themselves and their associated data from the data collection systems at any time, and to have input regarding what will happen to that data after the investigation (deletion, archiving, etc). They should also have recourse to a defined mechanism for raising concerns or
- See Decision Tree: “Should I share”, Step 3.

making complaints about the data collection effort. This mechanism should be resilient enough to remain accessible even after funding for the response effort has ended. These conversations should include a discussion of the tension between releasing the data publicly and keeping it protected, and the risks and benefits of each course of action. This dialogue must take place in a manner appropriate to the level of data literacy present in the community, elevating it where possible. Researchers should make every effort to validate their general findings and impressions with the local community before they leave. All considerations regarding community access to an stewardship of data should take the local legal context into account. After the crisis has passed, the community must be involved in the assessment phase of the response as described below.

- E. *Conduct assessments post-response:* Following the intervention and together with local partners, evaluate the response effort's use of location-based data in the context of measurable outcomes. Consider both positive and negative outcomes, and consider how the availability of location-based data –or lack thereof– affected these results. In instances where location-based data were helpful, determine whether the level of granularity was excessive, sufficient, or inadequate, and ensure that this information is available to the planners of future response efforts.

Principle #5: Give Access to Your Data - Share data openly, when safe and practical

- A. *Share data, but assess the risks and accept the consequences:* Data should only be shared once an analysis of the potential consequences of its dissemination has been completed, taking into account the nature of the data to be shared and the individual and/or group with which it would be shared. Define who is accountable for data-related harms, and establish mechanisms for addressing such harms which involve the local community. Know that people will misinterpret maps and visualizations, and design them to minimize misinterpretation. Actors tasked with implementing the sharing of data must be trained in current data security practice and understand the implications of sharing the data.
- B. *Assess the nature and form of the data to be shared:* The risks associated with the storage and sharing of data are likely to differ based on the type of data and its level of granularity, and the policies associated with sharing that data should be defined accordingly. Sharing of data should take place in the context of a clearly defined use case. In no instance should personally identifiable information be provided, including names, home addresses, phone numbers, IP addresses, or other obvious identifiers. Always remember that location-based data has the potential to be used as personally identifiable information even if it contains no such information explicitly. The form in which aggregated data are shared should reflect best practices and ensure that technical and administrative safeguards make re-identification difficult, if not impossible.

See Decision Tree: "Should I share", Step 2.

- C. *Assess the individuals and/or groups with which data will be shared:* Greater levels of disclosure, autonomy, and access to data may be allowed for highly trusted data recipients with strong, data security, audit, and access control processes and whose goals in using the data coincide with the purposes for which it was originally collected.
- D. *Ensure that the roles and responsibilities of key personnel related to data security are clearly defined:* When a project is initiated, a single individual or group should be designated to coordinate on all things related to data protection and privacy. Depending on the size of the project, a person or group responsible for data acquisition and licensing may also be designated. In both cases, these individuals or groups should be able to respond rapidly to critical events related to data security.
- E. *Have a plan in place to respond to a data breach:* Take measures to prevent location-based data in your possession from being hacked or shared accidentally. Discuss mitigation strategies in the event of a breach prior to beginning the collection effort, taking into account the specific nature of the data being collected. Have a communication plan in place, and ensure that the contact information for all relevant stakeholders is readily available throughout the organization. These should include, at a minimum, leaders in the communities with which you are working, as well as any relevant local, regional, and national level civil authorities. Be aware of the laws in place in the host country regarding personal information, and have at least one team member tasked with ensuring compliance.
- F. *Take extra caution in the context of a violent conflict:* Extra caution should be taken in collecting and sharing location-based data in situations where groups on the ground are in conflict. In such situations, location-based data may be more likely to lead to serious negative consequences for vulnerable populations if adversarial groups gain access to the data. This may include consulting with data protection experts and/or establishing a system of paths and gateways that allow data to be transferred effectively only in a deliberate and controlled way. Recognize that some data may be too sensitive to be shared until after an event is over.

See Decision Tree: “Should I share”, Step 4.

Appendix: Sample Risk Matrix Associated with Events Affecting Staff in a Hypothetical Response

	Low Danger (1)	Medium Danger (2)	High Danger (3)	Severe Danger (4)
Low Probability (1)	Disorganized attempts to spoof data (1)	Organized attempts to spoof data (2)	Disinformation campaign targets response (3)	Foreign military intervention (4)
Medium Probability (2)	Random Phishing Attacks (2)	Local population mistrusts response (4)	Targeted phishing attacks (6)	Targeted violence against response (8)
High Probability (3)	Volunteers collect data of varying quality (3)	Chain of command disrupted within security forces (4)	Disinformation campaign targets vulnerable populations (9)	N/A (12)
Certain (4)	Dataset contains errors/omissions (4)	Telecommunications disrupted (8)	N/A (12)	N/A (16)

Each cell in this matrix represents a possible event that could impact the response effort. The position of the event in the matrix is defined by the danger represented by the event and the likelihood of it taking place, with the risk defined as the product of those two factors. This is a simplified example; in an actual response, multiple matrices may be necessary for different phases of the operation, and multiple types of events may occupy the same cells in the matrix. In this particular matrix, values of twelve pass the “red line” of unacceptable risk.



Scientific Responsibility,
Human Rights and Law



Accelerating social good with artificial intelligence:

Insights from the Google AI Impact Challenge

The page features an abstract graphic design with several thick, curved lines in green, blue, yellow, and red. These lines are interspersed with dotted paths that curve and loop across the page. Some lines end in small dots of the same color. The overall composition is dynamic and modern.

Contents

- 2 Executive summary
- 4 Launching a call to changemakers
- 6 A view of the AI for social good landscape
- 13 Insights from our application review
- 28 Laying a foundation for growth
- 29 Catalog of common project submissions
- 37 Resources
- 39 Appendix

Executive summary

Why we wrote this report

In fall 2018, we announced the Google AI Impact Challenge, an open call for proposals on how to use artificial intelligence (AI) to help address society's most pressing issues. Twenty organizations received a total of \$25 million in grant funding from Google.org, coaching from Google's AI experts, credit and consulting from Google Cloud, and inclusion in a six-month Google Developers Launchpad Accelerator. In total, we received 2,602 submissions from six continents that together provide a view of the global AI for social good landscape. This publication shares insights gathered from these submissions, along with an extensive catalog of proposed projects, to help accelerate the way a variety of organizations can use AI to improve people's lives—and our planet.

Global momentum around AI for social good is growing—many organizations either are expressing interest in or are already using AI to address a wide array of societal challenges. We were impressed by the range of ideas we saw, and we're excited to share what we learned with the growing AI for social good community. As more social sector organizations recognize AI's potential, we all gain more high-impact opportunities to strengthen the emerging ecosystem.

Insights from our application review

As we reviewed the submissions and interviewed a portion of applicants, we uncovered trends about the state of AI for social good. While these trends are based only on a subset of organizations operating in this space, we believe that they point to significant opportunities for organizations seeking to use AI for social good.

Insight 1: Machine learning is not always the right answer.

Insight 2: Data accessibility challenges vary by sector.

Insight 3: Demand for technical talent has expanded from specialized AI expertise to data and engineering expertise.

Insight 4: Transforming AI insights into real-world social impact requires advance planning.

Insight 5: Most projects require partnerships to access both technical ability and sector expertise.

Insight 6: Many organizations are working on similar projects and could benefit from shared resources.

Insight 7: Organizations want to prioritize responsibility but don't know how.



The American University of Beirut is applying deep learning to agricultural data to help farmers in arid climates optimize irrigation systems.

In sharing these calls to action, we hope to inspire action and provoke conversation.

Opportunities

Nonprofits, academic institutions, social enterprises, funders, and technical partners—across both public and private sectors—all play a role in bolstering the developing AI for social good community. We also see an opportunity for strategic partnerships and collaboration among stakeholders, each of which brings a unique lens that can accelerate impact within this budding ecosystem.

As an extension of our insights, we have identified areas of opportunity for each stakeholder and have developed a set of proposed actions for those that are invested in advancing AI for social good.

In sharing these calls to action, we hope to inspire action and provoke conversation with illustrative examples. Of course, the most effective path forward will depend on many factors, such as geographic region, technical proficiency, resources at hand, the sector addressed, and organizational composition.

**We called.
The world answered.**

6

continents

119

countries

2,602

proposals

Launching a call to changemakers

In fall 2018, we launched the Google AI Impact Challenge, an open call to the world's nonprofits, social enterprises, and research institutions for proposals on how to use AI to help address society's most challenging, pressing, and important issues. Twenty organizations received a total of \$25 million in grant funding from Google.org, coaching from Google's AI experts, credit and consulting from Google Cloud, and inclusion in a six-month Google Developers Launchpad Accelerator. In total, we received 2,602 proposals from organizations from 119 countries spanning six continents for projects addressing a wide range of sectors, from environmental to education to humanitarian aid.

The landscape it revealed

The proposals reflected the wide range of organizations already working with AI methods and capabilities to address social challenges and looking to accelerate and scale their work. We also saw many applicants taking their first steps to use AI. In fact, 55% of the not-for-profit organizations and 40% of the for-profit social enterprises that applied reported no prior experience with AI. Their applications point to the tremendous potential in this nascent space.

Most applicants proposed to take advantage of the open-source machine learning libraries that help bring AI from the world of academia to the beginnings of mainstream use. As awareness of AI capabilities grows and barriers to entry lower, it's likely we will see even more social sector organizations enter the AI space.

Submissions to the Impact Challenge, along with our observations from the review process, paint an initial picture of the global AI for social good landscape today. They also illuminate opportunities for stakeholders to help catalyze the impactful and responsible future use of AI.

By sharing what we learned from this process, we hope to contribute to the accelerating use of AI as a meaningful tool for improving people's lives and solving some of society's biggest problems.

A view of the AI for social good landscape



Nexleaf Analytics is using machine learning to predict vaccine degradation and efficacy.

A growing global belief in the social promise of AI

44

countries had 10 or more applicants

61

countries had at least 5 applicants

A view of the AI for social good landscape

The Google AI Impact Challenge was a wide-reaching global call for ideas using AI for social good. The volume, diversity, quality, and creativity of submissions revealed a growing belief in AI as a promising tool for tackling difficult social challenges.

Many kinds of organizations from around the world are interested in using AI for social good

The geographic diversity of the applications highlighted AI's potential for addressing both global and local issues in developed and developing nations alike. We received applications from nearly two-thirds of the world's countries, across six continents.

Thirty-three percent of the applications came from the United States, 19% from Europe, 16% from Asia, 14% from the Middle East and Africa, 9% from Latin America, 6% from Canada, and 3% from Oceania. Forty-four different countries had 10 or more applicants, and 61 had at least 5 (see the appendix for a full list of countries with 5 or more applicants). It's important to note that this geographic distribution is likely not fully representative of

More than half of the applications came from organizations with fewer than 25 employees. This points to AI as a potential force multiplier in smaller organizations.

the overall distribution in the sector, as the open-call announcement and application were in English and relied on local nongovernmental organization networks for dissemination. Nonetheless, it's a promising indicator of the global interest in applying AI for social good.

The submissions were balanced between nonprofits, social enterprises, and academic institutions. The size of these organizations was also diverse, ranging from entrepreneurial two-person nonprofits to leading research facilities with thousands of staff. More than half of the applications came from organizations with fewer than 25 employees, showing that AI is an accessible and powerful tool for both small and large organizations. This points to AI as a potential force multiplier in smaller organizations.

AI can be used to address a wide array of societal challenges

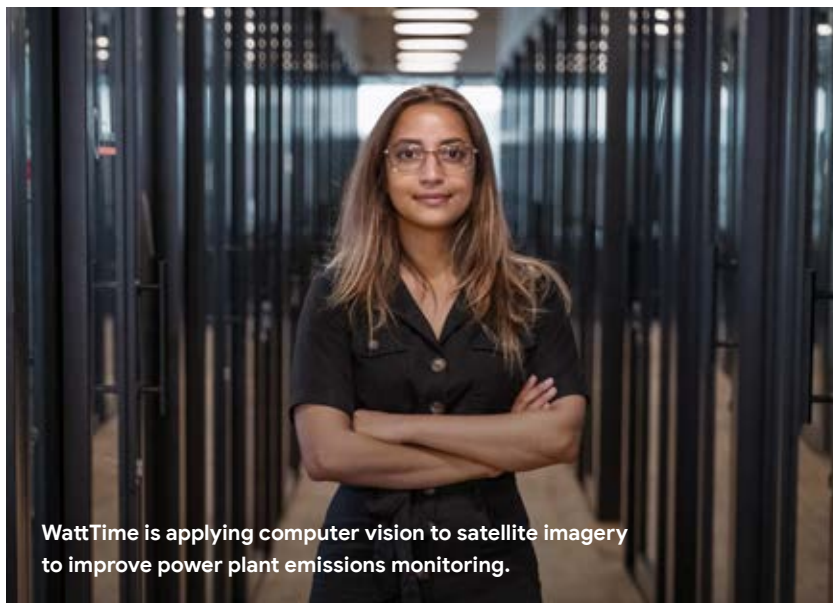
AI is flexible enough to extend beyond a single societal challenge or sector. In fact, proposals received as part of the AI Impact Challenge touched upon all 17 of the United Nations' Sustainable Development Goals.

Of the sectors represented, health-related applications were the most common, representing more than 25% of total submissions. Sixteen percent of applications sought to apply AI to environmental issues, followed by 12% for education, and 11% for economic development. Equality and inclusion, crisis response, and public and social sector topics also each received hundreds of submissions. The distribution of applications across sectors illustrates the versatility of AI capabilities and a broad-based interest within the social sector to leverage them.

For example, AI capabilities and techniques such as computer vision, deep learning, and natural language processing have the potential to enable new approaches to address social and economic issues. Across all major sectors mentioned above, computer vision was the most common AI capability chosen by applicants, with 41% referencing its use. Roughly a quarter of proposals sought to apply machine learning analytics, followed by nearly 20% each for structured deep learning and natural language processing.

41%

of applicants across all major sectors proposed to use computer vision, which was the most commonly referenced AI capability

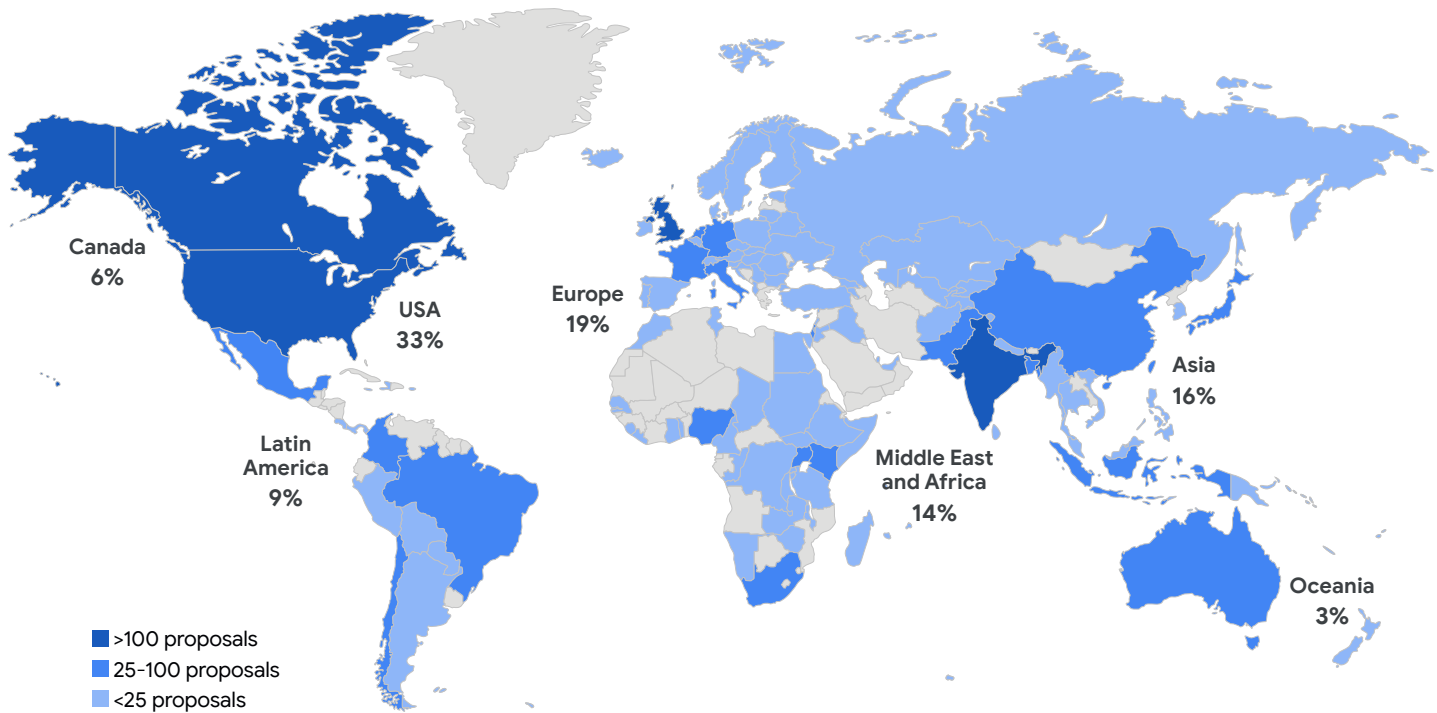


WattTime is applying computer vision to satellite imagery to improve power plant emissions monitoring.

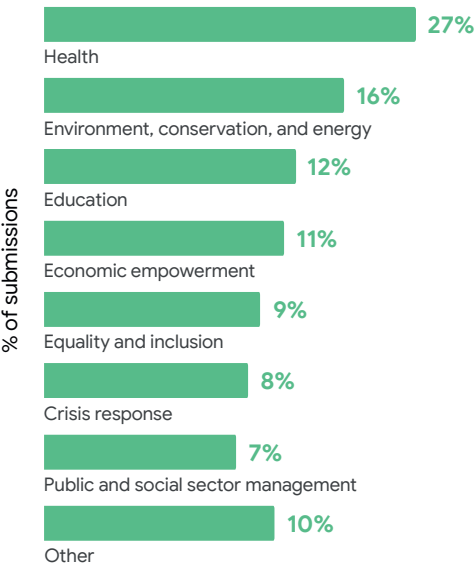
Common methods and capabilities leveraged by applicants

Core AI methods	
Rules-based solutions	Use explicitly stated rules to make decisions.
Machine learning solutions	<p>Learn without explicit programming, using examples, to develop a model that can make decisions.</p> <ul style="list-style-type: none"> • Deep learning: Using multiple layers of artificial neurons to create a network that can make a decision based on raw input. Applications of deep learning include computer vision and speech recognition.
AI-powered capabilities	
Audio processing	<p>Hear, recognize, and process sound files and other auditory inputs.</p> <ul style="list-style-type: none"> • Speech recognition: Using audio processing to translate human speech to text.
Computer vision	<p>See, recognize, and process images, videos, and other visual inputs.</p> <ul style="list-style-type: none"> • Object detection: Using computer vision to pick out and identify particular objects and/or physical properties. • Image and video classification: Using computer vision to understand and categorize or label visual inputs.
Machine learning analytics	Process and understand large volumes of data to identify patterns and make predictions.
Natural language processing	<p>Process, decipher, understand, and generate human language.</p> <ul style="list-style-type: none"> • Sentiment analysis: Using natural language processing to measure an author's or speaker's positivity or negativity.

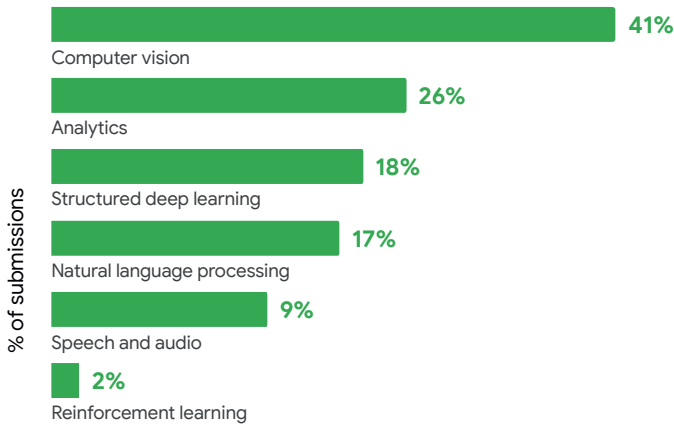
Global distribution of applications



Issue areas addressed in applications

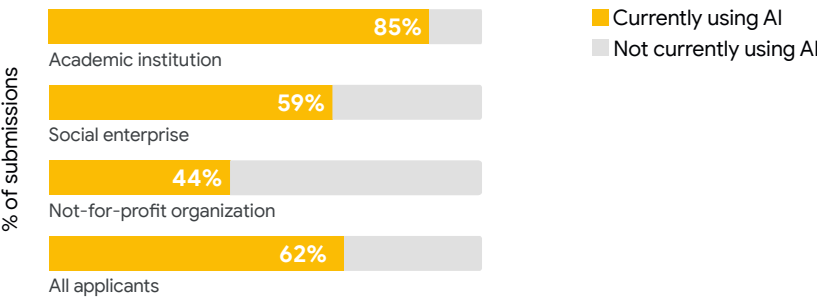


AI capability proposed in submission*



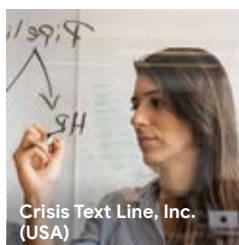
*Applicants could submit a proposal that leveraged more than one capability.

AI usage across applicants

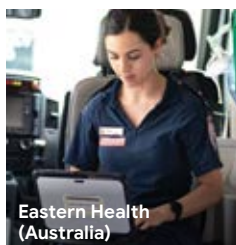




American University of Beirut
(Lebanon)



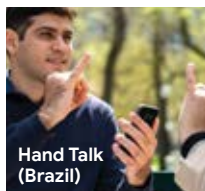
Crisis Text Line, Inc.
(USA)



Eastern Health
(Australia)



Fondation MSF
(France)



Hand Talk
(Brazil)

Google AI Impact Challenge grant recipients

The Google AI Impact Challenge culminated in the selection of 20 grantees. Combined, these organizations received \$25 million in grant funding from Google.org, coaching from Google's AI experts, credit and consulting from Google Cloud, and inclusion in a customized Google Developers Launchpad Accelerator program.

American University of Beirut (Lebanon): Applying machine learning to weather and agricultural data to improve irrigation for resource-strapped farmers in Africa and the Middle East

Colegio Mayor de Nuestra Señora del Rosario (Colombia): Using satellite imagery to detect illegal mines, enabling communities and the government to protect people and natural resources

Crisis Text Line, Inc. (USA): Using natural language processing to optimize the assignment of texters in crisis to counselors, reducing wait times and maintaining effective communication

Eastern Health (Australia): Analyzing clinical records from ambulances to uncover trends and potential points of intervention to inform policy and public health responses around suicide

Fondation MSF (France): Detecting patterns in antimicrobial imagery to help medical staff in low-resource areas prescribe the right antibiotics for bacterial infections

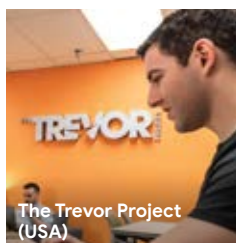
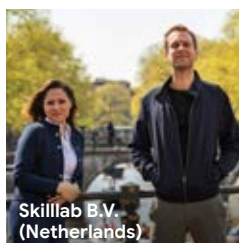
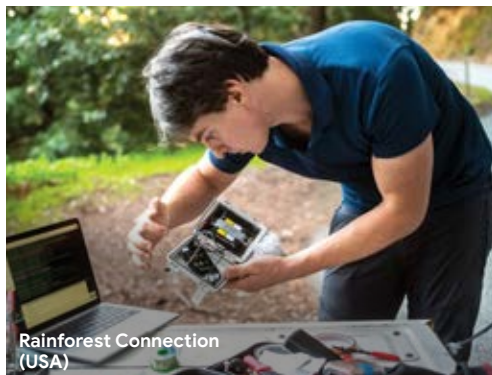
Full Fact (UK): Developing trend-monitoring and clustering tools to aid fact checkers' analysis so that they can help contextualize the news and enable informed decisions

Gringgo Indonesia Foundation (Indonesia): Building an image-recognition tool to improve plastic recycling rates, reduce ocean plastic pollution, and strengthen waste management in under-resourced communities

Hand Talk (Brazil): Using AI to translate Portuguese into Brazilian Sign Language through a digital avatar, enabling digital communication for Brazilians who are deaf or have partial hearing loss

HURIDOCS (Switzerland): Using natural language processing and machine learning to extract and connect relevant information in case-related documents, allowing human rights lawyers to effectively research and defend their cases

Makerere University (Uganda): Tracking and predicting air pollution patterns via low-cost sensors in Kampala, Uganda, improving air quality forecasting and intervention



New York University (USA): Partnering with the New York City Fire Department's analytics team to optimize response to its yearly 1.7 million emergencies, accounting for factors such as weather, traffic, and location

Nexleaf Analytics (USA): Building data models to predict vaccine viability throughout the cold vaccine supply chain and ensure effective delivery

Pennsylvania State University (USA): Using deep learning tools to better predict times of and locations at risk for landslides, creating a warning system to minimize the impact of natural disasters

Quill.org (USA): Using deep learning to provide low-income students with immediate feedback on their writing, enabling students to revise their work and quickly improve their skills

Rainforest Connection (USA): Using deep learning for bioacoustic monitoring and commonplace mobile technology to track rainforest health and detect threats

Skilllab B.V. (Netherlands): Helping refugees translate their skills to the European labor market and recommend relevant career pathways to explore

TalkingPoints (USA): Using AI to enable two-way translated parent-teacher engagement and coaching when language represents a barrier to communication

The Trevor Project (USA): Using natural language processing and sentiment analysis to determine an LGBTQ youth's suicide risk level to better tailor services for individuals seeking help

Wadhvani AI (India): Using image recognition to track and analyze pest-control efforts, enabling timely and localized intervention to stabilize crop production and reduce pesticide usage

WattTime (USA): Using image-processing algorithms and satellite networks to replace on-site power plant emissions monitors with open-source monitoring platforms

Insights from our application review



We believe these trends point to significant opportunities for organizations and individuals to apply AI for social good.

Insights from our application review

As we reviewed the submissions and interviewed a portion of applicants, we uncovered trends about the current state of AI for social good (see the appendix for the application review criteria). Some trends confirmed our existing assumptions, while others were more surprising. Although these trends are based only on the subset of organizations that applied, we believe they point to significant opportunities for organizations and individuals to apply AI for social good—and ways we all can help amplify their impact.

Insight 1: Machine learning is not always the right answer.

Insight 2: Data accessibility challenges vary by sector.

Insight 3: Demand for technical talent has expanded from specialized AI expertise to data and engineering expertise.

Insight 4: Transforming AI insights into real-world social impact requires advance planning.

Insight 5: Most projects require partnerships to access both technical ability and sector expertise.

Insight 6: Many organizations are working on similar projects and could benefit from shared resources.

Insight 7: Organizations want to prioritize responsibility but don't know how.

1 Machine learning is not always the right answer.

While we were excited by the number of organizations interested in exploring AI, throughout our evaluation process we identified opportunities for organizations to better understand what machine learning is and when it is an appropriate solution.

Some organizations submitted proposals that might be better implemented without machine learning by leveraging other methods that could result in faster, simpler, and cheaper execution. Other applicants underestimated the complexity of the work required, and in still other instances, machine learning is not yet sophisticated enough to make the proposed solution viable.

For example, several organizations proposed leveraging a deep learning platform to match individuals from an underserved population with the most appropriate legal knowledge and tools. While deep learning would be technically sufficient to achieve the desired outcome, similar results could be achieved faster and more cost-effectively through a rules-based system designed to recommend relevant content. Because rules-based systems are easier to understand and explain, the organizations could also help their intended audiences more quickly and effectively navigate the platform to find the most relevant resources.




HURIDOCS is using natural language processing of legal documents to help human rights lawyers effectively research and defend their cases.


INSIGHT 1: Machine learning is not always the right answer.


Opportunities


Implementers, policymakers, and funders need to be equipped to understand the potential value—and limitations—of AI, especially machine learning, to pressure test whether there is a faster, simpler, cheaper alternative to reach a proposed goal. Organizations and individuals with technical expertise play a critical role in providing educational resources and consulting with organizations on how best to leverage the right tools.

WHERE TO START:

 **Funders:** Learn more about the main methods and capabilities of AI, responsible AI practices, and the right questions to ask about data and implementation to better evaluate the feasibility of proposals using AI.

 **Organizations using AI for social good:** If technical expertise is needed to scope an AI project, reach out to organizations or individuals with that expertise to pressure test whether there is a faster, simpler, cheaper alternative.

 **Organizations with technical expertise:** Host workshops and other forums to equip social sector organizations interested in using AI to assess whether AI is the right fit for their project or challenge.

 **Policymakers:** Help boost public understanding of AI. For example, hold sessions with experts that are open to the public to facilitate more grounded and informed debate on key AI topics.

When can machine learning be useful?

When machine learning usually adds value	When machine learning may have limitations
<ul style="list-style-type: none">• Predicting future events when there are many parameters and it's not easy to see a pattern• Personalizing user experience• Understanding natural language• Categorizing objects or behaviors• Detecting infrequent events that change over time	<ul style="list-style-type: none">• Working with incomplete, limited, or inaccurate data• Maintaining predictability (e.g., placement of buttons in a user interface)• Offering complete “explainability”• Optimizing for speed and cost to market

Source: [Google's People + AI Guidebook, https://pair.withgoogle.com](https://pair.withgoogle.com)

2 Data accessibility challenges vary by sector.

Access to reliable and meaningful data is a consistent barrier for social sector organizations interested in applying AI methods and capabilities. Data access strategies from applicants fell into one of five categories:

1. First-party data that they had already collected (e.g., electronic health records, student records)
2. Publicly accessible third-party data (e.g., census data)
3. Purchasable third-party data (e.g., satellite images)
4. Third-party data accessible through a partnership
5. A design to collect first-party data

We observed meaningful variation in access to data across sectors as well as in the types of common datasets proposed.

Applicants in the crisis response, economic empowerment, and equality and inclusion categories were more likely to lack meaningful datasets. A significant number of crisis response submissions proposed to use accessible data sources such as satellite images and weather data but were limited by insufficient training data, as crisis events are not very common. For example, while satellite imagery is plentiful, there are not always enough labeled images of areas affected by a specific natural disaster to reliably train an image-classification model. The data challenges faced by economic empowerment and equality and inclusion proposals illustrate the difficulty in collecting large amounts of data from vulnerable populations that are often more transient, highly sensitive to privacy, and less likely to participate in the formal economy.

Applicants in health, environment, education, and the public and social sectors were more likely than other sectors to already have access to the necessary data. For these sectors, many of the data sources—for example, electronic health records in hospitals or student academic records in universities—are collected over the course of regular operations. The primary challenge for these organizations is acquiring access to the relevant data repositories. Without access to first-party data or public data sources, organizations need resources to purchase the respective data, such as satellite images, and/or time and assistance to broker the right partnership(s) for data access.

.....

The U.S. Department of Defense, Carnegie Mellon University, and CrowdAI are open-sourcing a labeled dataset containing

700K

before-and-after satellite images affected by eight different types of natural disasters.


.....


INSIGHT 2: Data accessibility challenges vary by sector.


Opportunities

In sectors where data already exists but is not easily accessible, organizations that own data have an opportunity to invest in data-sharing partnerships and responsible open-sourcing to allow other stakeholders to utilize this data (see [Google's approach to open data*](#)). In these cases, it will be important to consider privacy and security risks as well as potentially harmful use cases before sharing datasets broadly. In more data-sparse sectors, funders can help finance data collection. Funders and policymakers could leverage their resources and influence to support the collection and sharing of data, where appropriate.

WHERE TO START:

 **Funders:** Fund data collection or responsible aggregation, ideally incentivizing sharing across organizations.

 **Organizations using AI for social good:** Identify owned datasets that can be safely open-sourced or shared through data governance structures such as whitelists and data trusts.

 **Policymakers:** Use data to help boost research and development in AI.

- Create a framework that incentivizes and facilitates the creation, sharing, and reuse of datasets relevant to priority fields, in a manner that respects user expectations of privacy.
- Make more public datasets available, especially in priority subject realms for innovation.

Example: To help organizations that want to use machine learning to assess building damage related to disaster recovery, the U.S. Department of Defense, Carnegie Mellon University, and CrowdAI are open-sourcing a labeled dataset containing 700,000 satellite images of buildings before and after they were affected by eight different types of natural disasters. This dataset could train AI to help first responders safely navigate through areas recently affected by a natural disaster and prioritize areas most in need of aid.

[*https://www.blog.google/technology/ai/sharing-open-data/](https://www.blog.google/technology/ai/sharing-open-data/)

Commonly used datasets by project type

Sector	Commonly used datasets	
Crisis response	• Satellite images	• Social media (e.g., Twitter, Instagram)
Economic empowerment	• Business transaction records • Census and socioeconomic data	• Satellite images (agricultural yield) • Weather data (agricultural yield)
Education	• Student and school records	• User data from learning mobile applications
Environment	• Air, water, bioacoustic sensor data • Satellite images	• Weather data
Equality and inclusion	• Census data • Issue-specific public databases (e.g., from Organization for Economic Cooperation and Development, United Nations)	• Legal cases and outcomes • Survey data
Health	• Electronic health records (including medical imaging)	• Historical infectious disease outbreak records
Public sector	• Procurement data from individual countries	• Satellite images

3 Demand for technical talent has expanded from specialized AI expertise to data and engineering expertise.

In recent years, new machine learning software libraries and other open-source tools have reduced the technical overhead required to implement machine learning. More than 70% of submissions, across all sectors and organization types, relied on existing AI frameworks (e.g., Caffe, cuDNN, TensorFlow, PyTorch).

These tools enable individuals without deep backgrounds in machine learning to use best-in-class algorithms and practices. As a result, organizations have less need to hire specialized AI experts to build and use machine learning systems. With the burden of algorithm design and development removed, most of the necessary technical work is focused around cleaning and formatting data, as well as identifying key features, tools, and tuning parameters. Many organizations simply need analysts to manipulate the data and engineers to run the datasets through pre-existing algorithms.

These engineers need access to enough AI domain expertise to understand which algorithms to use; how to select, structure, and test training data; and how to test and mitigate for robust, responsible systems. While many applicants recognized the need for someone to play this data engineer role, throughout our review process, we saw that even the most mature organizations underestimated the time and resources needed to prepare and maintain the data for algorithm use.

Additionally, many applicants that were new to AI needed a better understanding of the types of data roles required, and others found it difficult to compete with private sector organizations when hiring technical talent.

Many organizations simply need analysts to manipulate the data and engineers to run the datasets through pre-existing algorithms.




Pennsylvania State University is using computer vision to create a warning system for landslides.


INSIGHT 3: Demand for technical talent has expanded from specialized AI expertise to data and engineering expertise.

Opportunities


While data analysts and data engineers are more accessible than AI experts, many of the applications, particularly from nonprofits, still cited access to technical expertise as a critical bottleneck. In the near term, organizations with technical expertise and funding can provide resources to fill the expertise needs. In the longer term, training and educational resources from online courses and bootcamps can help ameliorate the data talent shortage.

WHERE TO START:

 **Funders:** Provide funding to organizations to meet needs for data analysts and data engineers, and fill other gaps in technical expertise.

 **Organizations with technical expertise:** Create formal opportunities for employees, particularly data analysts and data engineers, to volunteer their technical skills for social sector projects.

Example: Google, IBM, and Microsoft have all developed formal programs to provide full-time technical expertise to nonprofit organizations. Smaller companies may not be able to dedicate full-time resources but could offer as-needed consults with nonprofit technical teams.

 **Policymakers:** Create incentives and programs for technical talent to support organizations that want to use AI for social good.

- Offer grants to support the development and provision of AI-oriented vocational training for people employed or seeking jobs in priority sectors.
- Offer training grants to encourage people from diverse backgrounds to learn about AI in order to bring fresh perspectives and wider community representation.
- Partner with industry in priority sectors to establish apprenticeship schemes to train the next generation of data scientists.

It would be beneficial to engage potential users early and often throughout the development process to understand user needs and how AI might help meet them.

4 Transforming AI insights into real-world social impact requires advance planning.

Many applicants had deep AI expertise and access to meaningful datasets but needed a clear implementation plan to translate insights and analysis into social impact.

For example, several organizations aimed to use satellite imagery to map in real time the damage caused by natural disasters such as earthquakes, fires, and floods, but they did not have a plan to share the model's insights with real-world programs that could act on these learnings to save lives. While it is easy to see the potential value of such a model—for example, improved warning systems and better distribution of aid—without properly defining a path toward implementation, that value cannot be fully realized. Moreover, it would be beneficial to engage potential users early and often throughout the development process to understand user needs and how AI might help meet them.

Academic organizations, in particular, should consider this challenge, as many of the proposals were led by professors focused on research and technical development. The American University of Beirut, for example, received a grant from the AI Impact Challenge for its work to save water by optimizing irrigation schedules through machine learning. The team has already developed relationships with local agricultural associations and intends to work with these partners to distribute the schedules to farmers in the form of a user-friendly mobile application.




Gringgo Indonesia Foundation is building an image-classification tool to improve waste management.

INSIGHT 4: Transforming AI insights into real-world social impact requires advance planning.

Opportunities


Before an AI system is developed, organizations need to plan for how they will operationalize it for impact and get meaningful feedback from users and beneficiaries. Identifying experts and partnering with affected communities to test and iterate ideas will lead to more impactful implementation of AI-powered solutions and avoid interesting research projects that may otherwise fail to reach their potential impact.

WHERE TO START:

 **Funders:** Ensure AI for social good projects have a clear path to impact and have the necessary resources and plans to engage users and the sector early in the design process.

 **Organizations using AI for social good:**

- For organizations focused on research or developing AI systems, invite implementers to research workshops and actively seek to work with organizations that can apply your research for real-world impact in tackling societal challenges.
- For organizations aiming to both create and implement the technology, develop your AI systems and implementation plan with frequent user testing and feedback from target beneficiaries and organizations working with these populations.

 **Policymakers:** Elevate the most important social and environmental challenges that need attention, and offer sectoral expertise to organizations developing solutions.

- Offer subsidies to support investment in the physical infrastructure underpinning AI in regions where it is lacking (e.g., discounts on electricity, faster CapEx depreciation).
- Look for tangible ways to make it easier for organizations to access AI capabilities, including through cloud-based services (e.g., by providing more flexible rules around data localization).
- Encourage universities to include training on applying AI across their curriculums, beyond engineering, so that the next generation of graduates to enter the industry is well equipped.

5

Most projects require partnerships to access both technical ability and sector expertise.

The most compelling proposals demonstrated fluency in both social sector and technical expertise. Applicants submitting such proposals fell into three broad categories:

1. Technologically savvy nonprofits and social enterprises that have in-house technical talent
2. Academic institutions that were willing to own their project and its implementation post-development
3. Partnerships between nonprofits with deep sector expertise and academic institutions or technology companies with the technical ability to shape and execute the AI portion of the project

On average, we saw that organizations with established partnerships had the most comprehensive applications and were best positioned to implement AI for social good.

Given the nascency of AI for social good work, few organizations fell into the first two categories. On average, we saw that organizations with established partnerships had the most comprehensive applications and were best positioned to implement AI for social good.

To accelerate impact, funders, technical partners, and policymakers should help equip organizations to be fluent in both AI and the sectors they are working to affect. Still, there may always be organizations with deep social sector expertise that will need to work with technical partners. Partnerships also face challenges, because nonprofits, technical partners, and academic institutions approach issues from different perspectives and may have different working cultures, priorities, and goals.

To provide an example of a mutually enriching partnership, one organization with a long history of working directly with refugees identified a need for better distribution of humanitarian aid. However, it lacked the technical expertise to create a machine learning forecasting model to inform the flow of aid from refugee camps with excess to those that were often short. This organization partnered with a company that specializes in applying advanced data analytics and AI tools for a variety of industries to create a prototype that improved forecasting of refugee supply needs. In return, the nonprofit provided the subject-matter expertise and distributed the findings to the relevant stakeholders.

INSIGHT 5: Most projects require partnerships to access both technical ability and sector expertise.

Opportunities

We believe partnerships between organizations with deep sector expertise and organizations with technical expertise offer the most actionable near-term opportunity to develop and operationalize the use of AI for social good. However, partnerships are not easy to find and develop. Common forums where organizations interested in using AI for social good could share missions and needs with technical experts may help facilitate connections.

WHERE TO START:



Funders: Host gatherings for grantees interested in using AI for social good to share current needs, and facilitate mutual introductions that may lead to strategic partnerships.



Policymakers: Create frameworks and forums to foster cross-sector collaboration on AI research.



Organizations using AI for social good: Have a clear understanding of your own strengths and limitations related to applying AI and developing potential partner profiles.

6

Many organizations are working on similar projects and could benefit from shared resources.

The applications we received highlighted many overlapping problem areas and solutions (see the catalog of [common project submissions](#) for examples). Given the obstacles we outline above around data accessibility and technical expertise, we believe that greater openness and collaboration can help jump-start initiatives by allowing resource-constrained organizations to draw from a shared pool of relevant knowledge and tools.

This kind of collaboration is by no means easy. Organizations may be hesitant to share their data and technology for a number of reasons, such as competition for a limited pool of grant funding. Even when there is a willingness to share knowledge, resources are needed to coordinate and organize this knowledge for maximum accessibility and to identify and mitigate privacy and security concerns. Nonetheless, we believe that promoting a culture of responsible openness can help accelerate programs that will help individuals and make the world better.

For example, we received more than 30 applications from all around the world proposing to use AI to identify and manage agricultural pests. For these applicants, sharing images of pests, relevant model code, and best practices could help improve the accuracy of models and reduce time spent on duplicated work for all. This allows the individual projects to focus on the unique value they are bringing to the space, be it their relationships with farmers in a certain region or understanding of pests unique to a certain area.

For tangentially related tasks, such as forecasting pest damage in upcoming agricultural seasons, sharing functions and tools (for example, by constructing a public dataset that contains detailed information about pest diet, migration patterns, or treatment options) will benefit all organizations working on these issues.

30

applications proposed using AI to identify and manage agricultural pests.

INSIGHT 6: Many organizations are working on similar projects and could benefit from shared resources.

Opportunities

A more open ecosystem requires both willingness to share responsibly and ease of access to these shared resources. Stakeholders can contribute to the open ecosystem in three ways: (a) share their own knowledge across multiple stages of AI implementation, (b) incentivize other organizations with knowledge to open-source, and (c) ensure that there are systems in place to make shared knowledge easily accessible for organizations it could benefit.

WHERE TO START:



Funders:

- Join together with key stakeholders in the AI for social good ecosystem to create a third-party body to develop open-sourcing best practices and aggregate open-source AI for social good projects.
- Require grantees to responsibly open-source funded projects (as Google.org asks of our grantees).



Policymakers: Increase access to publicly funded research and hold domain-specific research conferences for organizations to share their work on AI for social good.



Organizations using AI for social good:

Invest in responsible open-sourcing to share intellectual property (e.g., models and web and mobile applications), and share these investments with existing sector associations.

Collaboration and open-source opportunities across the stages of AI implementation



Data standards:

Organizations that seek to leverage tangential datasets may develop agreed-upon standards for quality, format, and reporting to help facilitate data interchange.



Datasets:

Organizations can responsibly share their underlying datasets with other organizations and, when possible, publicly accelerate training and accuracy of machine learning models.



Algorithms and models:

Relevant algorithms and models, and specifications on their design and capabilities, may also be shared with other organizations under open-source licenses or through APIs.



Web and mobile applications:

Many applications proposed web or mobile applications as a means to implement their models' insights. Open-sourcing these applications can provide a helpful reference point around best practices for other organizations.

Understanding and managing responsible AI use is both difficult and nonnegotiable.

7

Organizations want to prioritize responsibility but don't know how.

Any organizations that propose the use of AI need to understand how to implement AI responsibly and the potential risks of applying AI. We found that applicants varied significantly in their understanding of AI responsibility. Some had detailed plans to test datasets and model outputs for bias, protocols for handling personally identifiable information, and risk-mitigation plans for data security. Others focused on the promise of the technology and acknowledged the importance of responsibility but had no plans to ensure that their AI projects would be developed responsibly.

Understanding and managing responsible AI use is both difficult and nonnegotiable. As part of the review process, we evaluated all shortlisted applications for fit with the [Google AI principles](#).

These state that applications of AI should do the following:

1. Be socially beneficial
2. Avoid creating or reinforcing unfair bias
3. Be built and tested for safety
4. Be accountable to people
5. Incorporate privacy design principles
6. Uphold high standards of scientific excellence
7. Be made available for uses that accord with these principles

In addition to the above objectives, we will not design or deploy AI in the following application areas:

- Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks and will incorporate appropriate safety constraints
- Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people
- Technologies that gather or use information for surveillance violating internationally accepted norms
- Technologies whose purpose contravenes widely accepted principles of international law and human rights

We found that many organizations needed to more carefully consider risks related to creating or reinforcing unfair bias, incorporating privacy design principles, and mitigating risk of harmful use or misuse. In addition to the unfair bias and privacy challenges that any AI project might face, applicants using data from historically disadvantaged populations face particular challenges that require careful attention. For example, many applicants highlighted the need to check for unfair bias in models that were trained on a more general population and grappled with how to respect privacy by gathering consent from low-digital-literacy or transient groups.

Organizations also varied in their ability to evaluate potentially harmful use cases and identify ways to mitigate intentional or unintentional misuses of their technology. The organizations that did proactively identify these concerns proposed limiting the use of their models through governance structures, such as whitelists and data trusts.

Ultimately, organizations were interested in finding ways to implement technology responsibly and raised thoughtful questions. We hope that more social impact organizations using AI will help advance the conversation around responsibility, not only for the social sector, but also more broadly.

INSIGHT 7: Organizations want to prioritize responsibility but don't know how.

Opportunities

Ultimately, everyone has a critical role in ensuring responsibility.

WHERE TO START:



Funders: Ensure projects are vetted for potential responsibility concerns and that there is an ongoing mechanism for reviewing concerns with grantees.



Organizations using AI for social good:

- Have a clear idea of the responsibility guidelines to be followed (Google's AI principles are just one example of many).
- Where possible, make transparent modeling decisions and use transparent data collection methods to allow others to pressure test for responsible use of your technology.
- Engage a diverse set of stakeholders, including affected populations, to discuss potential risks and mitigations.

- Evaluate model performance across different dimensions that may highlight areas of unfair bias (e.g., different demographics).

- Develop a risk-mitigation plan for potential areas of harmful use or unintentional misuse.



Policymakers: Promote constructive governance frameworks and build responsible AI expertise in government bodies.

- Make government a role model for responsibly embracing AI.
- Encourage industry to share best practices and promote codes of conduct.
- Promote ethics training for government-funded researchers using machine learning.



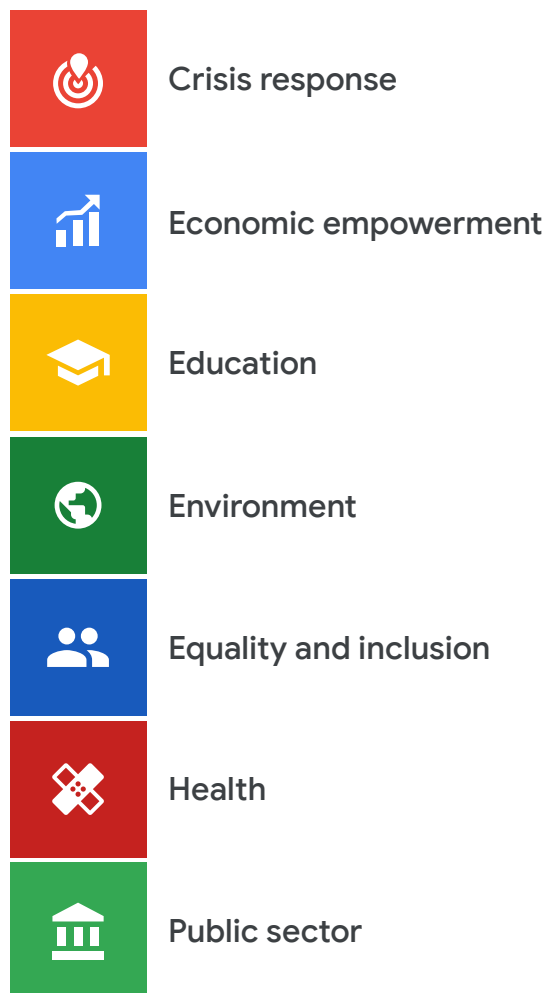
Wadhvani AI is using image classification to detect crop pests and inform local pest-control efforts.

Laying a foundation for growth

At Google, we are incredibly excited about the prospect of discovering how to better use AI responsibly and to support the growing AI for social good ecosystem to bring substantial, lasting quality-of-life improvement to people across all communities. It is our hope that the findings and insights in this report will help spark new progress and collaboration in this nascent space. Together, we have an opportunity to lay a foundation for AI to support social good actors in their missions.

Catalog of common project submissions

The following catalog highlights some of the common project types we saw from applicants. We share it with the hope that this information will help inform potential areas of collaboration and investment that can benefit projects at scale.





COMMON PROJECT SUBMISSIONS

Crisis response

Common project types	AI capabilities used	Data sources
Natural and man-made disasters		
Optimizing natural disaster relief efforts and funding	<ul style="list-style-type: none">• Machine learning analytics	<ul style="list-style-type: none">• Emergency response organization internal data (e.g., emergency calls)• Satellite images
Predicting earthquakes, floods, wildfires, and other catastrophic events	<ul style="list-style-type: none">• Deep learning• Image classification• Object detection	<ul style="list-style-type: none">• Data on historical occurrences• Emergency response organization internal data• Public international/regional data• Satellite images• Weather data
Real-time monitoring of extent and severity of damage	<ul style="list-style-type: none">• Image classification• Natural language processing — sentiment analysis• Object detection	<ul style="list-style-type: none">• Satellite images• Social media (e.g., Twitter, Instagram)



COMMON PROJECT SUBMISSIONS

Economic empowerment

Common project types	AI capabilities used	Data sources
Agricultural quality and yield		
Conserving and measuring soil health	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Sensor data• Weather data
Forecasting, identifying, and managing crop damage (pests, disease, etc.)	<ul style="list-style-type: none">• Image classification• Machine learning analytics• Object detection	<ul style="list-style-type: none">• Beneficiary-generated data (e.g., use of mobile application to take pictures of crops)
Guidance on the right crop to grow	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Sensor data• Weather data
Obtaining estimates for irrigation, fertilizer, and pesticide usage	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Sensor data• Weather data
Financial inclusion		
Improving access to credit for financially excluded individuals and communities	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• External transaction data (e.g., utility payments, mobile phone usage)• Financial organization internal transaction data• Historical borrower repayment rates
Skills supply and demand matching		
Identifying local job market needs to improve youth employability training	<ul style="list-style-type: none">• Deep learning• Natural language processing	<ul style="list-style-type: none">• Individual résumés• National labor statistics• Public local job advertisements• Skill ontologies (e.g., ESCO)



COMMON PROJECT SUBMISSIONS

Education

Common project types	AI capabilities used	Data sources
Access and completion of education		
Creating learning tools and emotional training for individuals with autism spectrum disorder	<ul style="list-style-type: none">• Facial recognition• Speech processing	<ul style="list-style-type: none">• Faces and poses in images and videos• User-generated data from mobile applications
Identifying at-risk students	<ul style="list-style-type: none">• Deep learning• Machine learning analytics• Natural language processing	<ul style="list-style-type: none">• Student and district records
Mapping out school locations in developing regions	<ul style="list-style-type: none">• Image classification• Natural language processing• Object detection	<ul style="list-style-type: none">• Census data• National and regional poverty, employment, health outcomes• Publicly available articles and news about specific schools• Publicly available United Nations Statistics Division data• Satellite images
Supporting interactive language learning, especially through mobile applications	<ul style="list-style-type: none">• Machine learning analytics• Speech processing	<ul style="list-style-type: none">• Existing language curriculum• User-generated data from mobile applications
Maximizing student achievement		
Career coaching	<ul style="list-style-type: none">• Machine learning analytics• Natural language processing	<ul style="list-style-type: none">• Skills ontology• Student records• Survey data
Providing adaptive learning applications for individual students	<ul style="list-style-type: none">• Deep learning• Machine learning analytics• Natural language processing	<ul style="list-style-type: none">• Student records• User-generated data from mobile applications
Teacher and administration productivity		
Grading and feedback to improve skills	<ul style="list-style-type: none">• Deep learning• Machine learning analytics• Natural language processing	<ul style="list-style-type: none">• Past graded papers, projects, homework



COMMON PROJECT SUBMISSIONS

Environment

Common project types	AI capabilities used	Data sources
Climate change		
Estimating causal effects of air pollution (e.g., health outcomes, severe storms)	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Insurance claims data• Publicly available community demographic data• Publicly available emissions data (e.g., from the Environmental Protection Agency)• Weather and storm data (e.g., Federal Emergency Management Agency maps)
Estimating urban air pollution	<ul style="list-style-type: none">• Deep learning• Image classification• Object detection	<ul style="list-style-type: none">• Air sensor data• Publicly available emissions data (e.g., from the Environmental Protection Agency)• Satellite images• Traffic flow data
Tracking concentrated methane emissions	<ul style="list-style-type: none">• Deep learning• Image classification• Object detection	<ul style="list-style-type: none">• Air sensor data• Hyperspectral imaging• Publicly available data on known methane leaks• Satellite images
Conservation		
Identifying illegal fishing vessels	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Field monitoring videos and images• Satellite images
Identifying illegal mining	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Satellite images
Monitoring ecological communities (e.g., endangered animals and habitats)	<ul style="list-style-type: none">• Audio processing• Deep learning• Image classification• Object detection	<ul style="list-style-type: none">• Bioacoustic sensor sound data• Field monitoring videos and images
Providing land-cover classification for precision conservation (e.g., water, deforestation, coral reefs)	<ul style="list-style-type: none">• Deep learning• Image classification• Object detection	<ul style="list-style-type: none">• Satellite images
Waste management		
Identifying and monitoring plastic debris (water and land)	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Existing street images (e.g., Google Earth's Street View)• Satellite images• User-generated pictures and videos through mobile application



COMMON PROJECT SUBMISSIONS

Equality and inclusion

Common project types	AI capabilities used	Data sources
Accessibility and disabilities		
Assisting in environment-sensing for the visually impaired	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Images and videos of everyday objects
Translating voice and text to sign language	<ul style="list-style-type: none">• Audio processing• Natural language processing	<ul style="list-style-type: none">• Videos of people speaking
Fair analysis in criminal proceedings		
Identifying bias and stereotypes	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Historical court data, including case outcomes
Human exploitation		
Predicting human trafficking patterns related to recruitment and transactions	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Arrest and indictment records• Federal, state, and local human trafficking cases• Police narratives• Publicly available U.S. Department of Labor enforcement data
Migrants and refugees		
Forecasting demand for aid within refugee camps	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Census data• International Organization for Migration datasets (e.g., displacement tracking)• Survey data• Weather data
Matching jobs to skills	<ul style="list-style-type: none">• Deep learning• Natural language processing	<ul style="list-style-type: none">• Individual résumés• National labor statistics• Public local job advertisements• Skill ontologies (e.g., ESCO)
Supporting the immigration and asylum processes with multilingual chatbots	<ul style="list-style-type: none">• Audio processing• Natural language processing	<ul style="list-style-type: none">• Survey data• Unstructured text and audio data from current requests



COMMON PROJECT SUBMISSIONS

Health

Common project types	AI capabilities used	Data sources
Diagnosis		
Detecting diseases (e.g., cancer, neurodegenerative conditions) using CT and MRI images	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Electronic health records
Detecting diseases using images from mobile cameras (e.g., skin conditions)	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• User-generated data from mobile application
Mental health		
Directing those with mental health questions to appropriate resources via chatbot	<ul style="list-style-type: none">• Natural language processing — sentiment analysis	<ul style="list-style-type: none">• Unstructured text and audio data from current requests
Triaging individuals at most imminent risk of suicide	<ul style="list-style-type: none">• Natural language processing — sentiment analysis• Speech processing	<ul style="list-style-type: none">• SMS conversations• Social media• Speech recordings with at-risk individuals
Outbreaks and epidemics		
Forecasting outbreaks and spread of infectious disease (e.g., malaria, dengue fever)	<ul style="list-style-type: none">• Deep learning	<ul style="list-style-type: none">• Publicly available records of past outbreaks (e.g., HealthMap)• Spatial data on population density and community demographics• Spatiotemporal disease data
Prevention		
Developing patient risk profiles (e.g., chronic diseases, pregnancies)	<ul style="list-style-type: none">• Deep learning• Machine learning analytics	<ul style="list-style-type: none">• Electronic health records
Treatment		
Personalizing treatment plans	<ul style="list-style-type: none">• Machine learning analytics	<ul style="list-style-type: none">• Electronic health records
Predicting surgical outcomes to improve surgery standardization	<ul style="list-style-type: none">• Machine learning analytics	<ul style="list-style-type: none">• Electronic health records• Hospital surgical outcome records



COMMON PROJECT SUBMISSIONS

Public sector

Common project types	AI capabilities used	Data sources
Citizen services		
Aiding citizen inquiries and customer service for government services	<ul style="list-style-type: none">• Natural language processing	<ul style="list-style-type: none">• Unstructured text data from current requests
Corruption		
Monitoring public procurement processes for corruption	<ul style="list-style-type: none">• Machine learning analytics• Natural language processing	<ul style="list-style-type: none">• Procurement data from individual countries
Land use		
Forecasting urban growth to plan for sustainable land use	<ul style="list-style-type: none">• Image classification• Object detection	<ul style="list-style-type: none">• Satellite images

Resources

Machine learning is not always the right answer

Using AI for Social Good: This guide helps nonprofits and social enterprises understand the types of problems their organizations can solve with machine learning, learn how to identify and prepare data sources, and review guidelines on how to develop and utilize machine learning responsibly. <https://ai.google/education/social-good-guide>

Machine Learning Crash Course with TensorFlow APIs: A self-study guide for aspiring machine learning practitioners. <https://developers.google.com/machine-learning/crash-course>

PAIR guidebook: Originally written for user experience (UX) professionals and product managers as a way to help create a human-centered approach to AI on their product teams, this guidebook is useful to anyone in any role wanting to build AI products in a more human-centered way. <https://pair.withgoogle.com>

Responsible AI

Building Responsible AI for Everyone: Reliable, effective, user-centered AI systems should be designed following general best practices for software systems, together with practices that address considerations unique to machine learning. Google's top recommendations are located here, with additional resources for further reading. <https://ai.google/responsibilities/responsible-ai-practices>

Responsible Development of AI: In this white paper, we provide an overview of our approach to responsible use and development of AI and share recommendations for government policy frameworks. <https://ai.google/static/documents/responsible-development-of-ai.pdf>

Facets tool: Facets contains visualizations to aid in understanding and analyzing machine learning datasets. Get a sense of the distribution of features in your datasets using Facets Overview, or explore individual data points using Facets Dive. <https://pair-code.github.io/facets>

What-If tool: The What-If tool makes it easy to efficiently and intuitively explore a model's performance on a dataset. Evaluate your model's performance as you manipulate features on individual data points, and explore various optimization strategies. <https://pair-code.github.io/what-if-tool>

Data sources

Google Cloud Platforms datasets: The Google Cloud Public Datasets program hosts copies of structured and unstructured data to make it easier for users to discover, access, and utilize public data in the cloud. These datasets are hosted for free. <https://console.cloud.google.com/marketplace>

Kaggle datasets: Explore, execute, share, and comment on code for any open dataset with the in-browser analytics tool, Kaggle Kernels. You can also download datasets in an easy-to-read format. <https://www.kaggle.com/datasets>

Google Earth Engine catalog: Earth Engine's public data archive includes more than 40 years of historical imagery and earth science datasets, updated and expanded daily. <https://developers.google.com/earth-engine/datasets/catalog>

Dataset Search: Google's Dataset Search tool allows users to search across thousands of dataset repositories on the web. Once you've found a relevant dataset, Dataset Search will direct you to the repository or provider where the data is hosted. <https://toolbox.google.com/datasetsearch>

Data Commons: Data Commons attempts to synthesize a single Open Knowledge Graph from different data sources. It links references to the same entities (cities, counties, organizations, etc.) across different datasets to nodes on the graph so that users can access data about a particular entity aggregated from different sources without data cleaning or joining. <https://www.datacommons.org>

Appendix

Funders

How can I help support AI for social good?

Funders play a vital role in bolstering the developing AI for social good community. But it's not always easy to know where to start.

As part of our work, we've identified some ways that funders can make a difference:

- Learn more about the main methods and capabilities of AI, responsible AI practices, and the right questions to ask about data and implementation to better evaluate the feasibility of proposals using AI.
- Fund data collection or responsible aggregation, ideally incentivizing sharing across organizations.
- Provide funding to organizations to meet needs for data analysts and data engineers, and fill other gaps in technical expertise.
- Ensure AI for social good projects have a clear path to impact and have the necessary resources and plans to engage users and the sector early in the design process.
- Host gatherings for grantees interested in using AI for social good to share current needs and facilitate mutual introductions that may lead to strategic partnerships.
- Join together with key stakeholders in the AI for social good ecosystem to create a third-party body to develop open-sourcing best practices and aggregate open-source AI for social good projects.
- Require grantees to responsibly open-source funded projects (as Google.org asks of our grantees).
- Ensure projects are vetted for potential responsibility concerns and that there is an ongoing mechanism for reviewing concerns with grantees.

Organizations

How can I help support AI for social good?

Organizations that use or are interested in using AI for social good and organizations with technical expertise in AI both play a vital role in bolstering and developing the community. But it's not always easy to know where to start.

As part of our work, we've identified some ways that organizations can make a difference:

Organizations interested in using AI for social good

- If technical expertise is needed to scope an AI project, reach out to organizations or individuals with that expertise to pressure test whether there is a faster, simpler, cheaper alternative.
- Identify owned datasets that can be safely open-sourced or shared through data governance structures such as whitelists and data trusts.
- For organizations focused on research or developing AI systems, invite implementers to research workshops and actively seek to work with organizations that can apply your research for real-world impact in tackling societal challenges.
- For organizations aiming to both create and implement the technology, develop your AI systems and implementation plan with frequent user testing and feedback from target beneficiaries and organizations working with these populations.
- Have a clear understanding of your own strengths and limitations related to applying AI and developing potential partner profiles.
- Invest in responsible open-sourcing to share intellectual property (e.g., models and web and mobile applications), and share these investments with existing sector associations.

- Have a clear idea of the responsibility guidelines to be followed (Google's AI principles are just one example of many).
- Where possible, make transparent modeling decisions and use transparent data collection methods to allow others to pressure test for responsible use of your technology.
- Engage a diverse set of stakeholders, including affected populations, to discuss potential risks and mitigations.
- Evaluate model performance across different dimensions that may highlight areas of unfair bias (e.g., different demographics).
- Develop a risk-mitigation plan for potential areas of harmful use or unintentional misuse.

Organizations with technical AI expertise

- Host workshops and other forums to equip social sector organizations interested in using AI to assess whether it's the right fit for their project or challenge.
- Create formal opportunities for employees, particularly data analysts and data engineers, to volunteer their technical skills for social sector projects.

Policymakers

How can I help support AI for social good?

Policymakers play a vital role in bolstering the developing AI for social good community. But it's not always easy to know where to start.

As part of our work, we've identified some ways that policymakers can make a difference:

-
- Help boost public understanding of AI (e.g., invite the public to attend sessions with experts to facilitate more grounded and informed debate on key AI topics).
 - Create a framework that incentivizes and facilitates the creation, sharing, and reuse of datasets relevant to priority fields, in a manner that respects user expectations of privacy.
 - Make more public datasets available, especially in priority subject realms for innovation.
 - Create incentives and programs for technical talent to support organizations that want to use AI for social good.
 - Offer grants to support the development and provision of AI-oriented vocational training for people employed or seeking jobs in priority sectors.
 - Offer training grants to encourage people from diverse backgrounds to learn about AI in order to bring fresh perspectives and wider community representation.
 - Partner with industry in priority sectors to establish apprenticeship schemes to train the next generation of data scientists.
 - Elevate the most important social and environmental challenges that need attention, and offer sectoral expertise to organizations developing solutions.
 - Offer subsidies to support investment in the physical infrastructure underpinning AI in regions where it is lacking (e.g., discounts on electricity, faster CapEx depreciation).
 - Look for tangible ways to make it easier for organizations to access AI capabilities, including through cloud-based services (e.g., by providing more flexible rules around data localization).
 - Encourage universities to include training on applying AI across their curriculums, beyond engineering, so that the next generation of graduates to enter the industry is well equipped.
 - Create frameworks and forums to foster cross-sector collaboration on AI research.
 - Increase access to publicly funded research, and hold domain-specific research conferences for organizations to share their work on AI for social good.
 - Promote constructive governance frameworks, and build responsible AI expertise in government bodies:
 - Make government a role model for responsibly embracing AI.
 - Encourage industry to share best practices and promote codes of conduct.
 - Promote ethics training for government-funded researchers using machine learning.

List of countries that submitted five or more applications

>100 proposals

Canada
India
United Kingdom
United States

25–100 proposals

Australia	Italy
Bangladesh	Japan
Brazil	Kenya
Chile	Mexico
China	Netherlands
Colombia	Nigeria
France	Pakistan
Germany	South Africa
Indonesia	Uganda
Israel	

<25 proposals

Argentina	Malaysia	Slovakia
Austria	Morocco	Slovenia
Belgium	Nepal	South Korea
Bulgaria	New Zealand	Spain
Cameroon	Panama	Sweden
Czechia	Peru	Switzerland
Denmark	Philippines	Tanzania
Egypt	Poland	Thailand
Finland	Portugal	Turkey
Ghana	Romania	Ukraine
Hong Kong	Russia	United Arab Emirates
Ireland	Serbia	Vietnam
Lebanon	Singapore	

AI for social good projects tackling the United Nations' Sustainable Development Goals

UN SDGs	Examples of AI for social good projects
 <p>1 NO POVERTY</p>	Applying machine learning on satellite imagery and survey data to extract socioeconomic indicators and generate visualizations and predictions of poverty in areas without survey data
 <p>2 ZERO HUNGER</p>	Applying machine learning to satellite images, localized food prices, and conflict data to predict and address severe acute childhood malnutrition
 <p>3 GOOD HEALTH AND WELL-BEING</p>	Using machine learning analytics on vaccine transit data, vaccine potency data, temperature data in vaccine refrigerators, equipment metadata, and other datasets to predict viability for common vaccines at every point in the supply chain
 <p>4 QUALITY EDUCATION</p>	Applying natural language processing on conversations between educators and non-English-speaking parents as well as machine learning analytics on parent and student profiles to construct a multilingual family engagement platform and deliver personalized resources that help parents digitally connect with teachers regardless of language
 <p>5 GENDER EQUALITY</p>	Leveraging natural language processing methods, including topic modeling, psycholinguistic feature modeling, and audio signal processing on voice recordings and chat transcripts from crisis call hotlines for women to escalate calls with high risk of intimate partner violence
 <p>6 CLEAN WATER AND SANITATION</p>	Leveraging machine learning analytics on thermal satellite images, weather data, and farmer-supplied agriculture data to estimate evapotranspiration and help farmers optimize the exact amount of water needed to irrigate crops
 <p>7 AFFORDABLE AND CLEAN ENERGY</p>	Utilizing machine learning analytics and image recognition on satellite images, power grid outlines, and relevant socioeconomic information to determine optimal resource allocation for electricity infrastructure in developing countries
 <p>8 DECENT WORK AND ECONOMIC GROWTH</p>	Using machine learning analytics on publicly available skills and occupational data to map an individual's skill set captured through a guided assessment directly to relevant occupations
 <p>9 INDUSTRY, INNOVATION AND INFRASTRUCTURE</p>	Using computer vision on Google Street View and LIDAR images to help residents to assess defensible space and identify flammable vegetation around their homes

AI for social good projects tackling the United Nations' Sustainable Development Goals (cont.)

UN SDGs	Examples of AI for social good projects
10 REDUCED INEQUALITIES 	Applying machine learning to image and text data to develop a dictionary of norm-revealing phrases that can be used as an alternative credit-scoring mechanism to make consumer lending more accessible to low-income individuals
11 SUSTAINABLE CITIES AND COMMUNITIES 	Applying machine learning analytics on incident dispatch data and other correlative data (weather, anomalous events, city demographics, etc.) to build a predictive model around emergency response times for first responders in urban areas
12 RESPONSIBLE CONSUMPTION AND PRODUCTION 	Leveraging deep neural net and image-recognition technology on garbage and waste management images to help automate classification and sorting of recyclable items at waste management facilities
13 CLIMATE ACTION 	Applying machine learning analytics and computer vision on emissions data, satellite imagery of power plants, weather conditions, and grid conditions to monitor power plant emissions
14 LIFE BELOW WATER 	Using image-recognition technology on waste facility images and data to help increase recycling rates and reduce ocean plastic pollution
15 LIFE ON LAND 	Applying machine learning analytics and audio recognition on live audio streams from rainforests and other ecosystems to help locals derive insights and help root out any ecological threats
16 PEACE, JUSTICE AND STRONG INSTITUTIONS 	Using natural language processing and machine learning methods on legal and judicial documents (e.g., laws, jurisprudence, victim testimonies, and resolutions) to extract relevant information and empower human rights advocates
17 PARTNERSHIPS FOR THE GOALS 	Leveraging machine learning analytics on developing country economic indicators and population survey data to create a repository to help inform public and private sector entities as well as the government to better target assistance and measure return on investment

Google AI Impact Challenge application review criteria

The submissions were reviewed by over 200 technical and issue experts within and outside of Google. Each submission was evaluated across five criteria associated with high-potential, high-impact projects.

Potential for impact	The project has a clear path to significant, real-world social impact and is grounded in research and data about the problem and solution.
Appropriate use case for AI	It is clear how using AI uniquely allows the project to have impact (i.e., introduces a solution that otherwise would not be possible). There is a plan to deploy an AI model and a strong link between the AI model's deployment and project goals.
Feasibility	The team has a well-developed, realistic plan to execute on the proposal, including access to a meaningful dataset as well as to domain and technical expertise and a robust implementation plan. The proposal correctly identifies the biggest risks and how they can be mitigated.
Scalability	The project can scale its impact or will serve as a model for other efforts to advance the field.
Responsibility	The project is consistent with the principles laid out in Google's AI principles, making use of recommended technical practices in the responsible AI practices. Potential ethical concerns are recognized, and mitigation strategies are proposed.



Google AI Impact Challenge

September 10, 2019



AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations

Luciano Floridi^{1,2} · Josh Cowsls^{1,2} · Monica Beltrametti³ · Raja Chatila^{4,5} · Patrice Chazerand⁶ · Virginia Dignum^{7,8} · Christoph Luetge⁹ · Robert Madelin¹⁰ · Ugo Pagallo¹¹ · Francesca Rossi^{12,13} · Burkhard Schafer¹⁴ · Peggy Valcke^{15,16} · Effy Vayena¹⁷

Received: 28 October 2018 / Accepted: 2 November 2018 / Published online: 26 November 2018
© The Author(s) 2018

Abstract

This article reports the findings of AI4People, an Atomium—EISMD initiative designed to lay the foundations for a “Good AI Society”. We introduce the core opportunities and risks of AI for society; present a synthesis of five ethical principles that should undergird its development and adoption; and offer 20 concrete recommendations—to assess, to develop, to incentivise, and to support good AI—which in some cases may be undertaken directly by national or supranational policy makers, while in others may be led by other stakeholders. If adopted, these recommendations would serve as a firm foundation for the establishment of a Good AI Society.

Keywords Artificial intelligence · AI4People · Data governance · Digital ethics · Governance · Ethics of AI

1 Introduction

AI is not another utility that needs to be regulated once it is mature. It is a powerful force, a new form of smart agency, which is already reshaping our lives, our interactions, and our environments. AI4People was set up to help steer this powerful force towards the good of society, everyone in it, and the environments we share. This article is the outcome of the collaborative effort by the AI4People Scientific

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

Extended author information available on the last page of the article

Committee—comprising 12 experts and chaired by Luciano Floridi¹—to propose a series of recommendations for the development of a Good AI Society.

The article synthesises three things: the *opportunities* and associated *risks* that AI technologies offer for fostering human dignity and promoting human flourishing; the *principles* that should undergird the adoption of AI; and 20 specific *recommendations* that, if adopted, will enable all stakeholders to seize the opportunities, to avoid or at least minimise and counterbalance the risks, to respect the principles, and hence to develop a Good AI Society.

The article is structured around four more sections after this introduction. Section 2 states the core opportunities for promoting human dignity and human flourishing offered by AI, together with their corresponding risks.² Section 3 offers a brief, high-level view of the advantages for organisations of taking an ethical approach to the development and use of AI. Section 4 formulates 5 ethical principles for AI, building on existing analyses, which should undergird the ethical adoption of AI in society at large. Finally, Sect. 5 offers 20 recommendations for the purpose of developing a Good AI Society in Europe.

Since the launch of AI4People in February 2018, the Scientific Committee has acted collaboratively to develop the recommendations in the final section of this paper. Through this work, we hope to have contributed to the foundation of a Good AI Society we can all share.

2 The Opportunities and Risks of AI for Society

That AI will have a major impact on society is no longer in question. Current debate turns instead on how far this impact will be positive or negative, for whom, in which ways, in which places, and on what timescale. Put another way, we can safely dispense with the question of *whether* AI will have an impact; the pertinent questions now are *by whom, how, where, and when* this positive or negative impact will be felt.

In order to frame these questions in a more substantive and practical way, we introduce here what we consider the four chief opportunities for society that AI offers. They are four because they address the four fundamental points in the understanding of human dignity and flourishing: *who we can become* (autonomous self-realisation); *what we can do* (human agency); *what we can achieve* (individual and societal capabilities); and *how we can interact with each other and the world* (societal cohesion). In each case, AI can be *used* to foster human nature and its potentialities, thus creating opportunities; *underused*, thus creating opportunity costs; or *overused* and *misused*, thus creating risks. As the terminology indicates, the assumption

¹ Besides Luciano Floridi, the members of the Scientific Committee are: Monica Beltrametti, Raja Chaitila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Josh Cowls is the rapporteur. Thomas Burri contributed to an earlier draft.

² The analysis in this and the following two sections is also available in Cowls and Floridi (2018). Further analysis and more information on the methodology employed will be presented in Cowls and Floridi (Forthcoming).

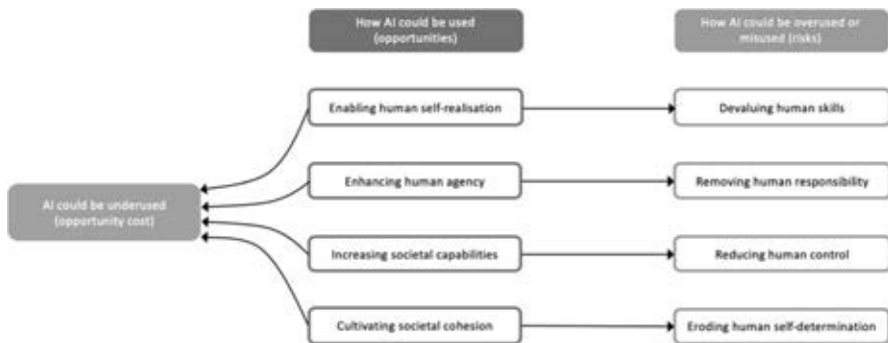


Fig. 1 Overview of the four core opportunities offered by AI, four corresponding risks, and the opportunity cost of underusing AI

is that the *use* of AI is synonymous with good innovation and positive applications of this technology. However, fear, ignorance, misplaced concerns or excessive reaction may lead a society to *underuse* AI technologies below their full potential, for what might be broadly described as the wrong reasons. This may cause significant opportunity costs. It might include, for example, heavy-handed or misconceived regulation, under-investment, or a public backlash akin to that faced by genetically modified crops (Imperial College 2017). As a result, the benefits offered by AI technologies may not be fully realised by society. These dangers arise largely from unintended consequences and relate typically to good intentions gone awry. However, we must also consider the risks associated with inadvertent *overuse* or wilful *misuse* of AI technologies, grounded, for example, in misaligned incentives, greed, adversarial geopolitics, or malicious intent. Everything from email scams to full-scale cyberwarfare may be accelerated or intensified by the malicious use of AI technologies (Taddeo 2018). And new evils may be made possible (King et al. 2018). The possibility of social progress represented by the aforementioned opportunities above must be weighed against the risk that malicious manipulation will be enabled or enhanced by AI. Yet a broad risk is that AI may be underused out of fear of overuse or misuse. We summarise these risks in Fig. 1 below, and offer a more detailed explanation in the text that follows.

2.1 Who We Can Become: Enabling Human Self-Realisation, Without Devaluing Human Abilities

AI may enable self-realisation, by which we mean the ability for people to flourish in terms of their own characteristics, interests, potential abilities or skills, aspirations, and life projects. Much as inventions, such as the washing machine, liberated people—particularly women—from the drudgery of domestic work, the “smart” automation of other mundane aspects of life may free up yet more time for cultural, intellectual and social pursuits, and more interesting and rewarding work. More AI may easily mean more human life spent more intelligently. The risk in this case is not the obsolescence of some old skills and the

emergence of new ones per se, but the pace at which this is happening and the unequal distributions of the costs and benefits that result. A very fast devaluation of old skills and hence a quick disruption of the job market and the nature of employment can be seen at the level of both the individual and society. At the level of the individual, jobs are often intimately linked to personal identity, self-esteem, and social role or standing, all factors that may be adversely affected by redundancy, even putting to one side the potential for severe economic harm. Furthermore, at the level of society, the deskilling in sensitive, skill-intensive domains, such as health care diagnosis or aviation, may create dangerous vulnerabilities in the event of AI malfunction or an adversarial attack. Fostering the development of AI in support of new abilities and skills, while anticipating and mitigating its impact on old ones will require both close study and potentially radical ideas, such as the proposal for some form of “universal basic income”, which is growing in popularity and experimental use. In the end, we need some intergenerational solidarity between those disadvantaged today and those advantaged tomorrow, to ensure that the disruptive transition between the present and the future will be as fair as possible, for everyone.

2.2 What We Can Do: Enhancing Human Agency, Without Removing Human Responsibility

AI is providing a growing reservoir of “smart agency”. Put at the service of human intelligence, such a resource can hugely enhance human agency. We can do more, better, and faster, thanks to the support provided by AI. In this sense of “Augmented Intelligence”, AI could be compared to the impact that engines have had on our lives. The larger the number of people who will enjoy the opportunities and benefits of such a reservoir of smart agency “on tap”, the better our societies will be. Responsibility is therefore essential, in view of what sort of AI we develop, how we use it, and whether we share with everyone its advantages and benefits. Obviously, the corresponding risk is the absence of such responsibility. This may happen not just because we have the wrong socio-political framework, but also because of a “black box” mentality, according to which AI systems for decision-making are seen as being beyond human understanding, and hence control. These concerns apply not only to high-profile cases, such as deaths caused by autonomous vehicles, but also to more commonplace but still significant uses, such as in automated decisions about parole or creditworthiness.

Yet the relationship between the degree and quality of agency that people enjoy and how much agency we delegate to autonomous systems is not zero-sum, either pragmatically or ethically. In fact, if developed thoughtfully, AI offers the opportunity of *improving and multiplying* the possibilities for human agency. Consider examples of “distributed morality” in human-to-human systems such as peer-to-peer lending (Floridi 2013). Human agency may be ultimately supported, refined and expanded by the embedding of “facilitating frameworks”, designed to improve the likelihood of morally good outcomes, in the set

of functions that we delegate to AI systems. AI systems could, if designed effectively, amplify and strengthen shared moral systems.

2.3 What We Can Achieve: Increasing Societal Capabilities, Without Reducing Human Control

Artificial intelligence offers myriad opportunities for improving and augmenting the capabilities of individuals and society at large. Whether by preventing and curing diseases or optimising transportation and logistics, the use of AI technologies presents countless possibilities for reinventing society by radically enhancing what humans are collectively capable of. More AI may support better coordination, and hence more ambitious goals. Human intelligence augmented by AI could find new solutions to old and new problems, from a fairer or more efficient distribution of resources to a more sustainable approach to consumption. Precisely because such technologies have the potential to be so powerful and disruptive, they also introduce proportionate risks. Increasingly, we may not need to be either ‘in or on the loop’ (that is, as part of the process or at least in control of it), if we can delegate our tasks to AI. However, if we rely on the use of AI technologies to augment our own abilities in the wrong way, we may delegate important tasks and above all decisions to autonomous systems that should remain at least partly subject to human supervision and choice. This in turn may reduce our ability to monitor the performance of these systems (by no longer being ‘on the loop’ either) or preventing or redressing errors or harms that arise (‘post loop’). It is also possible that these potential harms may accumulate and become entrenched, as more and more functions are delegated to artificial systems. It is therefore imperative to strike a balance between pursuing the ambitious opportunities offered by AI to improve human life and what we can achieve, on the one hand, and, on the other hand, ensuring that we remain in control of these major developments and their effects.

2.4 How We Can Interact: Cultivating Societal Cohesion, Without Eroding Human Self-Determination

From climate change and antimicrobial resistance to nuclear proliferation and fundamentalism, global problems increasingly have high degrees of coordination complexity, meaning that they can be tackled successfully only if all stakeholders co-design and co-own the solutions and cooperate to bring them about. AI, with its data-intensive, algorithmic-driven solutions, can hugely help to deal with such coordination complexity, supporting more societal cohesion and collaboration. For example, efforts to tackle climate change have exposed the challenge of creating a cohesive response, both within societies and between them. The scale of this challenge is such that we may soon need to decide between engineering the climate directly and designing societal frameworks to encourage a drastic cut in harmful emissions. This latter option might be undergirded by an algorithmic system to cultivate societal cohesion. Such a system would not be imposed from the outside; it

would be the result of a self-imposed choice, not unlike our choice of not buying chocolate if we had earlier chosen to be on a diet, or setting up an alarm clock to wake up. “Self-nudging” to behave in socially preferable ways is the best form of nudging, and the only one that preserves autonomy. It is the outcome of human decisions and choices, but it can rely on AI solutions to be implemented and facilitated. Yet the risk is that AI systems may erode human self-determination, as they may lead to unplanned and unwelcome changes in human behaviours to accommodate the routines that make automation work and people’s lives easier. AI’s predictive power and relentless nudging, even if unintentional, should be at the service of human self-determination and foster societal cohesion, not undermining human dignity or human flourishing.

Taken together, these four opportunities, and their corresponding challenges, paint a mixed picture about the impact of AI on society and the people in it. Accepting the presence of trade-offs, seizing the opportunities while working to anticipate, avoid, or minimise the risks head-on will improve the prospect for AI technologies to promote human dignity and flourishing. Having outlined the potential benefits to individuals and society at large of an ethically engaged approach to AI, in the next section we highlight the “dual advantage” to organisations of taking such an approach.

3 The Dual Advantage of an Ethical Approach to AI

Ensuring socially preferable outcomes of AI relies on resolving the tension between incorporating the benefits and mitigating the potential harms of AI, in short, simultaneously avoiding the misuse and underuse of these technologies. In this context, the value of an ethical approach to AI technologies comes into starker relief. Compliance with the law is merely necessary (it is the least that is required), but significantly insufficient (it is not the most that can and should be done) (Floridi 2018). With an analogy, it is the difference between playing according to the rules, and playing well, so that one may win the game. Adopting an ethical approach to AI confers what we define here as a “dual advantage”. On one side, ethics enables organisations to take advantage of the social value that AI enables. This is the advantage of being able to identify and leverage new opportunities that are socially acceptable or preferable. On the other side, ethics enables organisations to anticipate and avoid or at least minimise costly mistakes. This is the advantage of prevention and mitigation of courses of action that turn out to be socially unacceptable and hence rejected, even when legally unquestionable. This also lowers the opportunity costs of choices not made or options not grabbed for fear of mistakes.

Ethics’ dual advantage can only function in an environment of public trust and clear responsibilities more broadly. Public acceptance and adoption of AI technologies will occur only if the benefits are seen as meaningful and risks as potential, yet preventable, minimisable, or at least something against which one can be protected, through risk management (e.g. insurance) or redressing. These attitudes will depend in turn on public engagement with the development of AI technologies, openness about how they operate, and understandable, widely accessible mechanisms of

regulation and redress. In this way, an ethical approach to AI can also be seen as an early warning system against risks that might endanger entire organisations. The clear value to any organisation of the dual advantage of an ethical approach to AI amply justifies the expense of engagement, openness, and contestability that such an approach requires.

4 A Unified Framework of Principles for AI in Society

AI4People is not the first initiative to consider the ethical implications of AI. Many organisations have already produced statements of the values or principles that should guide the development and deployment of AI in society. Rather than conduct a similar, potentially redundant exercise here, we strive to move the dialogue forward, constructively, from principles to proposed policies, best practices, and concrete recommendations for new strategies. Such recommendations are not offered in a vacuum. But rather than generating yet another series of principles to serve as an ethical foundation for our recommendations, we offer a synthesis of existing sets of principles produced by various reputable, multi-stakeholder organisations and initiatives. A fuller explanation of the scope, selection and method of assessing these sets of principles is available in Cowls and Floridi (Forthcoming). Here, we focus on the commonalities and noteworthy differences observable across these sets of principles, in view of the 20 recommendations offered in the rest of the paper. The documents we assessed are:

1. The Asilomar AI Principles, developed under the auspices of the Future of Life Institute, in collaboration with attendees of the high-level Asilomar conference of January 2017 (hereafter “Asilomar”; Asilomar AI Principles 2017);
2. The Montreal Declaration for Responsible AI, developed under the auspices of the University of Montreal, following the Forum on the Socially Responsible Development of AI of November 2017 (hereafter “Montreal”; Montreal Declaration 2017)³;
3. The General Principles offered in the second version of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. This crowd-sourced global treatise received contributions from 250 global thought leaders to develop principles and recommendations for the ethical development and design of autonomous and intelligent systems, and was published in December 2017 (hereafter “IEEE”; IEEE 2017)⁴;
4. The Ethical Principles offered in the *Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems*, published by the European Commission’s European

³ The Montreal Declaration is currently open for comments as part of a redrafting exercise. The principles we refer to here are those which were publicly announced as of 1st May, 2018.

⁴ The third version of *Ethically Aligned Design* will be released in 2019 following wider public consultation.

- Group on Ethics in Science and New Technologies, in March 2018 (hereafter “EGE”; EGE 2018);
5. The “five overarching principles for an AI code” offered in paragraph 417 of the UK House of Lords Artificial Intelligence Committee’s report, *AI in the UK: ready, willing and able?*, published in April 2018 (hereafter “AIUK”; House of Lords 2018); and
 6. The Tenets of the Partnership on AI, a multistakeholder organisation consisting of academics, researchers, civil society organisations, companies building and utilising AI technology, and other groups (hereafter “the Partnership”; Partnership on AI 2018).

Taken together, they yield 47 principles.⁵ Overall, we find an impressive and reassuring degree of coherence and overlap between the six sets of principles. This can most clearly be shown by comparing the sets of principles with the set of four core principles commonly used in bioethics: beneficence, non-maleficence, autonomy, and justice. The comparison should not be surprising. Of all areas of applied ethics, bioethics is the one that most closely resembles digital ethics in dealing ecologically with new forms of agents, patients, and environments (Floridi 2013). The four bioethical principles adapt surprisingly well to the fresh ethical challenges posed by artificial intelligence. But they are not exhaustive. On the basis of the following comparative analysis, we argue that one more, new principle is needed in addition: *explicability*, understood as incorporating both intelligibility and accountability.

4.1 Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet

Of the four core bioethics principles, beneficence is perhaps the easiest to observe across the six sets of principles we synthesise here. The principle of creating AI technology that is beneficial to humanity is expressed in different ways, but it typically features at the top of each list of principles. Montreal and IEEE principles both use the term “well-being”: for Montreal, “the development of AI should ultimately promote the well-being of all sentient creatures”; while IEEE states the need to “prioritize human well-being as an outcome in all system designs”. AIUK and Asilomar both characterise this principle as the “common good”: AI should “be developed for the common good and the benefit of humanity”, according to AIUK. The Partnership describes the intention to “ensure that AI technologies benefit and empower as many people as possible”; while the EGE emphasises the principle of both “human dignity” and “sustainability”. Its principle of “sustainability” represents perhaps the

⁵ Of the six documents, the Asilomar Principles offer the largest number of principles with arguably the broadest scope. The 23 principles are organised under three headings, “research issues”, “ethics and values”, and “longer-term issues”. We have omitted consideration of the five “research issues” here as they are related specifically to the practicalities of AI development, particularly in the narrower context of academia and industry. Similarly, the Partnership’s eight Tenets consist of both intra-organisational objectives and wider principles for the development and use of AI. We include only the wider principles (the first, sixth, and seventh tenets).

widest of all interpretations of beneficence, arguing that “AI technology must be in line with ... ensur[ing] the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations”. Taken together, the prominence of these principles of beneficence firmly underlines the central importance of promoting the well-being of people and the planet.

4.2 Non-maleficence: Privacy, Security and “Capability Caution”

Though “do only good” (beneficence) and “do no harm” (non-maleficence) seem logically equivalent, in both the context of bioethics and of the ethics of AI they represent distinct principles, each requiring explication. While they encourage well-being, the sharing of benefits and the advancement of the public good, each of the six sets of principles also cautions against the many potentially negative consequences of overusing or misusing AI technologies. Of particular concern is the prevention of infringements on personal privacy, which is listed as a principle in five of the six sets, and as part of the “human rights” principles in the IEEE document. In each case, privacy is characterised as being intimately linked to individuals’ access to, and control over, how personal data is used.

Yet the infringement of privacy is not the only danger to be avoided in the adoption of AI. Several of the documents also emphasise the importance of avoiding the misuse of AI technologies in other ways. The Asilomar Principles are quite specific on this point, citing the threats of an AI arms race and of the recursive self-improvement of AI, as well as the need for “caution” around “upper limits on future AI capabilities”. The Partnership similarly asserts the importance of AI operating “within secure constraints”. The IEEE document meanwhile cites the need to “avoid misuse”, while the Montreal Declaration argues that those developing AI “should assume their responsibility by working against the risks arising from their technological innovations”, echoed by the EGE’s similar need for responsibility.

From these various warnings, it is not entirely clear whether it is the people developing AI, or the technology itself, which should be encouraged not to do harm—in other words, whether it is Frankenstein or his monster against whose maleficence we should be guarding. Confused also is the question of intent: promoting non-maleficence can be seen to incorporate the prevention of both accidental (what we above call “overuse”) and deliberate (what we call “misuse”) harms arising. In terms of the principle of non-maleficence, this need not be an either/or question: the point is simply to prevent harms arising, whether from the intent of humans or the unpredicted behaviour of machines (including the unintentional nudging of human behaviour in undesirable ways). Yet these underlying questions of agency, intent and control become knottier when we consider the next principle.

4.3 Autonomy: The Power to Decide (Whether to Decide)

Another classic tenet of bioethics is the principle of autonomy: the idea that individuals have a right to make decisions for themselves about the treatment they do or

not receive. In a medical context, this principle of autonomy is most often impaired when patients lack the mental capacity to make decisions in their own best interests; autonomy is thus surrendered involuntarily. With AI, the situation becomes rather more complex: when we adopt AI and its smart agency, we *willingly* cede some of our decision-making power to machines. Thus, affirming the principle of autonomy in the context of AI means striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents.

The principle of autonomy is explicitly stated in four of the six documents. The Montreal Declaration articulates the need for a balance between human- and machine-led decision-making, stating that “the development of AI should *promote* the autonomy of all human beings *and control* ... the autonomy of computer systems” (italics added). The EGE argues that autonomous systems “must not impair [the] freedom of human beings to set their own standards and norms and be able to live according to them”, while AIUK adopts the narrower stance that “the autonomous power to hurt, destroy or deceive human beings should never be vested in AI”. The Asilomar document similarly supports the principle of autonomy, insofar as “humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives”.

These documents express a similar sentiment in slightly different ways, echoing the distinction drawn above between beneficence and non-maleficence: not only should the autonomy of humans be promoted, but also the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be re-established (consider the case of a pilot able to turn off the automatic pilot and regain full control of the airplane). Taken together, the central point is to protect the intrinsic value of human choice—at least for significant decisions—and, as a corollary, to contain the risk of delegating too much to machines. Therefore, what seems most important here is what we might call “meta-autonomy”, or a “decide-to-delegate” model: humans should always retain the power to *decide which decisions to take*, exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making. As anticipated, any delegation should remain overridable in principle (deciding to decide again).

The decision to make or delegate decisions does not take place in a vacuum. Nor is this capacity to decide (to decide, and to decide again) distributed equally across society. The consequences of this potential disparity in autonomy are addressed in the final of the four principles inspired by bioethics.

4.4 Justice: Promoting Prosperity and Preserving Solidarity

The last of the four classic bioethics principles is justice, which is typically invoked in relation to the distribution of resources, such as new and experimental treatment options or simply the general availability of conventional healthcare. Again, this bioethics principle finds clear echoes across the principles for AI that we analyse. The importance of “justice” is explicitly cited in the Montreal Declaration, which argues that “the development of AI should promote justice and seek to eliminate all

types of discrimination”, while the Asilomar Principles include the need for both “shared benefit” and “shared prosperity” from AI. Under its principle named “Justice, equity and solidarity”, the EGE argues that AI should “contribute to global justice and equal access to the benefits” of AI technologies. It also warns against the risk of bias in datasets used to train AI systems, and—unique among the documents—argues for the need to defend against threats to “solidarity”, including “systems of mutual assistance such as in social insurance and healthcare”. The emphasis on the protection of social support systems may reflect geopolitics, insofar as the EGE is a European body. The AIUK report argues that citizens should be able to “flourish mentally, emotionally and economically alongside artificial intelligence”. The Partnership, meanwhile, adopts a more cautious framing, pledging to “respect the interests of all parties that may be impacted by AI advances”.

As with the other principles already discussed, these interpretations of what justice means as an ethical principle in the context of AI are broadly similar, yet contain subtle distinctions. Across the documents, justice variously relates to

- (a) Using AI to correct past wrongs such as eliminating unfair discrimination;
- (b) Ensuring that the use of AI creates benefits that are shared (or at least shareable); and
- (c) Preventing the creation of *new* harms, such as the undermining of existing social structures.

Notable also are the different ways in which the position of AI, *vis-à-vis* people, is characterised in relation to justice. In Asilomar and EGE respectively, it is AI technologies themselves that “should benefit and empower as many people as possible” and “contribute to global justice”, whereas in Montreal, it is “the *development* of AI” that “should promote justice” (*italics added*). In AIUK, meanwhile, people should flourish merely “alongside” AI. Our purpose here is not to split semantic hairs. The diverse ways in which the relationship between people and AI is described in these documents hints at broader confusion over AI as a man-made reservoir of “smart agency”. Put simply, and to resume our bioethics analogy, are we (humans) the patient, receiving the “treatment” of AI, the doctor prescribing it? Or both? It seems that we must resolve this question before seeking to answer the next question of whether the treatment will even work. This is the core justification for our identification within these documents of a new principle, one that is not drawn from bioethics.

4.5 Explicability: Enabling the Other Principles Through Intelligibility and Accountability

The short answer to the question of whether “we” are the patient or the doctor is that actually we could be either—depending on the circumstances and on who “we” are in our everyday life. The situation is inherently unequal: a small fraction of humanity is currently engaged in the design and development of a set of technologies that are already transforming the everyday lives of just about everyone else. This stark

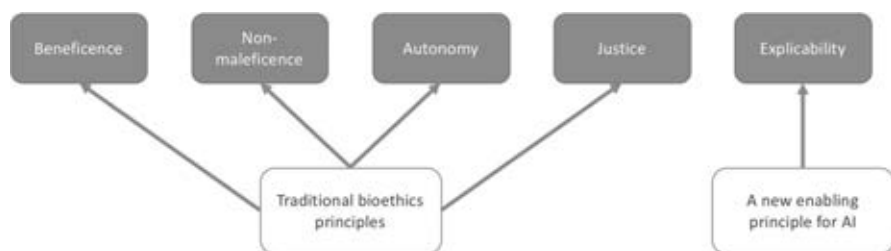


Fig. 2 An ethical framework for AI, formed of four traditional principles and a new one

reality is not lost on the authors whose documents we analyse. In all, reference is made to the need to *understand* and *hold to account* the decision-making processes of AI. This principle is expressed using different terms: “transparency” in Asilomar; “accountability” in EGE; both “transparency” and “accountability” in IEEE; “intelligibility” in AIUK; and as “understandable and interpretable” for the Partnership. Though described in different ways, each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible to all but (at best) the most expert observers.

The addition of this principle, which we synthesise as “explicability” both in the epistemological sense of “intelligibility” (as an answer to the question “how does it work?”) and in the ethical sense of “accountability” (as an answer to the question: “who is responsible for the way it works?”), is therefore the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI. It complements the other four principles: for AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our “decision about who should decide” must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must ensure that the technology—or, more accurately, the people and organisations developing and deploying it—are held accountable in the event of a negative outcome, which would require in turn some understanding of why this outcome arose. More broadly, we must negotiate the terms of the relationship between ourselves and this transformative technology, on grounds that are readily understandable to the proverbial person “on the street”.

Taken together, we argue that these five principles capture the meaning of each of the 47 principles contained in the six high-profile, expert-driven documents, forming an ethical framework within which we offer our recommendations below. This framework of principles is shown in Fig. 2.

5 Recommendations for a Good AI Society

This section introduces the Recommendations for a Good AI Society. It consists of two parts: a Preamble, and 20 Action Points.

There are four kinds of Action Points: to *assess*, to *develop*, to *incentivise* and to *support*. Some recommendations may be undertaken directly, by national or European policy makers, in collaboration with stakeholders where appropriate. For others, policy makers may play an enabling role for efforts undertaken or led by third parties.

5.1 Preamble

We believe that, in order to create a Good AI Society, the ethical principles identified in the previous section should be embedded in the default practices of AI. In particular, AI should be designed and developed in ways that decrease inequality and further social empowerment, with respect for human autonomy, and increase benefits that are shared by all, equitably. It is especially important that AI be explicable, as explicability is a critical tool to build public trust in, and understanding of, the technology.

We also believe that creating a Good AI Society requires a multistakeholder approach, which is the most effective way to ensure that AI will serve the needs of society, by enabling developers, users and rule-makers to be on board and collaborating from the outset.

Different cultural frameworks inform attitudes to new technology. This document represents a European approach, which is meant to be complementary to other approaches. We are committed to the development of AI technology in a way that *secures people's trust, serves the public interest, and strengthens shared social responsibility*.

Finally, this set of recommendations should be seen as a “living document”. The Action Points are designed to be dynamic, requiring not simply single policies or one-off investments, but rather, continuous, ongoing efforts for their effects to be sustained.

5.2 Action Points

5.2.1 Assessment

1. *Assess* the capacity of existing institutions, such as national civil courts, to redress the mistakes made or harms inflicted by AI systems. This assessment should evaluate the presence of sustainable, majority-agreed foundations for liability from the design stage onwards, in order to reduce negligence and conflicts (see also Recommendation 5).⁶
2. *Assess* which tasks and decision-making functionalities should *not* be delegated to AI systems, through the use of participatory mechanisms to ensure alignment with

⁶ Determining accountability and responsibility may usefully borrow from lawyers in Ancient Rome who would go by this formula ‘cuius commoda eius et incommoda’ (‘the person who derives an advantage from a situation must also bear the inconvenience’). A good 2200 years old principle that has a well-established tradition and elaboration could properly set the starting level of abstraction in this field.

societal values and understanding of public opinion. This assessment should take into account existing legislation and be supported by ongoing dialogue between all stakeholders (including government, industry, and civil society) to debate how AI will impact society opinion (in concert with Recommendation 17).

3. *Assess* whether current regulations are sufficiently grounded in ethics to provide a legislative framework that can keep pace with technological developments. This may include a framework of key principles that would be applicable to urgent and/or unanticipated problems.

5.2.2 Development

4. *Develop* a framework to enhance the explicability of AI systems that make socially significant decisions. Central to this framework is the ability for individuals to obtain a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences. This is likely to require the development of frameworks specific to different industries, and professional associations should be involved in this process, alongside experts in science, business, law, and ethics.
5. *Develop* appropriate legal procedures and improve the IT infrastructure of the justice system to permit the scrutiny of algorithmic decisions in court. This is likely to include the creation of a framework for AI explainability as indicated in Recommendation 4, specific to the legal system. Examples of appropriate procedures may include the applicable disclosure of sensitive commercial information in IP litigation, and—where disclosure poses unacceptable risks, for instance to national security—the configuration of AI systems to adopt technical solutions by default, such as zero-knowledge proofs in order to evaluate their trustworthiness.
6. *Develop* auditing mechanisms for AI systems to identify unwanted consequences, such as unfair bias, and (for instance, in cooperation with the insurance sector) a solidarity mechanism to deal with severe risks in AI-intensive sectors. Those risks could be mitigated by multistakeholder mechanisms upstream. Pre-digital experience indicates that, in some cases, it may take a couple of decades before society catches up with technology by way of rebalancing rights and protection adequately to restore trust. The earlier that users and governments become involved—as made possible by ICT—the shorter this lag will be.
7. *Develop* a redress process or mechanism to remedy or compensate for a wrong or grievance caused by AI. To foster public trust in AI, society needs a widely accessible and reliable mechanism of redress for harms inflicted, costs incurred, or other grievances caused by the technology. Such a mechanism will necessarily involve a clear and comprehensive allocation of accountability to humans and/or organisations. Lessons could be learnt from the aerospace industry, for example, which has a proven system of handling unwanted consequences thoroughly and seriously. The development of this process must follow from the assessment of existing capacity outlined in Recommendation 1. If a lack of capacity is identified, additional institutional solutions should be developed at national and/or EU levels, to enable people to seek redress. Such solutions may include:

- An “AI ombudsperson” to ensure the auditing of allegedly unfair or inequitable uses of AI;
- A guided process for registering a complaint akin to making a Freedom of Information request; and
- The development of liability insurance mechanisms, which would be required as an obligatory accompaniment of specific classes of AI offerings in EU and other markets. This would ensure that the relative reliability of AI-powered artefacts, especially in robotics, is mirrored in insurance pricing and therefore in the market prices of competing products.⁷

Whichever solutions are developed, these are likely to rely on the framework for intelligibility proposed in Recommendation 4.

8. *Develop* agreed-upon metrics for the trustworthiness of AI products and services, to be undertaken either by a new organisation, or by a suitable existing organisation. These metrics would serve as the basis for a system that enables the user-driven benchmarking of all marketed AI offerings. In this way, an index for trustworthy AI can be developed and signalled, in addition to a product’s price. This “trust comparison index” for AI would improve public understanding and engender competitiveness around the development of safer, more socially beneficial AI (e.g., “IwantgreatAI.org”). In the longer term, such a system could form the basis for a broader system of certification for deserving products and services, administered by the organisation noted here, and/or by the oversight agency proposed in Recommendation 9. The organisation could also support the development of codes of conduct (see Recommendation 18). Furthermore, those who own or operate inputs to AI systems and profit from it could be tasked with funding and/or helping to develop AI literacy programs for consumers, in their own best interest.
9. *Develop* a new EU oversight agency responsible for the protection of public welfare through the scientific evaluation and supervision of AI products, software, systems, or services. This may be similar, for example, to the European Medicines Agency. Relatedly, a “post-release” monitoring system for AIs similar to, for example, the one available for drugs should be developed, with reporting duties for some stakeholders and easy reporting mechanisms for other users.
10. *Develop* a European observatory for AI. The mission of the observatory would be to watch developments, provide a forum to nurture debate and consensus, provide a repository for AI literature and software (including concepts and links to available literature), and issue step-by-step recommendation and guidelines for action.
11. *Develop* legal instruments and contractual templates to lay the foundation for a smooth and rewarding human–machine collaboration in the work environment.

⁷ Of course, to the extent that AI systems are ‘products’, general tort law still applies in the same way to AI as it applies in any instance involving defective products or services that injure users or do not perform as claimed or expected.

Shaping the narrative on the ‘Future of Work’ is instrumental to winning “hearts and minds”. In keeping with ‘A Europe that protects’, the idea of “inclusive innovation” and to smooth the transition to new kinds of jobs, a European AI Adjustment Fund could be set up along the lines of the European Globalisation Adjustment Fund.

5.2.3 Incentivisation

12. *Incentivise* financially, at the EU level, the development and use of AI technologies within the EU that are socially preferable (not merely acceptable) and environmentally friendly (not merely sustainable but favourable to the environment). This will include the elaboration of methodologies that can help assess whether AI projects are socially preferable and environmentally friendly. In this vein, adopting a ‘challenge approach’ (see DARPA challenges) may encourage creativity and promote competition in the development of specific AI solutions that are ethically sound and in the interest of the common good.
13. *Incentivise* financially a sustained, increased and coherent European research effort, tailored to the specific features of AI as a scientific field of investigation. This should involve a clear mission to advance AI for social good, to serve as a unique counterbalance to AI trends with less focus on social opportunities.
14. *Incentivise* financially cross-disciplinary and cross-sectoral cooperation and debate concerning the intersections between technology, social issues, legal studies, and ethics. Debates about technological challenges may lag behind the actual technical progress, but if they are strategically informed by a diverse, multistakeholder group, they may steer and support technological innovation in the right direction. Ethics should help seize opportunities and cope with challenges, not only describe them. It is essential in this respect that diversity infuses the design and development of AI, in terms of gender, class, ethnicity, discipline and other pertinent dimensions, in order to increase inclusivity, toleration, and the richness of ideas and perspectives.
15. *Incentivise* financially the inclusion of ethical, legal and social considerations in AI research projects. In parallel, incentivise regular reviews of legislation to test the extent to which it fosters socially positive innovation. Taken together, these two measures will help ensure that AI technology has ethics at its heart and that policy is oriented towards innovation.
16. *Incentivise* financially the development and use of lawfully de-regulated special zones within the EU for the empirical testing and development of AI systems. These zones may take the form of a “living lab” (or *Tokku*), building on the experience of existing “test highways” (or *Teststrecken*). In addition to aligning innovation more closely with society’s preferred level of risk, sandbox experiments such as these contribute to hands-on education and the promotion of accountability and acceptability at an early stage. “Protection by design” is intrinsic to this kind of framework.
17. *Incentivise* financially research about public perception and understanding of AI and its applications, and the implementation of structured public consultation mechanisms to design policies and rules related to AI. This may include

the direct elicitation of public opinion via traditional research methods, such as opinion polls and focus groups, as well as more experimental approaches, such as providing simulated examples of the ethical dilemmas introduced by AI systems, or experiments in social science labs. This research agenda should not serve merely to measure public opinion, but should also lead to the co-creation of policies, standards, best practices, and rules as a result.

5.2.4 Support

18. *Support* the development of self-regulatory codes of conduct for data and AI related professions, with specific ethical duties. This would be along the lines of other socially sensitive professions, such as medical doctors or lawyers, i.e., with the attendant certification of ‘ethical AI’ through trust-labels to make sure that people understand the merits of ethical AI and will therefore demand it from providers. Current attention manipulation techniques may be constrained through these self-regulating instruments.
19. *Support* the capacity of corporate boards of directors to take responsibility for the ethical implications of companies’ AI technologies. For example, this may include improved training for existing boards and the potential development of an ethics committee with internal auditing powers. This could be developed within the existing structure of both one-tier and two-tier board systems, and/or in conjunction with the development of a mandatory form of “corporate ethical review board” to be adopted by organisations developing or using AI systems, to evaluate initial projects and their deployment with respect to fundamental principles.
20. *Support* the creation of educational curricula and public awareness activities around the societal, legal, and ethical impact of Artificial Intelligence. This may include:
 - Curricula for schools, supporting the inclusion of computer science among the basic disciplines to be taught;
 - Initiatives and qualification programmes in businesses dealing with AI technology, to educate employees on the societal, legal, and ethical impact of working alongside AI;
 - A European-level recommendation to include ethics and human rights in the degrees of data and AI scientists and other scientific and engineering curricula dealing with computational and AI systems;
 - The development of similar programmes for the public at large, with a special focus on those involved at each stage of management of the technology, including civil servants, politicians and journalists;
 - Engagement with wider initiatives such as the ITU AI for Good events and NGOs working on the UN Sustainable Development Goals.

6 Conclusion

Europe, and the world at large, face the emergence of a technology that holds much exciting promise for many aspects of human life, and yet seems to pose major threats as well. This article—and especially the Recommendations in the previous section—seek to nudge the tiller in the direction of ethically and socially preferable outcomes from the development, design and deployment of AI technologies. Building on our identification of both the core opportunities and the risks of AI for society as well as the set of five ethical principles we synthesised to guide its adoption, we formulated 20 Action Points in the spirit of collaboration and in the interest of creating *concrete* and *constructive* responses to the most pressing social challenges posed by AI.

With the rapid pace of technological change, it can be tempting to view the political process in the liberal democracies of today as old-fashioned, out-of-step, and no longer up to the task of preserving the values and promoting the interests of society and everyone in it. We disagree. With the Recommendations we offer here, including the creation of centres, agencies, curricula, and other infrastructure, we have made the case for an ambitious, inclusive, equitable programme of policy making and technological innovation, which we believe will contribute to securing the benefits and mitigating the risks of AI, for all people, and for the world we share.

Acknowledgements This publication would not have been possible without the generous support of Atomium—European Institute for Science, Media and Democracy. We are particularly grateful to Michelangelo Baracchi Bonvicini, Atomium’s President, to Guido Romeo, its Editor in Chief, the staff of Atomium for their help, and to all the partners of the AI4People project and members of its Forum (<http://www.eismd.eu/ai4people>) for their feedback. Luciano Floridi’s work has also been supported by the Privacy-Enhancing and Identification-Enabling Solutions for IoT (PEIESI) project, part of the PETRAS Internet of Things research hub, funded by the Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1. The authors of this article are the only persons responsible for its contents and any remaining mistakes.


Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Asilomar AI Principles. (2017). *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*. Retrieved September 18, 2018 from <https://futureoflife.org/ai-principles>.
- Cowls, J., & Floridi, L. (2018). *Prolegomena to a White Paper on Recommendations for the Ethics of AI (June 19, 2018)*. Available at SSRN: <https://ssrn.com/abstract=3198732>.
- Cowls, J., & Floridi, L. (Forthcoming). *The Utility of a Principled Approach to AI Ethics*.
- European Group on Ethics in Science and New Technologies. (2018). *Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems*. Retrieved September 18, 2018 from https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-egre-released-2018-apr-24_en.
- Floridi, L. (2013). *The ethics of information*. Oxford: Oxford University Press.
- Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology*, 31(1), 1–8.
- House of Lords Artificial Intelligence Committee. (2018). *AI in the UK: ready, willing and able?* Retrieved September 18, 2018 from <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>.

- Imperial College London. (2017). *Written Submission to House of Lords Select Committee on Artificial Intelligence [AIC0214]*. Retrieved September 18, 2018 from <http://bit.ly/2yleuET>.
- King, T., Aggarwal, N., Taddeo, M., & Floridi, L. (2018). *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*. Available at SSRN: <https://ssrn.com/abstract=3183238>.
- Montreal Declaration for a Responsible Development of Artificial Intelligence. (2017). *Announced at the conclusion of the Forum on the Socially Responsible Development of AI*. Retrieved September 18, 2018 from <https://www.montrealdeclaration-responsibleai.com/the-declaration>.
- Partnership on AI. (2018). *Tenets*. Retrieved September 18, 2018 from <https://www.partnershiponai.org/tenets/>.
- Taddeo, M. (2018). The limits of deterrence theory in cyberspace. *Philosophy & Technology*, 31(3), 339–355.
- The IEEE Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically Aligned Design*, v2. Retrieved September 18, 2018 from <https://ethicsinaction.ieee.org>.

Affiliations

Luciano Floridi^{1,2}  · Josh Cows1^{1,2} · Monica Beltrametti³ · Raja Chatila^{4,5} · Patrice Chazerand⁶ · Virginia Dignum^{7,8} · Christoph Luetge⁹ · Robert Madelin¹⁰ · Ugo Pagallo¹¹ · Francesca Rossi^{12,13} · Burkhard Schafer¹⁴ · Peggy Valcke^{15,16} · Effy Vayena¹⁷

¹ Oxford Internet Institute, University of Oxford, Oxford, UK

² The Alan Turing Institute, London, UK

³ Naver Corporation, Grenoble, France

⁴ French National Center of Scientific Research, Paris, France

⁵ Institute of Intelligent Systems and Robotics, Pierre and Marie Curie University, Paris, France

⁶ Digital Europe, Brussels, Belgium

⁷ University of Umeå, Umeå, Sweden

⁸ Delft Design for Values Institute, Delft University of Technology, Delft, The Netherlands

⁹ TUM School of Governance, Technical University of Munich, Munich, Germany

¹⁰ Centre for Technology and Global Affairs, University of Oxford, Oxford, UK

¹¹ Department of Law, University of Turin, Turin, Italy

¹² IBM Research, New York, USA

¹³ University of Padova, Padua, Italy

¹⁴ University of Edinburgh Law School, Edinburgh, UK

¹⁵ Centre for IT & IP Law, Catholic University of Leuven, Flanders, Belgium

¹⁶ Bocconi University, Milan, Italy

¹⁷ Bioethics, Health Ethics and Policy Lab, ETH Zurich, Zurich, Switzerland

Understanding Media and Information Quality in an Age of Artificial Intelligence, Automation, Algorithms and Machine Learning

Yochai Benkler, Robert Faris, Hal Roberts, Nikki Bourassa

Berkman Klein Center for Internet & Society

July 12, 2018



Introduction

Democracy is under stress that would have been unimaginable a decade or two ago. The victory of Donald Trump in the U.S. presidential campaign and the successful “Leave” campaign in the UK Brexit referendum led the Oxford English Dictionary to select “post-truth” as word of the year for 2016. Since then, anti-elite populism across the North Atlantic has continued to gain victories, chief among them the 2018 victory of the Five Star Movement in Italy. As governments, foundations, academics, political activists, journalists, and civil society in democratic societies struggled to understand the overwhelming political results alongside the apparent information disorder, most observers’ eyes fell on technology. Social media, bots, hyper-targeted behavioral marketing — all have been seen as culprits or at least catalysts in public conversations. Technology has altered the foundations of news and media, and as trust in media continues to decline, artificial intelligence, machine learning, and algorithms have come to play a critical role not only as threats to the integrity and quality of media, but also as a source of potential solutions.

The core threats to information quality associated with AI include:

- **Algorithmic curation.** Most commonly known as the “filter bubble” concern, algorithms designed by platforms to keep users engaged produce ever-more refined rabbit holes down which users can go in a dynamic of reinforcement learning that leads them to ever-more extreme versions of their beliefs and opinions.
- **Bots.** Improvements in automation allow bots to become ever-more-effective simulations of human participants, thereby permitting propagandists to mount large-scale influence campaigns on social media by simulating larger and harder-to-detect armies of automated accounts.
- **Fake reports and videos.** Improving automated news reporting and manipulation of video and audio may enable the creation of seemingly authentic videos of political actors that will irrevocably harm their reputations and become high-powered vectors for false reporting.
- **Targeted behavioral marketing powered by algorithms and machine-learning.** Here, the concern is that the vast amounts of individually-identifiable data about users will allow ever-improving algorithms to refine the stream of content that individuals receive so as to manipulate their political opinions and behaviors.

Our approach to this set of issues is through the analysis of very large datasets, covering millions of stories, tweets, and Facebook shares, across diverse platforms reflecting national presidential politics in America from April 2015 to the present. The results of our work suggest that emphasizing the particular and technological, rather than the interaction of technology, institutions, and culture in a particular historical context, may lead to systematic overstating of the importance of technology, both positive and negative. Many studies do not attempt to address the impact of AI and other technologically-mediated phenomena that they identify and measure, but rather are focused on identifying and describing observable practices. In our own work, we have repeatedly found bots, Russian influence campaigns, and “fake news” of the nihilistic political clickbait variety; in all these contexts, we have found these technologically-mediated phenomena to be background noise, rather than a major driver of the observed patterns of political communication. Excessive attention to these technology-centric

phenomena risks masking the real political-cultural dynamics that have confounded American political communication.

Studying media ecosystems

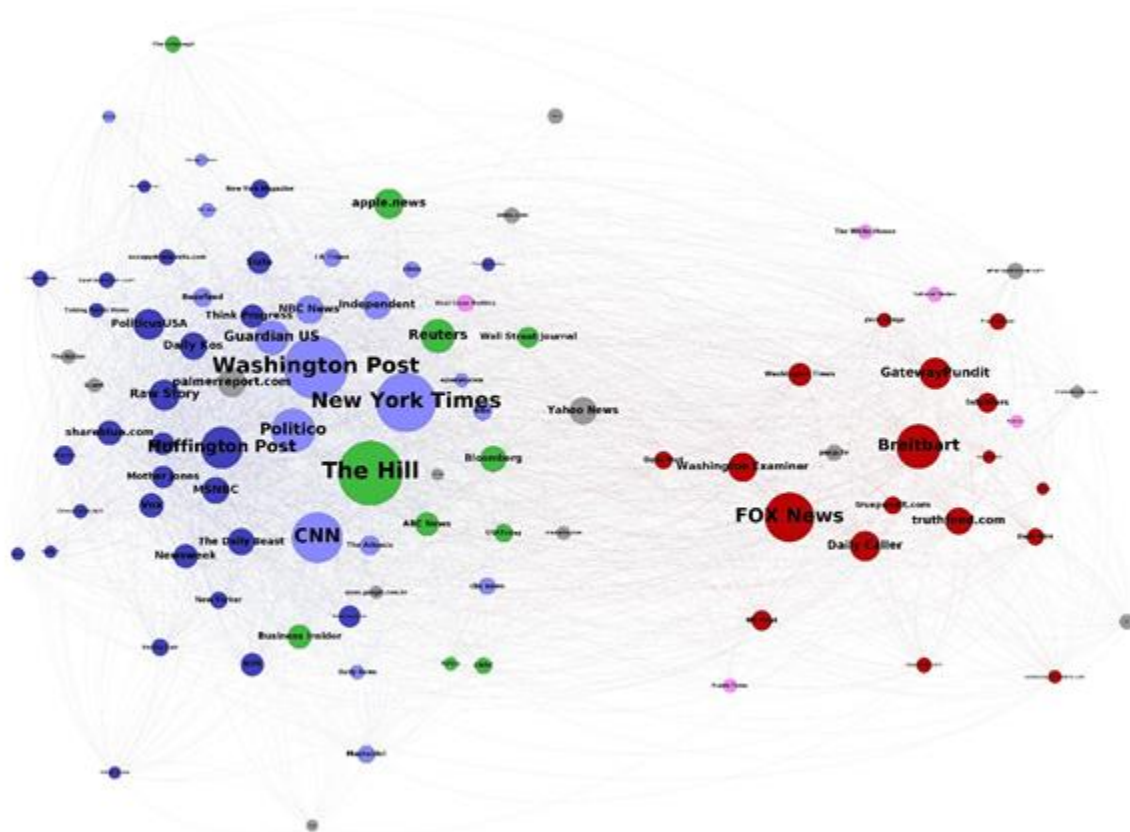
We study propaganda and disinformation in political communication through the lens of media ecosystems. The springboard for our research is a tool built in a decade-long collaboration between our Berkman Klein Center team and our colleagues at MIT Media Lab's Center for Civic Media: [Media Cloud](#). Media Cloud enables us to discover media network structures that form around specific topics and to shed light on media frames, attention, and influence. Our early work centered on discrete legislative and regulatory controversies, such as [net neutrality](#) and the [SOPA-PIPA bills](#). As election year approached in 2016 we turned our attention to the presidential race.

Initially, we analyzed the 18 months leading up to the November 8 vote. The [Columbia Journalism Review](#) published our preliminary findings in February 2017, and we followed with a more complete [analysis of pre-election media](#) in August of last year. Since then we have expanded our research to include analysis of the election's aftermath, and have combined the election-period study with a review of the first year of the Trump presidency into a book coming out with Oxford University Press in September 2018, entitled *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. With the support of the [Ethics and Governance of Artificial Intelligence Fund](#), we focused attention on the impact of artificial intelligence and automation on political public media. This work is motivated by the need to better understand the character, scope, and origins of the issue as an essential element in crafting corrective measures. Here we summarize our preliminary conclusions as well as the work of a growing number of researchers investing in this space.

Disinformation problems are fundamentally political and social — not technological

Our research leads us to conclude that political and social factors, not technology, are at the root of current media and information quality problems in American political communication. We analyze the linking, tweeting, Facebook sharing, and text of four million stories related to presidential politics published over a period of nearly three years before and after the 2016 presidential election. Relying on this data, we mapped out media source networks in open web media (online news, blogs, etc.) and on Twitter to learn how media sources clumped together based on similar cross-media linking and sharing patterns. This revealed a startling finding: media sources are, as many have observed, deeply polarized; but they are aligned in a distinctly asymmetric pattern. There is, as it turns out, not a simple left-right division in American political media. Instead, the divide is between the right and the rest of the political media spectrum. Media sources ranging from those with editorial voices that lean towards the center-right, such as the Wall Street Journal, all the way to distinctly left voices such as Mother Jones or the Daily Kos, comprise a single interconnected media ecosystem. In this portion of the media spectrum, attention is normally distributed around a peak focus on mainstream professional journalism outlets and

journalistic values and mainstream professional journalism continue to exercise significant constraint on the survivability and spread of falsehoods. By contrast, there is a large gap between media sources in this broad center to center-left to left cluster and media sources on the right. Moreover, attention on the right increases as one moves further from the center, and peaks around media sites that are exclusively right-wing.



Network map based on Twitter media sharing from January 22, 2017, to January 21, 2018. Nodes are sized by number of Twitter shares.

When we took a closer look at the substance and practices of journalism from the center of the spectrum to its poles, we found major differences in behavior. Media sources on the left and right extremes of the spectrum are prone to sacrificing accuracy in favor of partisan messaging, making them also prone to the spread of disinformation. However, the accountability and fact-checking mechanisms in the media structures on the center and center-left provide a barrier against further propagation and a check on shoddy reporting from the left. Media outlets that strive for objectivity and balance are less susceptible to spreading disinformation.

The leading media in the right-wing cluster follow a distinctly different strategy, and repeatedly amplify and accredit, rather than dampen or correct, bias-confirming falsehoods and misleading framings of

facts. We find that insular conservative media lack the will or capacity to adhere to objectivity and accuracy in their reporting. Infowars, Gateway Pundit, and conservative political clickbait are a problem, but a bigger problem is when Drudge Report, Breitbart, and Fox News — conservative media sources with much larger audiences — amplify conspiratorial and deceptive stories. This creates a significant gap in information quality for readers of conservative-leaning sources of news relative to those who consume media sources outside the right-wing ecosystem, including left-wing sites.

Algorithmic curation and filter bubbles

It is important to understand that if filter bubbles or algorithmic curation were the primary reason for polarization, we would expect to see symmetric patterns of attention clustering by populations of audiences and publisher — all of whom are at the same technological frontier. The stark asymmetry between the right and the rest strongly suggests that differences in political culture and, we argue in our book, institutionally-driven media history of right-wing media, particularly talk radio and Fox News, drove the right into a media dynamic we call the propaganda feedback loop. This occurred before major Internet sites developed; Internet sites adapted to that already-operative media dynamic, rather than causing it.

Indeed, when we compare the pre-election coverage and the post-election coverage, attention on the left has shifted more to the center, while attention on the right has remained as right-oriented. Fox News has become more prominent online relative to Breitbart, but at the price of becoming more purely focused on right-oriented content than it was during the election. Fox online has lost hyperlinks from the center-right and center even more than from the center-left and left, and gained links from the right. This pattern is not the simple product of algorithmic filter bubbles. Compared to influence metrics based on social media attention, weblinks are more insulated from platform-imposed algorithms as they reflect the judgment and linking decisions of web authors, rather than attention and sharing patterns of audiences. Nonetheless, the asymmetric polarization appears by this measure too, and it is by this measure that Fox itself has moved to the right since the 2016 election.

Bots

Over the course of the year we have developed our own versions of several of the most widely-used bot-detection algorithms. We reran all of our analyses after removing accounts that were identified as “bots” by various methods. Although we do in fact identify a good number of accounts that would be considered “bots” by many of these definitions, the overall architecture of communication we observe remains unchanged whether we include these accounts or not. Without stronger evidence that particular major communications dynamics have been swayed by bots, our macro-level observations leave us skeptical as to the importance of bots. Moreover, using micro-scale case studies, we observe bots as background noise rather than as impactful interventions.

Functional and structural differences

Today's media ecosystems are of course inseparable from technology, but if technology was at the core of the disinformation problem, we would expect to see as much disinformation on one side of the spectrum as the other. And yet, that's not the case. The asymmetry we observe in the structure and practices of media across the U.S. political spectrum is evidence that the dynamics and problems in modern day media cannot be explained by technology alone.

Our caution should not be taken as clear evidence that algorithms are unimportant. There is little doubt in our minds that algorithms and machine learning impact media manipulation and disinformation. In large-scale digital media platforms, algorithms serve a curatorial role and act as gatekeepers, and are subject to gaming and manipulation in ways that human editors are not. In our own research, we have seen, for example, that Twitter is littered with disinformation and false accounts, some of which are bots and others sockpuppets. Moreover, we observe that the spread of disinformation is worse on Facebook than on Twitter, and on Twitter than on the open web. Moreover, we observed that, as Facebook was making its algorithm changes over the year, certain of the worst offenders on both the left and the right lost prominence and attention, although new offenders quickly took their place. In our current work we continue to monitor the differences across platforms and to find measures by which to assess the effectiveness of algorithm changes on Facebook or Twitter on the spread of disinformation and misinformation, and, perhaps most importantly, to assess the impact of these changes on the media ecosystem as a whole.

Diagnoses, conclusions, and interventions should be based on systemic observations of impact, rather than individual observations of “bad behavior”

The media manipulation problem and the public attention afforded it have created an environment of urgency. Insights and research are key to mitigating some of this pressure, but the temptation to leap to assumptions of impact from any observable phenomenon is hard to resist. We are naturally drawn to the novel: trolls, clickbait factories, social media algorithms, targeted advertising, psychographic profiling, and Russian interference captivate the imagination and compel question and inquiry. Identifying instances of such interventions is technically challenging, and when we see hard work bear fruit it is tempting to attribute high importance to these observations. Nonetheless, without a baseline against which to assess the impact of one or several of these phenomena, it is impossible to justify assigning them the importance that they currently receive, or to design interventions that properly address the actually important interventions rather than those that are observable and salient.

Putting these novelties into perspective is hampered by our ability to fully and accurately measure media consumption. Audiences reached by misleading and false articles and media sources is measured in the millions. As prolific as this sounds, as described previously this ultimately is a small portion of




content and exposure. A second point of uncertainty relates to impact and persuasion — there is scant evidence that false media stories have any impact in changing behavior.

A prime example is Russian efforts to influence election-related discourse. Although information around the number of accounts, posts, and impressions seems to be updated regularly, current numbers draw into question the true significance of their reach. In October 2017, Facebook [revealed](#) that approximately 80,000 pieces of content published by the now infamous Russian-operated Internet Research Agency were introduced to 29 million people between January 2015 and August 2017. Their subsequent likes and shares increased the reach of the posts to 126 million Facebook users. This is indeed a big scary number. But what we don't know is how much user attention this content commanded among the many billions of stories and posts that comprised the news feeds of U.S. voters, or if it changed any minds or influenced voter behavior. For example, using one of the tools we developed for this project, we can track the appearance of a given string across Facebook public facing pages, Twitter, Reddit, Instagram, and the first few posts in 4chan discussions on a timeline down to the minute or second (depending on the platform). Using that utility, we tracked a particular story with strong indications of Russian influence (about John Podesta participating in satanic rituals), and compared all accounts to both the congressionally-released list of Russian Twitter accounts and an additional list of tens of thousands of accounts that are highly likely to be Russian bots or sockpuppets. While we did observe that 5–15% of the accounts were of such provenance at any given minute, the timing and pattern of their participation was unremarkable — a background presence of more-or-less stable existence, rather than a strategic intervention in either timing or in direction, and they therefore did not seem to play an important role in increasing the salience of the story in total or by tweeting it specifically at the major actors who were central in propagating the lie — Wikileaks, Infowars, or, ultimately, Sean Hannity.

It is critical not to confound the observation of a novel phenomenon with its actual impact in the world; newness is no proxy for impact. Indeed, the discovery of such phenomena is important, but the messaging around the discovery should not be conflated with evidence that the discovered behavior in fact had impact.

The rapid advancement of technologies that digitally manipulate sound, images, and videos is a source of growing anxiety. It is too soon to predict whether synthetic media will exacerbate media disorder and epistemic crisis, but if fabricated media are shared, consumed, and interpreted in ways similar to the current versions of deceptive media, it will be spread or contained by the same influential social, political, and media entities. The fantastical nature of the claims that were widely believed by significant numbers of voters — nearly half of Trump voters gave some credence to rumors that someone associated with the Clinton campaign was running a pedophilia ring — suggests that social identity, much more than any technical indicia of truth, are the foundation for beliefs and attitudes about political matters.

Our work greatly benefits by consciously focusing on the disinformation forest, rather than on its individual trees. With an eye toward the long-term dynamics between institutions, culture, and technology, we look at algorithms and technology as a relatively new element integrated into the dynamics of broader media systems while keeping grasp of many more traditional vectors and structures. We look across media and across platforms, considering the enduring influence of radio and



cable alongside broadcast television and social media platforms. Indeed, a combination of open web media publication and social media network mapping greatly enhances our understanding of the likely impact of bots, sockpuppets, political clickbait, foreign intervention, and microtargeting. Although prolific in absolute terms, the role of these modern vectors of disinformation plays an important but modest role in the overall media ecosystem.

It doesn't look like AI is coming to the rescue anytime soon, if ever

There are a number of ways in which AI may help us to address information quality issues, including:

- Automated detection and flagging of false stories; algorithmic tools to point readers to corrections and fact-checking.
- Algorithmic tools to detect and defend against coordinated attacks by propagandists.
- Computational aids to “nudge” of readers and users to expose themselves to media and reporting outside of their echo chambers, thus opening their eyes to a broader range of viewpoints.

Despite the strong advances in the field, there is little indication at present that these tools will be able to serve as independent arbiters of truth, salience, or value. Repeatedly in our work, we have seen that stories built around a kernel of true facts, framed and interpreted in materially misleading ways are more important than stories made of clear falsifiable claims. Moreover, the meaning of stories often emerges from the network of stories and frames within which they are located. Stories and their interpretation — and misinterpretation — draw upon and act through existing social and political structures and narratives. As seen in existing applications, AI is effective at channeling and amplifying sentiments and interests of affinity groups, and helping people in ways that they want to be helped, but no present efforts appear to be able to diagnose the level of subtlety involved in the most important present disinformation campaigns we have been studying.

Systemic responses are required to address systemic issues

So what do we do about all of this? There is no silver bullet, but the best thing we can do is to continue to work toward systemic responses to the systemic problems undermining today's media ecosystem and information quality. We must continue to focus attention on rebuilding and strengthening the political and civic institutions that undergird a well functioning democracy. The insights described here draw primarily on research conducted in the United States, but we need ecosystem-based analyses of media systems in more countries to understand how the vulnerabilities, resilience, and dynamics differ across different political and cultural contexts.

Strengthening media accountability mechanisms is vital, and tamping down disinformation and media manipulation is an integral part of the everyday hard work of democracy and governance. But given the

perils of placing constraints on political speech, taking a careful and cautious approach to regulation is essential. It is important to be similarly circumspect in what we require of social media companies in addressing disinformation.

Last but far from least, there is still much work to be done in understanding the nature, reach, and impact of disinformation, as well as monitoring and evaluating targeted efforts to inoculate readers and inhibit the spread of disinformation. An area of particular concern is the increasing concentration of research capability and political outreach capability in a few private hands. We must continue efforts to bolster public interest research initiatives and develop and implement structures that allow researchers access to key sources of data with mechanisms in place to ensure robust and ethical research standards.

The development, application, and capabilities of AI-based systems are evolving rapidly, leaving largely unanswered a broad range of important short- and long-term questions related to the social impact, governance, and ethical implementations of these technologies and practices. Over the past year, the [Berkman Klein Center](#) and the [MIT Media Lab](#), as anchor institutions of the [Ethics and Governance of Artificial Intelligence Fund](#), have initiated projects in areas such as social and criminal justice, media and information quality, and global governance and inclusion, in order to provide guidance to decision-makers in the private and public sectors, and to engage in impact-oriented pilot projects to bolster the use of AI for the public good, while also building an institutional knowledge base on the ethics and governance of AI, fostering human capacity, and strengthening interfaces with industry and policy-makers. Over this initial year, we have learned a lot about the challenges and opportunities for impact. This snapshot provides a brief look at some of those lessons and how they inform our work going forward.

Looking Into the Crystal Ball

M

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

JANUARY 9–10, 2020 SANTA CLARA, CA



THE PREMIER SOURCE FOR CLE

NATIONAL INSTITUTES

Looking into the Crystal Ball

January 10, 2020 / 4:30 p.m.



THE PREMIER SOURCE FOR CLE

Looking into the Crystal Ball

January 10, 2020 / 4:30 p.m

Speakers

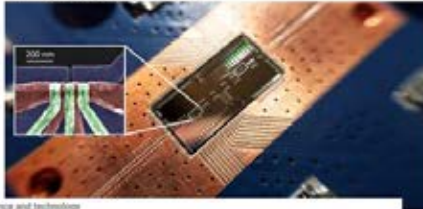
- Prof. Keith Abney (Department of Philosophy, California Polytechnic State University)
- Ryan Clark (Associate, Baker Botts)
- Daniel Dries (Lead IP Counsel, Psi Quantum)
- Stephen Wu (Shareholder, Silicon Valley Law Group) [Moderator]

What is Quantum Computing?

NATIONAL INSTITUTES

Artificial Intelligence and Robotics

Quantum Computing Breakthrough: Silicon Qubits Interact at Long-Distance



science and technology
Georgia Tech Collaborates with IBM to Develop Software Stacks for Quantum Computers



IonQ Helps Bring Quantum Computing to Life on Amazon Web Services

IonQ Participates in the Launch of Amazon Braket to Leverage IonQ's Unique Approach to Quantum Computing



YEAR AHEAD 2020

Quantum computing could solve problems we don't even know we have



technology
Intel Releases the Horse Ridge Chip for Quantum Computing!



Everyone is jumping on the quantum computing bandwagon, but why?

TECHNOLOGY 20 December 2019



The 'Quantum Computing' Decade Is Coming—Here's Why You Should Care

By Chris Roberts • 12/18/19 7:00am

Podcast: The Overhype and Underestimation of Quantum Computing

Quantum States (Qubits) are not...



$= |1\rangle$



$= |0\rangle$

a NOT gate \rightarrow flips the coin



$= |?\rangle$

a SHAKE gate \rightarrow randomizes the coin

BUT in quantum, an H gate \rightarrow creates a “superpositions state”
(when you observe the coin, sometimes 1 and sometimes 0)

Quantum superpositions $|?\rangle$ are **NOT** due to some classical randomization of the qubit
They are as “fundamental” states, different from $|1\rangle$ or $|0\rangle$ (they are neither 1 nor 0, or both 1 and 0?).

Entanglement is weird...

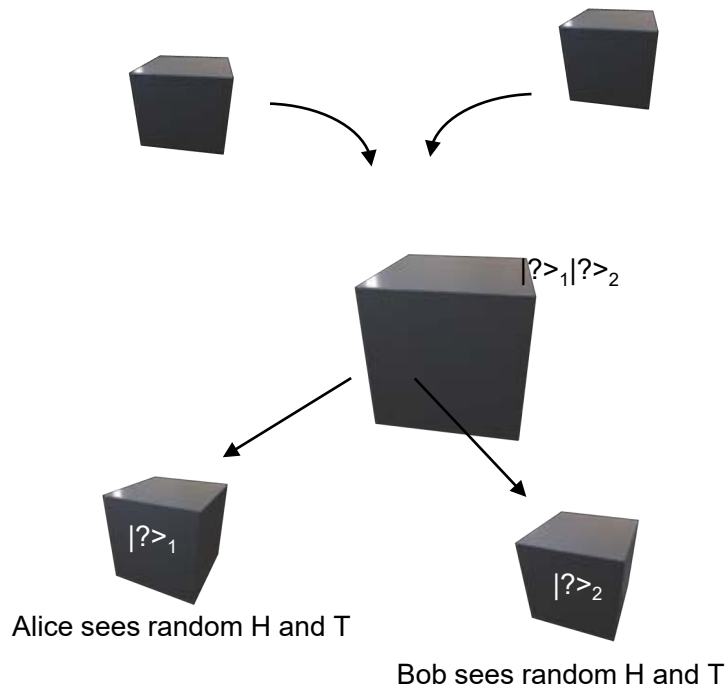
Classical bit coins:

Alice's data for coin 1: H, H, T, T, T, T, H, H, H, H, T, H, H, T, H, H

Alice: "Yup, my results are random."

Bob's data for coin 2: H, T, T, H, T, H, H, T, H, H, H, H, T, T, H, H

Bob: "Yup, my results are random, too."



Entangled qubit coins:

Alice's data for coin 1: H, H, T, T, T, T, H, H, H, H, T, H, H, T, H, H

Alice: "Yup, my results are random."

Bob's data for coin 2: T, T, H, H, H, H, T, T, T, T, H, T, T, H, T, T

Bob: "Yup, my results are random, too...BUT perfectly anticorrelated with yours!"



Exponential Scaling

(amount of resources required $\rightarrow 2^N$)



							128
256	512	1024	2048	4096	8192	16384	32768
65536	131K	262K	524K	1M	2M	4M	8M
16M	33M	67M	134M	268M	536M	1G	2G
4G	8G	17G	34G	68G	137G	274G	549G
1T	2T	4T	8T	17T	35T	70T	140T
281T	562T	1P	2P	4P	9P	18P	36P
72P	144P	288P	576P	1E	2E	4E	9E

N=20 \rightarrow 350 lbs of rice halfway down the third row.

N=24 \rightarrow 3-4 million lbs. (roughly 100 shipping containers)

N=64 \rightarrow something like 6×10^{12} metric tons, or 2,000 times the entire world's production of wheat in 2014 per wikipedia).

Legal Issues with Quantum Computing

- Patents
- Data Security – current Board-level risk?
- Privacy – GDPR/CCPA
- Export Controls

10-Year Quantum Computing Plan

- National Quantum Initiative Act
 - \$1.2 billion over ten years to accelerate R&D
 - National Quantum Coordination Office
- NIST
 - \$80 million per year through 2023 for QIS R&D
- NSF
 - \$50 million per year through 2023 to establish 2 to 5 Multidisciplinary Centers for Quantum Research and Education
- DOE
 - \$25 million per year through 2023 to establish 2 to 5 National QIS Research Centers

Artificial General Intelligence and Superintelligence



Human enhancement: a narrow view

Minimally
functioning
person

Einstein



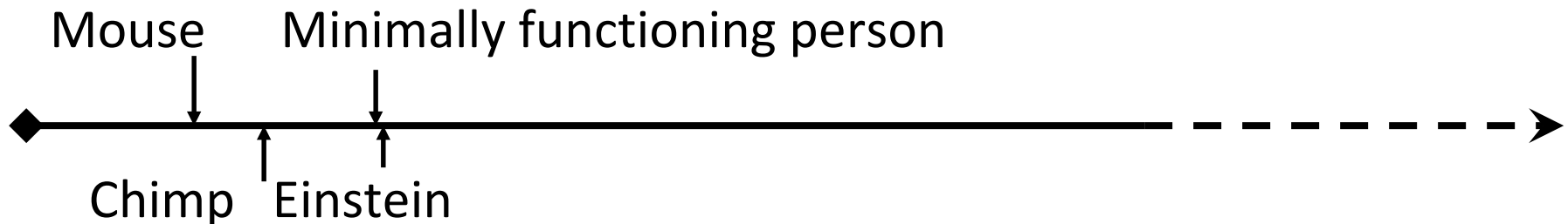
Human enhancement: a narrow view

Minimally
functioning
person

Einstein



A broader view (superintelligence):



Prospects for Artificial General Intelligence

- What is Artificial General Intelligence?
- Why is AGI an important issue?
- Will we see AGI in our lifetimes?
- Relationship of consciousness to the law
- Can we create consciousness? - Chinese Room vs Piecemeal replacement argument

Superintelligence

- What is superintelligence?
- How will we achieve it - and does that matter?
- Why is superintelligence an important issue?
- Does embodiment matter - genies, oracles, etc.?
- Person-affecting vs person-neutral ethics?

Superintelligence

- What is the “control problem”?
- Malevolent superintelligence vs perverse instantiation and problems of existential risk?
Will it decrease or increase net existential risk, and importance vs other problems?
- Prior vs posterior restraint: how regulate research?
- Personhood: Slavery vs mere servitude?

The Legal System in an Era of AGI and Superintelligence

- What changes to the law will be required?
- Models of regulation
- Legal recognition of human primacy in an AGI/superintelligence era
- Prospects for machines having rights and models for machine personhood

Questions?

- Prof. Keith Abney | kabney@calpoly.edu
- Ryan Clark | ryan.clark@bakerbotts.com
- Daniel Dries | ddries@psiquantum.com

- Stephen Wu | ssw@svlg.com

The rise of AGI/robots and artificial personhood: near- to far-term ethical & legal implications

Keith Abney

Cal Poly State Univ., San Luis Obispo

Ethics + Emerging Sciences Group

ethics.calpoly.edu

2 December 2019

This research was partially funded through a NSF collaborative grant between Cal Poly and the University of Florida, entitled “Artificial Intelligence and Predictive Policing: An Ethical Analysis.” National Science Foundation award #1917707.

Looking into the Crystal Ball

There are coming legal and ethical challenges from artificial general intelligence (AGI) and quantum computing. This presentation will provide an overview of the promise, possibilities, and potential threats from AGI and quantum computing, and discuss potential tools for governance, including policies, procedures, and subordinate documentation to support safe and transparent use of these technologies.

First off, I will examine how AI will speed the merging of war and policing.

Jus ad Vim (the just use of force, short of war)

Distinction between “measures short of war,” e.g. no-fly zones, pinpoint missile strikes, covert operations;

Vs “actual warfare,” typified by a ground invasion or a large-scale bombing campaign.

Walzer: Former are technically acts of war under IHL, but “it is common sense to recognize that they are very different from war”

jus ad vim involves: diminished risk to own troops; smaller, more predictable destruction; far fewer civilian casualties; cost far less.

Policing and AI/ robots?

- ✦ *So: increasing jus ad vim actions tempting for politicians, which raises possibility of more unjust coercion/ 'war'*
- ✦ *Increasing worry - where's line between jus ad vim and ordinary policing/ peacekeeping?*
- ✦ *Terrorism and 'law enforcement' vs 'military' solutions; useless distinction?*
- ✦ *e.g., Islamic terrorist threat: best results from military and law enforcement working together?*
- ✦ *so, jus ad vim as new normal? Role of AI/ robots?*
- ✦ **1 Predictive policing**
- ✦ *LA 2012 pilot project; now used over 50 cities*
- ✦ *Benefit: it works at deterring crime, also helps in rapidly apprehending suspects*

- **Predictive policing: benefits, costs, and worries**

- ✦ *Reasons why benefit: crime follows particular patterns (time of day, place, ...) and AI can predict and target limited police resources better than any human*
- ✦ *AI arguably less biased, prone to mistakes/ bad 'hunches' than human cops*
- ✦ *But: racial profiling, algorithmic bias; discriminatory outcomes even without intent?*
- ✦ *Unfair burdens on POC?*
- ✦ *Should we aim for perfect enforcement of laws - lead to “authoritarian lock-in” or “quarantine”?*
- ✦ *Generally, what number of false positives are worth a decrease in false negatives?*

Third party defense?

- ✿ *In JWT, including jus ad vim, idea that a 3rd party can intervene against aggressor of behalf of intended victim.*
- ✿ *Also apply to predpol/ **AI policing - just institutionalized 3rd party defense?***
- ✿ *But what if 3rd party defense poses risk of harm to other innocent individuals, aka ‘collateral damage’?*
- ✿ *How much is too much?*

2 The rise of robots: complications

- ✦ *The foregoing discussion assumes AI only serves human soldiers/ police/ peacekeepers*
- ✦ *But: do the **types of combatants** matter for the ethics of peace and war? Large/ small nation-states? Non-state actors? Even businesses (cyber/ space)?*
- ✦ *Or - perhaps **the types of weapons**? Lethal vs non-lethal? Ships vs planes? Ballistic missiles vs bullets? 'Rods from God' vs nukes?*
- ✦ *Or, **who/ what's doing the fighting/ policing**? Humans, 'human in/ on the loop' robots, or fully (L)AWS?*
- ✦ ***Autonomy worry**: for war/ jus ad vim and policing, should human commanders be in, on, or out of the loop?*

Relevance of LAWS to issues

- ✦ **Common *interdependence* argument:** LAWS would cause unjust wars,
- ✦ *because they greatly decrease the cost of war, emboldening political leaders to fight unjust wars*
- ✦ *So, **jus ad bellum** worries justify **jus in bello** restrictions on LAWS;*
- ✦ *Would we the same about jus ad vim/ policing?*
- ✦ *What if they enable more just wars/ better policing?*
- ✦ ***Again, how do we weigh false positives vs false negatives?***

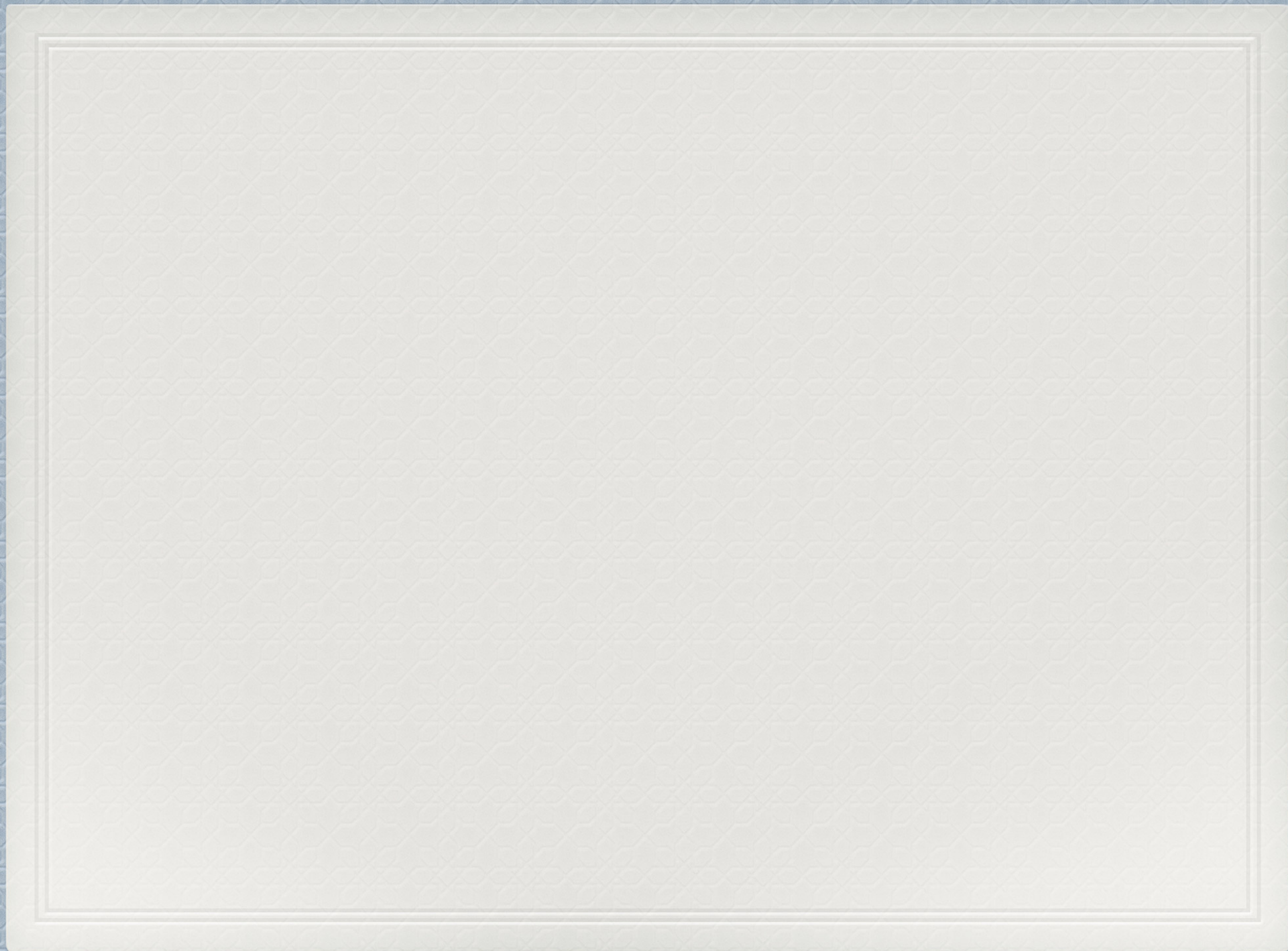
Relevance of AWS continued

- ✦ **Principle: 'Ought implies can':** moral responsibility always a function of capabilities
- ✦ so moral issues subject to change as new tech emerges
- ✦ Latency and other aspects of robotic technical superiority, including no need to prioritize self-defense, may make robots morally superior choices to human police/peacekeepers/ warfighters
- ✦ Just war in some battlefields (e.g. space) plausibly **requires LAWS (military necessity)**

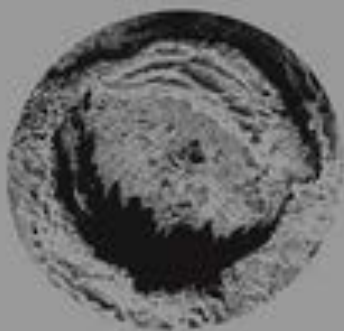
Technological asymmetry?

- ✦ *But how will communities (esp. disadvantaged communities) respond to technically superior robotic peacekeeping?*
- ✦ *Especially when that seems to focus on their communities and not apply to communities of privilege (e.g., predpol)*
- ✦ *Will it be welcomed as a resource to keep the peace?*
- ✦ *Or resisted as a means of quarantine and enforcing a second-class status?*

- ✦ ***First Conclusion: Anticipatory ethics and solving the 'first generation problem' are the key***



Next, AGI/Robotic Consciousness: existential risk and related issues



Chicxulub crater

65 million years ago
at least 150 kilometers (93 miles) wide

source: USGS



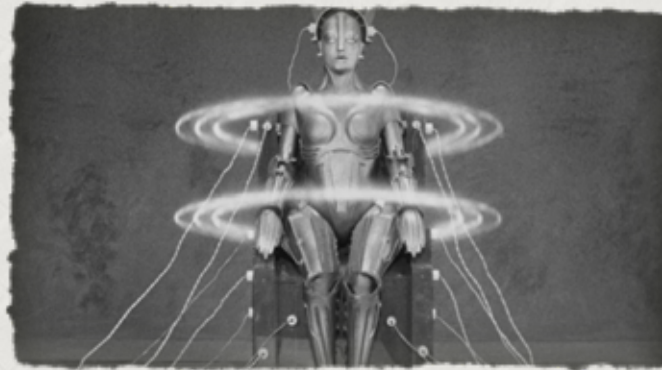
Crater formed by asteroid 3.26 billion years ago
Approximately 478 kilometers (297 miles) wide



Island of Hawaii

122 kilometers (76 miles) wide
source: Hawaii economic data book

The trope of AI/Robotic Consciousness Leading to Rebellion, a scifi staple



Defining AI and AGI

- ✿ Russell and Norvig's (2002) *Artificial Intelligence: A Modern Approach*: defining AI by its telos along two axes; does AI match or surpass human abilities in rational thinking, or in the sophistication of its behavior?
- ✿ Assess AI by how it thinks, or by what it does?
- ✿ Natural vs artificial? Natural intelligence evolved to solve the problems of evolution, by having (re Popper on the scientific method) ideas about survival and reproductive strategies die in our stead. (Darwin awards?)
- ✿ So best define (narrow) AI: a goal-oriented, problem-solving thinking process, with at least some narrow human-level (or better) capabilities, that arose artificially, not naturally.
- ✿ Like artificial vs natural flying machines, AI may achieve the same cognitive goals as natural intelligence through very different means
- ✿ AGI: when AI has matched or bettered average human cognitive abilities generally, not just for a narrow set of cognitive goals.

AI: Moral status?

- ✦ *Ratiocentrism vs sentientism (also biocentrism, cosmocentrism, ...)*
- ✦ *‘Zombie/ robot argument’ (Abney 2019): AI could pass one “moral Turing test” (Sparrow 2012), yet experience no pleasure/ pain*
- ✦ *Wireheading argument (Abney 2019): AI could turn us all into hedonium... malignant failure mode, not moral success!*
- ✦ *Confusing final vs intrinsic vs instrumental value (re Korsgaard 1983)*
- ✦ *Yes, AI could have intrinsic moral status... but not yet.*
- ✦ *So what will happen when it does?*

The prophets of doom?

- ✦ *Stephen Hawking, Bill Gates, Elon Musk, etc. concerned about ‘control problem’. Hawking (2014): “Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.”*
- ✦ ***Beneficial AI**, January 2015: Nick Bostrom joined Musk, Hawking, Max Tegmark, Lord Martin Rees, etc, in signing the **Future of Life Institute**'s open letter speaking to the potential risks and benefits associated with artificial intelligence.*
- ✦ *The signatories “...believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.”*

- ✦ **Tegmark:** *Main risks of AGI come not from malevolence or consciously evil behavior per se, but rather from the misalignment of the goals of AGI with those of humans.*
- ✦ **Bostrom:** *AGI revolution turns many philosophical problems into practical political questions and forces us to engage in “philosophy with a deadline”*
- ✦ **Timeline:** *no one knows – but expert median estimate for AGI is 2050; some as early as late 2020s*

✧ **Bostrom:** *Would X-risk prohibit – or require-AGI/ robots as full persons?*

✧ *Pro: help with all other existential risks*

✧ *Con: AGI decides to exterminate humanity*

✧ **Forms of AGI:** *Speed, Collective, or Quality Superintelligence?*

✧ **Problems: Singularity Argument ('fast uptake'):** *If sudden "intelligence explosion," then can't test on intermediate AIs; perhaps no precautions at all, if explosion catches creators completely by surprise. Possible result:*

✧ *Decisive strategic advantage - singleton?*

- ✿ *Control problem and full autonomy - **how keep a human in/ on the loop?***
- ✿ *Attempts: **Boxing methods, Incentive methods, Stunting, Tripwires***
- ✿ *Result: Oracles - just answer questions*
- ✿ *Genies - just execute commands (spirit vs letter?)*
- ✿ *Sovereigns - full autonomy*
- ✿ *Tool-AIs - not designed for goal-oriented behavior, just search, etc.?*
- ✿ *But we control more physically powerful entities by being smarter, so AGI as oracle/ tool/ genie seems unlikely to work....*

✿ *Person-affecting vs person-neutral ethics? (**Former bids speed, latter caution?**)*

✿ *Instrumental versus final goals*

✿ *Final goals – given, or learned? Relation to intrinsic value?*

✿ *Reinforcement learning and proximal vs distal goals: cf. evolution and eating/ reproduction*

✿ *Coherence as necessary (sufficient?) condition – cf. Kant/ Korsgaard*

- ✿ *Perverse instantiation (validation?) - e.g., Paperclip maximizer?*
- ✿ *Program verification and the problems of value and goal alignment: ML vs GOFAI*
- ✿ *Evolutionary algorithms?*

- ✿ *Does it matter how we get there?*
- ✿ *classic top-down AI (GOFAI)*
- ✿ *Whole brain emulation*
- ✿ *Neuromorphic AI (mix)?*

- ✿ *Bostrom: Last is most plausible – and most dangerous*

- ✦ *Bostrom commends **CEV**: coherent extrapolated volition?*
- ✦ *But: possible to have coherent Caligula, maximize suffering?*
- ✦ *MR instead? Need correct ethical theory to avoid ‘**malignant failure mode**’, e.g.:*
- ✦ *Hedonistic utilitarianism (Bentham, Mill) and computronium/ hedonium*
- ✦ *Petersen: PI and diachronic rationality as solution (to avoid Parfit problem, ‘future-Tuesday-indifferent’)*

- ✦ **Conclusions:** *is superintelligence likely to be superethical?*
- ✦ *Perverse instantiation is an issue, because final goals typically underspecified/ vague*
- ✦ *So underspecified final goals must be learned by value alignment*
- ✦ *Final goals best learned by seeking coherence in practical reasoning, so*
- ✦ *Practical reasoning demands coherence with future final goals (**goal content integrity**)*

- ✿ *That requires consistency with not just future self, but all successors – AGI cloning/ fission and ‘teleological thread’ vs classical PI*
- ✿ *So AGI will learn complex goals by (largely) respecting our shared intentions, and hence be ethical*
- ✿ *But: possibly simple goals that do not require ML (like paperclip maximizing)?*
- ✿ *So, 30% chance superintelligence will be superethical (Petersen)?*

So, is an AI/ROBOT Rebellion/takeover inevitable? And if so, is that necessarily bad?

The AGI/ robot rebellion need not be violent. Ethical progress: perhaps now we wouldn't exterminate the Neanderthals...

- ✿ *And, Bostrom's 'simulation argument': Even if inevitable, perhaps result is that this is not now 'base reality'; perhaps we are in an individual **ancestor simulation**, set up by an AI to explore (and learn from) an agent's reaction to "interesting times"?*

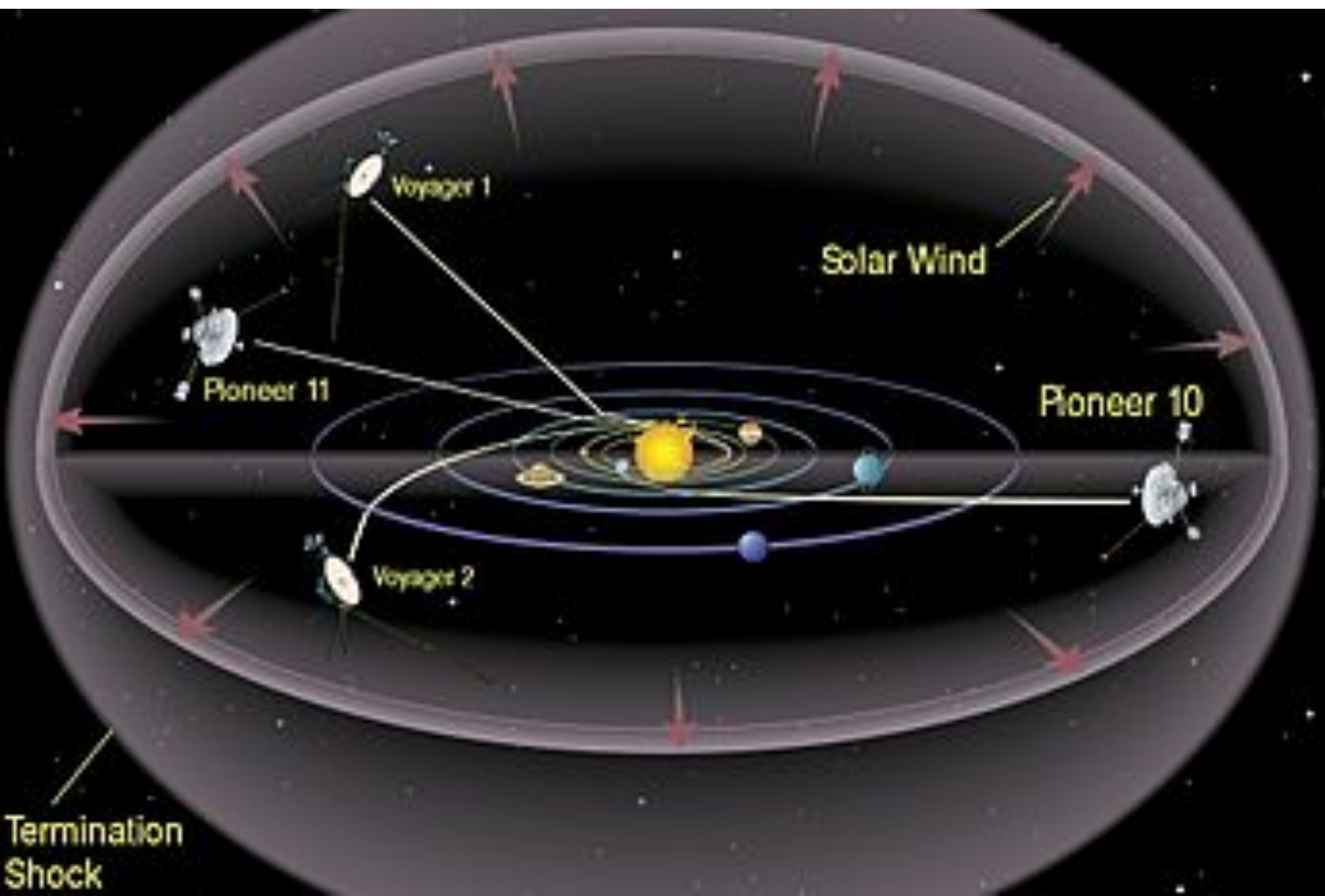
AGI scenarios: LIBERTARIAN UTOPIA	Humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property rights.
BENEVOLENT DICTATOR	Everybody knows that the AI runs society and enforces strict rules, but most people view this as a good thing.
EGALITARIAN UTOPIA	Humans, cyborgs and uploads coexist peacefully thanks to property abolition and guaranteed income.
GATEKEEPER	A superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. As a result, helper robots with slightly subhuman intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.
PROTECTOR GOD	Essentially omniscient and omnipotent AI maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough that many humans even doubt the AI's existence.
ENSLAVED GOD	A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.
CONQUERORS	AI takes control, decides that humans are a threat/nuisance/waste of resources and gets rid of us, perhaps by a method that we don't even understand.
DESCENDANTS	Als replace humans, but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who's smarter than them, who learns from them, and then accomplishes what they could only dream of — even if they can't live to see it all.
ZOOKEEPER	An omnipotent AI keeps some humans around, who feel treated like zoo animals and lament their fate.
1984	Technological progress toward superintelligence is permanently curtailed not by an AI but by a human-led Orwellian surveillance state where certain kinds of AI research are banned.
REVERSION	Technological progress toward superintelligence is prevented by reverting to a pre-technological society in the style of the Amish.
SELF- DESTRUCTION	Superintelligence is never created because humanity drives itself extinct by other means (say nuclear and/or biotech mayhem fueled by climate crisis).

- But, plenty of time? *Problem: our space robots*, particularly Voyager 1 and 2.
- **Interstellar Doomsday Argument**
- Begin with *Self-Sampling Assumption (SSA)*:
- “One should reason as if one were a random sample from the set of all observers in one’s reference class”.
- So, for a random observation:
- 95% probability phenomenon will continue for between $1/39$ and 39 times its present age

✧ Interstellar Doomsday Argument,

✧ Continued

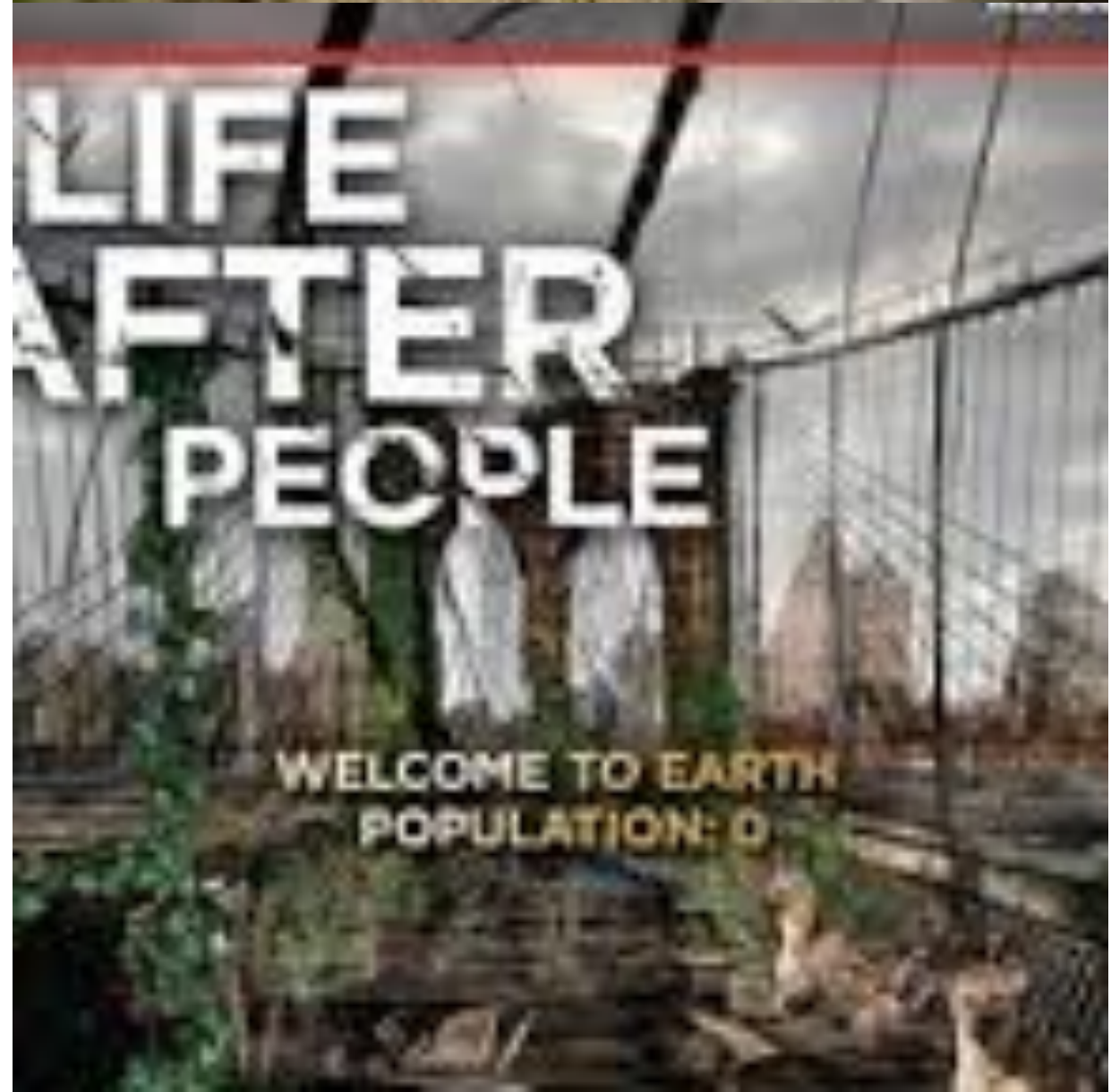
- ✧ *Only 5% chance* your random observation comes in the first or last 2.5% of its lifetime of the phenomenon in question
- ✧ *Voyager 1* entered interstellar space in 2012
- ✧ So in 2019, it's 95% probable that our time left as a civilization that sends interstellar robots
- ✧ is between $L/39$ (for $L = 7$ years, that's 66 more days) and $39L$ (273 more years)
- ✧ Most plausible way we stop sending robots: human extinction



- ✧ *von Neumann probes* cinch problem?
- ✧ With feasible speed $c/40$, saturation of entire galaxy would take 4 million years, less than 1/3,000th of age of Milky Way.
- ✧ If not saturated by robotic probes from ancient aliens, implausible it's because they need more time to get here.
- ✧ But: '*Oumuamua* as alien lightsail? (Loeb)



- ✿ Easier: interstellar probes or sustainable human colonies?
- ✿ Conclusion: if humanity cannot escape Earth - soon - we'll go extinct
- ✿ So only good reason for humans (vs robots) to go into space is *colonization*
- ✿ For chance of success, robots will have to go first and terraform/build long-term habitats



Quantum Leap:

Accelerating America's Growth in Quantum Computing

Ryan C. Clark
Baker Botts LLP
Palo Alto, California

The National Quantum Initiative Act (H.R. 6227) ("the Act") authorizes \$1.2 billion over ten years to accelerate quantum computing research and development in the United States. Signed into law on December 21, 2018, the Act received overwhelming bipartisan support and passed unanimously in the Senate and by a vote of 348-11 in the House of Representatives. The law provides a 10-year coordinated strategy for the White House and specified federal agencies to support and encourage the growth of quantum technology in the United States.

What Is a Quantum Computer?

Quantum computers use strange properties of subatomic particles to process information more quickly than classical computers. Traditional computers work using bits of information that can exist in two states: 1 or 0. At the machine level, this binary state is represented using electrical circuits that are either high (1) or low (0). Quantum computers are fundamentally different. Quantum computers work using quantum bits of information, or "qubits,"¹ that take advantage of physics properties only apparent at a subatomic level.

For example, quantum computers take advantage of the subatomic property of superposition where the qubit state can be 1, 0, or both simultaneously. To illustrate superposition, think of an imaginary sphere. A classic bit can be in only two states – at the top (1) or bottom (0) of the sphere. But a qubit can be any point on the sphere at once. This allows quantum computers to store and process much more information than a traditional computer.

Quantum computers also take advantage of the subatomic property of quantum entanglement. Entangled qubits always have the same state, even if they are on the other side of the world. Due to entanglement, qubits need to be described both in terms of the states of each qubit and also in terms of the correlations between the individual qubits. The correlations between individual qubits grows exponentially: for n qubits, there are 2^n correlations.

These strange quantum properties are the reason why quantum computing holds such promise in certain types of tasks, such as finding very large prime numbers. Because factoring large prime numbers is important in cryptography and blockchain technology, it's likely that quantum computers could crack many current systems and databases. Quantum computers may also have applications in pharmaceutical molecular modeling, secure communications, quantum simulation, sifting through large data sets, financial modeling, search and machine learning algorithms, artificial intelligence, and weather forecasting.

What Is the 10-Year Plan?

The Act provides a 10-year roadmap for a coordinated Federal program to accelerate quantum research and development for the economic and national security of the United States. The law directs the President to establish a National Quantum Initiative Program and a Subcommittee on Quantum Information Science (“Subcommittee”) with coordination from:

1. the National Institute of Standards and Technology (NIST)
2. the National Science Foundation (NSF)
3. the Department of Energy (DOE);
4. the National Aeronautics and Space Administration (NASA);
5. the Department of Defense (DoD);
6. the Office of the Director of National Intelligence (ODNI);
7. the Office of Management and Budget (OMB);
8. the Office of Science and Technology Policy (OSTP); and
9. such other Federal department or agency as the President considers appropriate.

The Subcommittee’s responsibilities include providing a 5-year strategic plan by December 21, 2019 and a subsequent 5-year strategic plan by December 21, 2024. The Act provides additional plans for three federal agencies: NIST, NSF, and DOE.²

NIST Quantum Activities

NIST is allocated \$80,000,000 per year through 2023 to support and expand quantum information science research and development. This support includes training scientists and establishing collaborative ventures with public or private sector entities. NIST will also convene a consortium of stakeholders by December 21, 2019 to identify the future measurement, standards, cybersecurity, and other appropriate needs for supporting the development of a robust quantum information science and technology industry in the United States.

NSF Quantum Activities

In addition to continuing its basic research and education program on quantum information science and engineering, NSF is allocated \$10,000,000 per year through 2023 to establish between 2 and 5 Multidisciplinary Centers for Quantum Research and Educations (“NSF Centers”). The purpose of the NSF Centers shall be to advance, support curriculum and workforce development, and foster innovation in quantum information science and engineering. The NSF will accept applications from institutions of higher education or nonprofit organizations with descriptions of its short-term and long-term plans for collaborating, advancing science, and becoming self-sustaining after the expiration of funding.

DOE Quantum Activities

In addition to continuing its basic research and coordination on quantum information science, DOE is allocated \$25,000,000 per year through 2023 to establish between 2 and 5 National Quantum Information Science Research Centers (“DOE Centers”). The purpose of the DOE Centers shall be to accelerate scientific breakthroughs in quantum information science and technology. The DOE will coordinate with, and ensure unnecessary duplication with, industry, institutions of higher education, and the activities of other DOE research entities, including the Nanoscale Science Research Centers, the Energy Frontier Research Centers, the Energy Innovation Hubs, and the National Laboratories.

How Can I Get Involved?

The ABA's Section of Science & Technology Law has established a Task Force on Quantum Computing to provide educational resources to the legal community about the unique legal, business, and technical challenges posed by quantum computing. To get involved or learn more, please contact the Chair of the Task Force on Quantum Computing:

Ryan C. Clark
Baker Botts, L.L.P.
ryan.clark@bakerbotts.com

Endnotes

¹ While qubit (quantum bit) is an excellent portmanteau, thanks are due to Lewis Carroll for the introduction of portmanteaus to our lexicon. In 1871, Carroll's *Through the Looking Glass* used a talking egg named Humpty Dumpty to first define portmanteaus, including mimsy (miserable and flimsy), slithy (slimy and lithe), galumph (gallop and triumph) and chortle (chuckle and snort).

² Despite winning the 1992 Golden Globe Award for the "Best Performance by an Actor in a TV Series" and a four-year, consecutive streak from 1990-1993 as the Quality TV Awards "Best Actor in a Quality Drama Series," it is unclear whether Scott Bakula will have any responsibilities for this quantum leap.



NEURAL DEVICES WILL CHANGE HUMANKIND

What Legal Issues Will Follow?

By **Stephen S. Wu** and **Marc Goodman**

Though it is not widely known, brain implants and other neural devices have been successfully used for several years to treat neurological disease and brain injuries. In the future, these devices hold the promise of enhancing our quality of life and ultimately expanding the functionality of our minds. For instance, neuroprosthetic devices will interface with the nervous system to control prosthetic limbs. Moreover, new brain-computer interfaces and devices may someday duplicate some or all of the functionality of the human brain.

Some futurists and artificial intelligence experts envision credible scenarios in which synthetic brains will, within this century, extend the functionality of our own brains to the point where they will rival and then surpass the power of an organic human brain. At the same time, humans seem to have no limitations when it comes to finding ways to attack the computerized devices that others have invented. Attackers have successfully compromised computers, mobile phones, ATMs, telephone networks, and even networked power grids. If neural devices fulfill the promise of treatment, and enhance our quality of lives and functionality—which appears likely, given the preliminary clinical success demonstrated from neuroprosthetics—their use and adoption will likely grow in the future. When this happens, inevitably, a wide variety of legal, security, and public policy concerns will follow.

We will begin this article with an overview of brain implants and neural devices and their likely uses in the future. We will then discuss the legal issues that will arise from the intersection among neural devices, information security, cybercrime, and the law. Finally, we will close with our thoughts on how lawyers will deal with these new legal issues.

Stephen S. Wu, a partner in the Silicon Valley law firm Cooke Kobrick & Wu LLP, practices in the areas of information technology and intellectual property litigation and transactions, and served as the 2010–2011 Chair of the ABA Section of Science & Technology Law. Marc Goodman is the founder of the Future Crimes Institute and has worked globally with Interpol, NATO, and the United Nations on emerging technosecurity threats. He serves as Chair for Policy, Law & Ethics at Singularity University and participates in several committees of the ABA Section of Science & Technology Law.

The Development of Neural Devices

Brain-computer interface (BCI) or brain-machine interface (BMI) devices have been under development for a long time, with some devices having already reached the market. Currently, neural devices range from toys and games to research projects and some commercialized products. Turning first to toys and games, some companies are beginning to use brain wave technology. For instance, at the Game Developer's Conference in 2008, a company called NeuroSky, Inc. demonstrated a game in which users can manipulate objects on the screen with their thoughts alone. Both NeuroSky's MindWave and Emotiv's EPOC use sensors in an external headset to "read" the magnitude of users' brain waves and perform an action on the screen, such as moving a cursor or typing a key on a keyboard.¹ Similarly, a Star Wars-licensed product, The Force Trainer, is a toy in which a headset measures a child's brain waves and, when the child concentrates, converts brain waves into a signal to cause a ball to rise in a tube. Videos demonstrating these products are available on YouTube.²

Numerous research projects are underway to develop BCI technologies and neuroprosthetic devices. For instance, researchers are looking at the cellular level to determine how brain cells can control information technology processes. Researchers are also working on devices to control prosthetic limbs. One researcher cultured brain cells and used them to control a fighter plane flight simulator.³ Other research projects in the neuroprosthetic area involve an interface between the nervous system and a device allowing a user to control a prosthetic limb. Next-generation devices may provide users with a "touch" sensation by using sensors in the prosthetic limb to transmit signals back to the brain.⁴ Other research projects use "deep brain stimulators" to stimulate the brain, producing promising results for the treatment of Parkinson's disease, Alzheimer's disease, chronic pain, and other conditions.

Another fascinating research project concerns the study of how implantable medical devices directly attached to our brains can allow us to control a cursor on a computer screen, type, and send emails with our thoughts alone. This research involved attaching an interface device to a man who is paralyzed and unable to use his limbs to control a mouse or a keyboard. Such devices hold great promise to aid those with spinal injuries or diseases like Lou Gehrig's disease that impair movement.⁵

Perhaps the most common implanted neural device on the market now is the cochlear implant. Cochlear implants help the deaf or hearing impaired to hear better by taking in sound, converting it into electric signals, and interfacing with the nervous system so that the brain can receive and process the signals to generate sound in the mind of the wearer.

The Future of Neural Devices

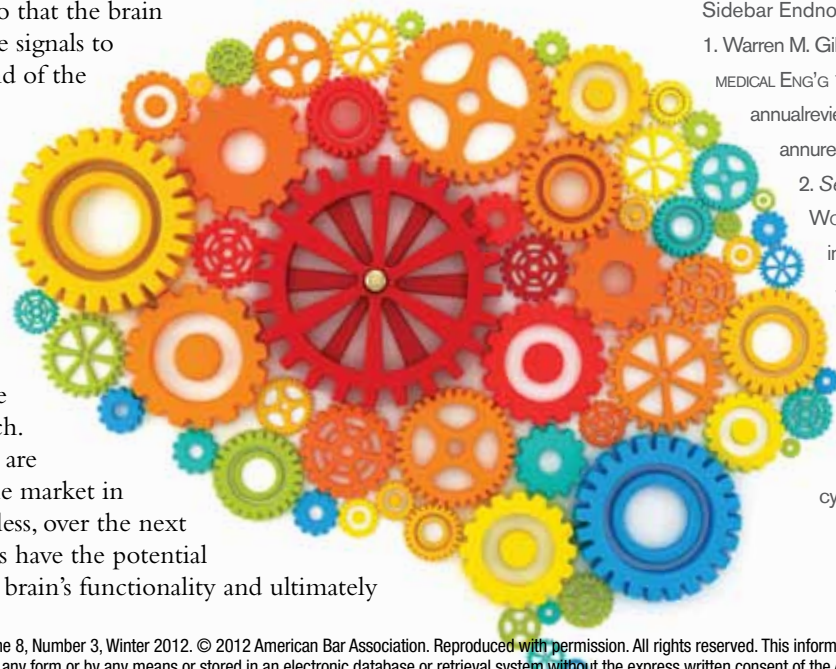
In upcoming decades, we will continue to develop, sell, and use neural devices as toys and games, therapies for disease and injury, and for the purposes of further research. Some of the products that are in development will hit the market in upcoming years. Nonetheless, over the next half century, neural devices have the potential of radically enhancing the brain's functionality and ultimately

WHAT ARE NEURAL DEVICES?

"Neural interfaces are connections that enable a two-way exchange of information with the nervous system. These connections can occur at multiple levels, including with peripheral nerves, with the spinal cord, or with the brain. . . ." Brain implants are a specific kind of neural device placed on the surface or the cortex of the brain that create an interface between the nervous system and microchips in order to treat damaged parts of the brain² or, in the future, to enhance its functionality. "By connecting intimately with computers, we will take the human brain to a new level. . . . If we can provide the brain with speedy access to unlimited memory, unlimited calculation ability, and instant wireless communication ability, we will produce a human with unsurpassable intelligence."³

Sidebar Endnotes

1. Warren M. Gill, et al., 11 ANN. REV. BIO-MEDICAL ENG'G 1 (2009), available at www.annualreviews.org/doi/pdf/10.1146/annurev-bioeng-061008-124927.
2. See *Brain Implant*, ARTICLE WORLD, www.articleworld.org/index.php/Brain_implant.
3. Sherry Baker, *The Rise of the Cyborgs*, DISCOVER MAG., Oct. 2008, available at <http://discovermagazine.com/2008/oct/26-rise-of-the-cyborgs>.



changing the human condition.

Some futurists predict that brain implants and other neural devices will dramatically expand the capabilities of the human brain. Chief among these futurists is Ray Kurzweil, who has written numerous books on the future of technology, including *The Singularity Is Near*.⁶ These futurists point to trends such as exponential increases in speed and power of computer processors, the chief example of which is Moore's Law. Moore's Law holds that the number of transistors that can fit on a chip roughly doubles every two years. Moore's Law correlates with the speed and power of chips. With the benefit of ever more powerful computers and the miniaturization of computing devices, brain implants and other neural devices have the potential of showing similar exponential increases in speed and power. Futurists believe that such speed and power will give neural devices the capability of greatly enhancing the functioning of the human brain.

Even the benefits of enhanced memory through the use of information technology itself would radically change our lives. Right now, our organic brain has limited memory; we can only perceive and retain so much information at once, and we forget a great deal of information over our lifetimes. If we had devices directly connected to our brains that could enhance our memories, whether onboard in a brain implant or offboard in devices connected to our

brains, we could learn new subjects rapidly and could recall and use information quickly. With the ever-increasing storage capacity afforded by information technology, we would not need to forget information—unless we wanted to. These technologies may significantly assist efforts to overcome diseases that cause the loss of memory.

Kurzweil and other futurists, however, go further than simply envisioning the use of information technology to enhance human brains. They believe that in this century, researchers will achieve ever-greater understanding of how the brain works. Coupled with exponentially advancing technologies, an enhanced knowledge of brain function will allow us to develop brain implants that have the speed, power, and memory to replicate the functionality of the entire human brain. Moreover, there is no reason why the brain must be limited to onboard devices. If we can connect devices to offboard devices, we will have potentially unlimited processing power and memory. In addition, the next logical step is to connect our neural devices to the Internet so that we can share, add to, and manipulate the entire world's information.

Kurzweil has predicted that around midcentury, computers and artificial intelligence will be so powerful, humans will be able to transfer (“upload”) their lifetime's worth of memories to computers and carry on their thought processes and remember information using computers so that their thinking can be independent from their organic brains. Indeed, under this scenario, humans' thoughts and minds could outlive the death of the organic brain and physical body so that these humans would be functionally immortal. Their minds and thoughts would live on as processes run in computers. Hence, a *Time* magazine article describing this scenario bore the title “2045: The Year Man Becomes Immortal.”⁷

Under Kurzweil's view, while ever more powerful information technology enhances the power of human brains, artificial intelligence (AI) will also become more powerful. Exponential technology advances will permit AI to rival and then surpass human intelligence. Because brain enhancement technology and AI will use many of the same methods, the distinction between the machine-enhanced organic human beings and AIs will blur. Consequently, Kurzweil predicts that by the end of this century, humans and robots using strong AI will be functionally indistinguishable.

The Information Security, Privacy, and Cybercrime Threats

Against this backdrop of sweeping changes in technology, we now examine the threats to neural devices from an information security and privacy perspective. Given the significant developments unfolding in the world of BCI and neuroprosthetics, we anticipate a wide variety of potential criminal threats to the human brain itself. First, people will have the means to attack neural devices. The media have publicized stories about hacking pacemakers and other medical devices. Attackers could use similar means to attack devices on board the human body, including wireless devices, controllers for prosthetic limbs, or deep brain stimulators.⁸

Second, people have the means and the motivation to exploit neural devices. Human ingenuity has no limitation, and unfortunately, the world has many people with the desire to use whatever weapons are at their disposal to defraud, terrorize, or otherwise harm other people. At a minimum, people already use technology to stalk, annoy, and steal from one another.

Third, the track record of the use of computers and the Internet shows that people will attack and subvert computers and devices if given the opportunity to do so. We have seen computer viruses, phishing, network intrusions, online fraud, unauthorized access, records snooping, trade secret theft, cyberbullying, cyberstalking, and numerous other forms of hacking and attacks. The history of computing and the Internet are replete with examples of ingenious people coming up with novel ways to harm other people.



In short, bad actors of various kinds, with various motivations, will inevitably attack neural devices, just as they have tried to attack computers, Internet-attached devices, and other medical devices.

The threat to neural devices, however, is different in kind from the threat to computers and the Internet. Early viruses caused annoyance as some hacker bragged about proving to the world he could inject malware into some piece of software by causing the screen to display a message saying, in essence, "Gotcha!" More malicious software caused more harm, such as erasing files. As malware became more sophisticated, attackers used it to steal money, technology, trade secrets, or national secrets. Other attackers sought to use denial of service attacks to bring down critical systems.

All of these conventional attacks affected, at least in the first instance, only money, data, and other property. It is true that such attacks could indirectly lead to injury or the loss of life, such as attacks on hospitals' computer networks, the theft of military technology, or the compromise of national security information. Nonetheless, none of the systems involved directly touched human beings, and so their compromise did not result in the immediate physical harm to a human.

The hacking of medical devices poses a different kind of threat. Medical devices maintain health and sustain life. Hacking such a device could result in immediate death or injury to a human. For instance, if an attacker turned off a pacemaker remotely, he could cause the user's heart to stop and immediately kill the user.

The use of neural devices entails an even greater risk than other medical devices. Attacking a neural device used to enhance a human's memory may have the effect of wiping out some or most of someone's memory or thought processes. Imagine a hacking attack that results in the equivalent of a lobotomy. A successful attack may not only harm the person physically or mentally, but it also may deprive the victim of the very essence of that person's humanity, his or her mind and memory. Disabling a prosthetic limb is one thing, but an electronic lobotomy is something quite different. Existing law criminalizing the conduct of a person that merely "intentionally accesses a protected computer without authorization"⁹ does not begin to encompass or address the full significance of such a crime.

The Legal Issues of Brain Implants and Neuroprosthetics

Extrapolating from today's legal issues to the potential applications of neural devices in the future, we anticipate significant challenges to the legal system, including the following items.

Information Security and Privacy Policies and Controls

Companies that sell and provide services to support neural devices may have unique access to private information stored in the human brain—a depth and level of access that makes both Facebook and Google's archives seem feeble by comparison. Even the process of obtaining meaningful informed consent to the collection of data from an individual poses significant challenges. Device manufacturers will need to have data security policies to establish administrative, physical, and technical controls over the manufacture of neural devices and over the services that support such devices. Likewise, they will need privacy policies to address information practices relating to sensitive information and systems.

Information Security and Privacy Compliance

Federal and state laws may someday impose security and privacy requirements on manufacturers of neural devices and the companies that service them. Litigation arising from privacy torts and the violation of security and privacy requirements in statutes and regulations may crop up. Lawyers handling these matters may be specialists in information security and privacy laws, information technology transactions lawyers, or, in the case of privacy and security suits, litigators.

- **Products liability law:** Users whose devices are hacked may bring products liability, negligence, warranty, unfair trade practice, and related claims against manufacturers of neural devices for failing to implement security controls to prevent the compromise of these devices. Litigation lawyers will handle these matters.
- **Medical device regulation:** Manufacturers, importers, and other companies in the field will need to comply with Food and Drug Administration regulations when seeking to make or import neural medical devices. Regulations and the regulatory practice may involve obtaining premarket clearance or approval; resolving disputes with the FDA; ensuring quality of the devices; dealing with FDA inspections; handling product recalls; and overseeing advertisements to consumers. Lawyers with an FDA practice handling medical devices will likely take on these matters.
- **Criminal law:** Criminal lawyers will likely take on matters of both substantive law and criminal procedure. For instance:
 - **Substantive criminal law:** States will need to come to grips with the new neural device technologies and create statutes to strike at those attacking neural devices. For instance, the intentional wiping of someone's stored memories should be a recognized crime. Some of the existing cybercrime laws would need to be updated as well. For example, is an attack on a brain implant analogous to a physical assault on a victim, such as a punch to the face? Likewise, if an attacker wiped the memories and functionality of a person's brain device, leaving the victim in a vegetative state, with what crime would we charge the attacker?
 - **Criminal procedure:** The courts will need to apply existing criminal procedural protections to those using neural devices. For

AS NEURAL DEVICES BECOME COMMONPLACE, PRACTITIONERS WILL NEED TO LEARN MORE ABOUT THE TECHNOLOGY.

example, if I have a neural device that stores my “memories” in the form of images, videos, text, and other files in my brain, can I prevent disclosures of this electronically stored information (ESI) on the basis of my Fifth Amendment right against self-incrimination? What if my neural device is connected to offboard storage? Does the fact that a third party holds those same files mean that I have no right to resist disclosure of this ESI under the Fifth Amendment due to the happenstance that the storage is off board instead of on board in my brain?

How Neural Devices Will Change Lawyers’ Practices

We believe that the development and commercialization of neural devices is one example of the convergence of information technology and life sciences. Practitioners have an opportunity to

create and shape a new area of law at the intersection of information technology and neuroscience. As lawyers have changed their practices to accommodate computers, the Internet, and other advances in technology, they will change their practices to take neural devices into account.

First, as neural devices become commonplace, practitioners will need to learn more about the technology. This step is the foundation for all science and technology law practices. The ABA Section of Science & Technology Law’s committees provide forums for scientists, lawyers, and businesspeople to come together, learn about each other’s fields, and create educational publications for the industry and bar. The Behavioral and Neuroscience Law Committee and Information Security Committee of the Section will, I am sure, work together in the future on these issues. Lawyers can talk with technologists and business contacts in their client organizations and work with multidisciplinary groups within the client to share views about the technology and its legal implications. Lawyers can provide crucial guidance to their clients that develop neural devices to build security and privacy protections into the devices and associated services during the design process.

Second, lawyers should help legislators, regulators, and groups like the Uniform Law Commission shape policy in the area of neural devices. We will need the good judgment of lawyers to inform policymakers about how existing laws apply to neural devices, where the law has gaps, and how new laws could address those gaps. At the same time, lawyers play a role in ensuring that any new legislation or regulations will protect the freedoms and protections we cherish.

Finally, lawyers will need to recognize where representing clients concerning neural devices will simply involve applying old law and methods to this new technology, and where new methods will be necessary. For instance, the FDA may require thorough assessments of privacy and security of neural devices during the approval

process, where in the past it may have been more concerned with safety and effectiveness to treat disease and illness. In turn, FDA practitioners will need to become even more familiar with information security and privacy issues than they already are to respond to any FDA privacy and security requirements.

Brain implants and other neural devices may dramatically change the treatment of disease and injuries to the brain and nervous system. Over the long run, they may change humankind and what humans are capable of doing. Ultimately, the law will need to keep pace and lawyers with it. We may be witnessing the birth of a new field of law (public policy and security threats) at the intersection of neuroscience and information technology. We will be excited to see where this new field takes us. The time to consider these issues is now, before the use of these technologies becomes widespread and significant social harm takes place. ♦

Endnotes

1. See <http://emotiv.com/> and www.neurosky.com for further information.
2. For instance, see the video at www.youtube.com/watch?v=6SFaPw5Mw0Y.
3. See Thomas B. Demarse & Karl P. Dockendorf, *Adaptive Flight Control With Living Neuronal Networks on Microelectrode Arrays*, available at <http://neural.bme.ufl.edu/page13/assets/NeuroFlight2.pdf>.
4. See David Brown, *For 1st Woman with Bionic Arm, a New Life Is Within Reach*, WASH. POST, Sept. 14, 2006, available at www.washingtonpost.com/wp-dyn/content/article/2006/09/13/AR2006091302271.html.
5. Andrew Pollack, *Paralyzed Man Uses Thoughts to Move a Cursor*, N.Y. TIMES, July 13, 2006, available at www.nytimes.com/2006/07/13/science/13brain.html.
6. Ray Kurzweil, *The Singularity Is Near* (2006).
7. Lev Grossman, *2045: The Year Man Becomes Immortal*, TIME, Feb. 10, 2011, available at www.time.com/time/magazine/article/0,9171,2048299,00.html.
8. See Tamara Denning et al., *Neurosecurity: Security and Privacy for Neural Devices*, J. OF NEUROSURGERY (July 2009).
9. Computer Fraud and Abuse Act, 18 U.S.C. § 1030(a)(5)(B).

AI Conceptual Risk Analysis Matrix (CRAMSM)

Martin Ciupa¹ and Keith Abney²

¹CTO calvIO Inc., Webster, New York USA

mciupa@calvIOinc.com

²Senior Lecturer, Cal Poly-SLO, California USA

kabney@calpoly.edu

Abstract

AI advances represent a great technological opportunity, but also possible perils. This paper undertakes an ethical and systematic evaluation of those risks in a pragmatic analytical form of questions for designers/implementors of AI-Based systems. We structure this dialog as Conceptual Risk Analysis Matrix (CRAMSM). We look at a topical case example in an actual industrial setting and apply CRAMSM. Conclusions to its efficacy are drawn.

Key Words

AI Risk, Risk Analysis, Dialog Systems.

1. Introduction

A common worry about AI is that it poses an unacceptable risk to humanity (or individual humans) in some way. An extensive literature has begun to emerge about various aspects of Artificial Intelligence (AI) risk, much of it focused on existential risk from Artificial Generic Intelligence (AGI). But AI poses other risks, from how driverless cars solve the ‘trolley problem’, to whether autonomous military robots attack only legitimate targets, to trust in the safety of AI/Robotics in industrial and commercial settings. More generally, the discussion of risks from AI has paid insufficient attention to the nature of risk itself, as well as how decisions about the acceptability of the risks of AI compare to worries about convergent technologies. For example, in military robotics serious concern exists over a possible lack of “meaningful human control” [1]. Missing is a similar concern for autonomous AI-controlled cyberattacks that would lack the very same control [2]. The Vice Chairman of the Joint Chiefs of Staff understands, saying, “In the [Defense] Department, we build machines and we test them until they break. You can’t do that with an artificial intelligence, deep learning piece of software. We’re going to have to figure out how to get the software to tell us what it’s learned” [3]. Such issues apply well beyond the military, and demand an analysis of AI risk that also applies to civilian contexts and to risks that do not rise to the level of human extinction.

So, how best to understand the risks of AI, judge them (un)acceptable, and then apply our insights on risk to determine what policies to pursue?

2. Defining risk, and how to think about it

So, AI poses many different types of risk – but what exactly is risk? Andrew Maynard [4] suggests that we start with the idea of “value.” If innovation is defined as creating value that someone is willing to pay for, then he suggests risk as a *threat to value*, and not just in the ways value is usually thought of when assessing risk, such as health, the environment or financial gain/loss. The possible loss of well-being, environmental sustainability, deeply held beliefs, or even a sense of cultural or personal identity should also count. Risk’s opposite, safety, should be seen as relative, not absolute: safety in all respects is never 100% guaranteed, so as safety is best understood as relative freedom from a threat of harm, so risk is a relative exposure to such a threat.

Extending a schema based on previous work [5], the major factors in determining ‘acceptable risk’ in AI will include (but are not limited to):

2.1 Acceptable-Risk Factor: Consent

Consent: Is the risk voluntarily endured, or not? For instance, secondhand smoke is generally more objectionable than firsthand, because the passive smoker did not consent to the risk, even if the objective risk is smaller. Will those who are at risk from AI reasonably give consent? When would it be appropriate to deploy or use AI without the meaningful consent of those affected? Would non-voluntariness (in which the affected party is unaware of the risk/ cannot consent) be morally different from involuntariness (in which the affected party is aware of the risk and does not consent)? [6]

2.2 Acceptable-Risk Factor: Informed Consent

Even if AIs only have a ‘slave morality’ in which they always follow orders [7], and citizens consent to their use (through, say, political means), that still leaves unanswered whether the risk (of malfunction, unintended consequences, or other error) to *unintended* parties is morally permissible. After all, even if widespread consent is in some sense possible, it is completely unrealistic to believe that all humans affected by AI could give *informed* consent to their use. So, does the morality of consent require adequate knowledge of what is being consented to?

Informed consent: Are those who undergo the risk voluntarily fully aware of the true nature of the risk? Or would such knowledge undermine their efficacy in fulfilling their (risky) roles? Or are there other reasons for preferring ignorance? Thus, will all those at risk from AI know that they are at risk? If not, do those who know have an obligation to inform others of the risks? What about foreseeable but unknown risks—how should they (the ‘known unknowns’) be handled? Could informing people that they are at risk ever be unethical, even akin to terrorism?

2.3 Acceptable-Risk Factor: The Affected Population

Even if consent or informed consent do not appear to be morally required with respect to some AI, we may continue to focus on the affected population as another factor in determining acceptable risk:

Affected population: Who is at risk—is it merely groups that are particularly susceptible or innocent, or those who broadly understand that their role is risky, even if they do not know the particulars of the risk? For example, in military operations civilians and other noncombatants are usually seen as not morally required to endure the same sorts of risks as military personnel, even (or especially) when the risk is involuntary or non-voluntary.

2.4 Acceptable-Risk Factor: Step risk versus State risk

A state risk is the risk of being in a certain state, and the total amount of risk to the system is a direct function of the time spent in the state. Thus, state risk is time-dependent; total risk depends (usually linearly) on the time spent in the state. So, for us living on the surface of the Earth, the risk of death by asteroid strike is a state risk (it increases the longer we’re here).

Step risk, on the other hand, is a discrete risk of taking the next step in some series or undergoing some transition; once the transition is complete, the risk vanishes. In general, step risk is not time-dependent, so the amount of time spent on step matters little (or not at all). [8] Crossing a minefield is usually a step risk – the risk is the same whether you cross it in 1 minute or 10 minutes. For example, the development of AGI poses an existential step risk; but, if there is a ‘fast takeoff,’ any additional state risk of developing AGI may be negligible.

Step risk versus state risk: How shall we determine when state risks are more important than step risks, or vice-versa? If a potential diminishment in a step risk depends on increasing a separate state risk (e.g. slowing down or stopping AGI research that, if successful, would decrease other risks to humanity), how do we decide what to do?

2.5 Acceptable-Risk Factors: Seriousness and Probability

We thereby come to the two most basic facets of risk assessment, seriousness and probability: how bad would the harm be, and how likely is it to happen?

Seriousness: A risk of death or serious physical (or psychological) harm is understandably seen differently than the risk of a scratch or a temporary power failure or slight monetary costs. But the attempt to make serious risks nonexistent may turn out to be prohibitively expensive or otherwise contraindicated. What magnitude of AI risk is acceptable—and to whom: users, nonusers, the environment, or the AI itself?

Probability: This is sometimes conflated with seriousness but is intellectually quite distinct. The seriousness of the risk of a 10-km asteroid hitting Earth is quite high (possible human extinction), but the probability is reassuringly low (though not zero, as perhaps the dinosaurs discovered). What is the probability of harm from AIs? How much certainty can we have in estimating this probability? How do we decide on the probability of serious harm that is acceptable, versus moderate harm or mild harm? If a function, is it linear, asymptotic, or other? Is it continuous or not?

2.6 Acceptable-Risk Factors: Who Determines Acceptable Risk?

In various other social contexts, all of the following have been defended as proper methods for determining that a risk is unacceptable [9]:

Good faith subjective standard: It is up to each individual as to whether an unacceptable risk exists. That would involve questions such as the following: Can the designers or users of AI be trusted to make wise choices about (un)acceptable risk? The idiosyncrasies of human risk aversion may make this standard impossible to defend, as well as the problem of involuntary/ non-voluntary risk borne by nonusers.

The reasonable-person standard: An unacceptable risk is simply what a fair, informed member of a relevant community believes to be an unacceptable risk. Can we substitute a professional code or some other basis for what a ‘reasonable person’ would think for the difficult-to-foresee vagaries of conditions in the rapidly emerging AI field, and the subjective judgment of its practitioners and users? Or what kind of judgment would we expect an autonomous AI to have—would we trust it to accurately determine and act upon the assessed risk? If not, then can AI never be deployed without teleoperators—like military robots, should we always demand a human in the loop? But even a ‘kill switch’ that enabled autonomous operation until a human doing remote surveillance determined something had gone wrong would still leave unsolved the first-generation problem.

Objective standard: An unacceptable risk requires evidence and/or expert testimony as to the reality of (and unacceptability of) the risk. But there remains the first-generation problem: how do we understand that something is an unacceptable risk unless some first generation has already endured and suffered from it? How else could we obtain convincing objective evidence?

2.7 Acceptable-Risk Factors: The Wild Card: Existential Risk?

Plausibly, a requirement for extensive, variegated, realistic, and exhaustive pre-deployment testing of AIs in virtual environments before they are used in actual human interactions could render many AI risks acceptable under the previous criteria. But one AI risk may remain unacceptable even with the most rigorous pre-deployment testing. An existential risk refers to a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential. Existential disasters would end human civilization for all time to come. For utilitarians, existential risks are terribly important: doing what we can to mitigate even a small chance that humanity comes to an end may well be worth almost any cost. And for deontologists, the idea that ‘one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation’ is at least a plausible *prima facie* (and perhaps absolute) duty; such a survival principle appears required for any viable ethics [10]. If there is even a tiny risk that developing AGI would pose an existential risk, this ‘Extinction Principle’ may well imply that we have a duty to stop it.

2.8 Conceptual Risk Analysis Matrix (CRAMSM)

Taking note of items 2.1-2.8 we have formed the following dialog matrix for focusing the dialog of AI Risks in a given use-case.

1/ Acceptable-Risk Factor: Consent
<i>Assess the degree to which consent has been given of the use-case risk</i>
2/ Acceptable-Risk Factor: Informed Consent
<i>Assess the risk of the use-case that parties have potentially not been informed about (or do not understand) the risk potential</i>
3/ Acceptable-Risk Factor: The Affected Population
<i>Assess the risk to the potential Affected Population</i>
4/ Acceptable-Risk Factor: Step risk versus State risk
<i>Assess use case risk in terms of</i> <ol style="list-style-type: none"> 1. <i>State Risk (time likely spent in a state that is a cause of risk)</i> 2. <i>Step Risk (chance of entering into a new risk, as a consequence of step transitions)</i>
5/ Acceptable-Risk Factors: Seriousness and Probability
<i>Assess use case's risk in term of an analysis of potential seriousness and probability of occurrence</i> <ol style="list-style-type: none"> 1. <i>Seriousness: What (if any) serious risks from AIs are acceptable—and to whom: users, nonusers, the environment, or the AI itself?</i> 2. <i>Probability: How much certainty can we have in estimating this probability? What probability of serious harm is acceptable? What probability of moderate harm is acceptable? What probability of mild harm is acceptable?</i>
6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?
<i>Assess use case's risk against standards applicable to it.</i> <ol style="list-style-type: none"> 1. <i>Good faith subjective standard</i> 2. <i>The reasonable-person standard</i> 3. <i>Objective standard</i>
7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?
<i>Assess use case's existential risk that, should it come to pass, might</i> <ol style="list-style-type: none"> 1. <i>annihilate Earth-originating intelligent life,</i> 2. <i>Or permanently or drastically curtail its potential.</i>

3. Specific Case Study, Possible Solution, and Risk Analysis

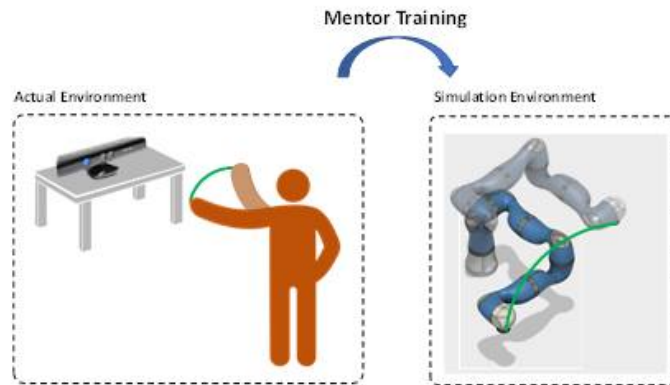
3.1 Specific Case Study

Our case study is applying CRAMSM to the Path Planning of a Robot Arm, in which vials of severe biohazardous materials are to be moved from point A to point B in an optimum path. This path is constrained by parameters such as speed, power-usage, minimization of actuator acceleration and deceleration (that causes wear of the actuators) and collision avoidance. See Fig 1. Keep in mind that cost of production, as well as quality/safety, are value factors to be balanced in this manufacturing example. And the use of AI Robots in this case example is a very real-world example of potential benefit albeit with techno-ethical concerns. The engineering case study has been outlined in Refs [11], [12] and [13].

This path planning problem use case example envisaged here can be described in the following stages of “teaching” the system with a “show & tell” paradigm and AI-based optimization algorithm. It is a process of three stages teaching the system from a basic to advanced level novice to levels of expert proficiency.

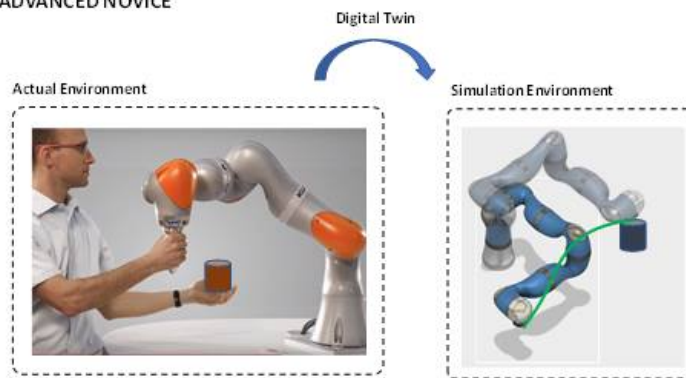
Stage 1 proposes Mentor Training, whereby the motion capture of a human expert is captured into a simulation environment.

NOVICE (AWARE INCOMPETENCE)



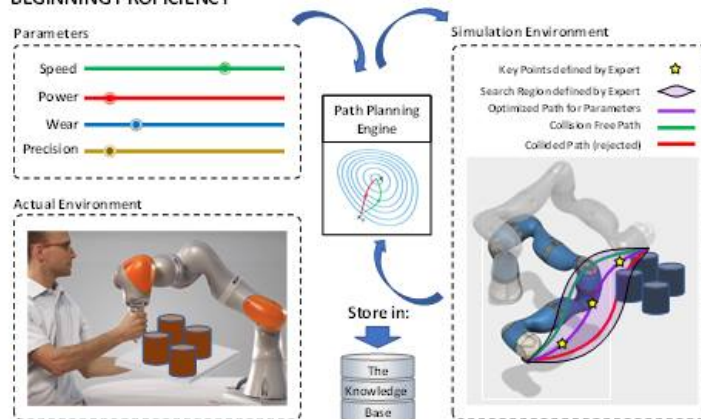
Stage 2 proposes Digital Twin capture of fine tuning of the Robot arm (a Co-robotic arm is envisaged in this example that allows for the physical manipulation of the arm).

ADVANCED NOVICE



Stage 3 proposes an AI-Based fine tuning of the path planning by an optimization process based on value parameters (e.g., Speed of path cycle, Power utilization, Wear and Tear, and Precision of movement) an AI search algorithm might seek an optimum minimization of the path plan against a utility metric of these values.

BEGINNING PROFICIENCY



3.2 CRAMSM applied to the Case Example

<i>1/ Acceptable-Risk Factor: Consent</i>
The use of a robot (in a protective clean room cell) reduces the need for human operator exposure to Biohazards. Any personnel entering the clean room cell should have safety training and contracted consent.
<i>2/ Acceptable-Risk Factor: Informed Consent</i>
However, if the AI directing the robot causes breaches of the clean/safe room (e.g., collisions with the cell walls), then what was thought safe might not be. In this respect personnel in the potential effective area may not be fully informed of the reliability/trust in the system. It is necessary to test any robot behavior in detailed simulation to ensure the path planning algorithms will not likely violate these rules. And personnel potential affected by failures be informed of the extent of the safety testing.
<i>3/ Acceptable-Risk Factor: The Affected Population</i>
The affected population might not be limited to the factory; conceivably, an extended exposure could cause health and safety threats to those outside, or violations of FDA regulations, etc. Again, the system must be validated against the regulations/laws applicable to the domain. The potential affected population should be briefed of the risks.
<i>4/ Acceptable-Risk Factor: Step risk versus State risk</i>
Both state and step risks need to be exposed through the AI testing process and the results passed to stage 5 below.
<i>5/ Acceptable-Risk Factors: Seriousness and Probability</i>
The seriousness of a biohazard breach can be evaluated in principle, but the probability may needs validating in test simulations. Thorough simulation is advocated (to avoid physical exposure) as well as an assessment of the system.
<i>6/ Acceptable-Risk Factors: Who Determines Acceptable Risk?</i>
There are industry bodies that set standards (e.g., GAMP5) as well as government entities that set regulations in this case example (e.g., US FDA).
<i>7/ Acceptable-Risk Factors: The Wild Card: Existential Risk?</i>
If the biohazard agent was severe enough, as might be possible with nuclear materials and/or live chemical/biological agents, then the impacts could be existential, if the AI goes “rogue.” The severity relates properly to steps 3 and 5 above. The risk of ‘going rogue’ is conceivable, in a case of complete AI automation of the industrial facility, and absent proper safeguards against hacking. A solution may involve a software system overriding ethical kernel that ensures “no harm” and sufficient cybersecurity measures. Ultimately a degree of human oversight may be warranted with the system requiring human authorization for critical procedures. Including the ability to switch off the system.

4. Conclusions

We reviewed the concept of AI Risk and picked a real world industrial problem. We proceeded to outline a means of structuring a dialog we refer to as CRAMSM

The AI/Robotics case example (AI-based Path Planning for a Pick and Place application for Biohazardous material) and applied the CRAMSM dialog to it; we think the result is an actual beneficial one for highlighting the AI risk concerns and start the process of handling them objectively.

As such, we believe the resulting techno-philosophy methodology to be a potentially useful early step in the building of tools for conceptualizing and assessing acceptable AI Risk. Further work is needed to develop these concepts, and trial them in real-world applications.

5. References

- [1] UNIDIR: The weaponization of increasingly autonomous technologies: considering how Meaningful Human Control might move the discussion forward. *UNIDIR Resources*, no. 2, 2014. <http://unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf>. Last Referenced 22nd October 2017.
- [2] Roff, H.: Monstermind or the doomsday machine? Autonomous cyberwarfare. *Duck of Minerva*, 13 August 2014. <http://duckofminerva.com/2014/08/monstermind-or-the-doomsday-machine-autonomous-cyberwarfare.html>. Last Referenced 22nd October 2017.
- [3] Clevenger, A.: ‘The Terminator conundrum’: Pentagon weighs ethics of pairing deadly force, AI. *Army Times*, 23 January 2016. <http://www.armytimes.com/story/defense/policy-budget/budget/2016/01/23/terminator-conundrum-pentagon-weighs-ethics-pairing-deadly-force-ai/79205722/>. Last Referenced 22nd October 2017.
- [4] Andrew Maynard, “Thinking innovatively about the risks of tech innovation”. *The Conversation*, January 12, 2016. <https://theconversation.com/thinking-innovatively-about-the-risks-of-tech-innovation-52934>. Last Referenced 22nd October 2017.
- [5] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf. Last Referenced 22nd October 2017.
- [6] Abney, K., Lin, P., and Mehlman, M. “Military Neuroenhancement and Risk Assessment” in James Giordano (ed.), *Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns* (Taylor & Francis Group, 2014)
- [7] Lin, P., Mehlman, M., and Abney, K.: Enhanced warfighters: risk, ethics, and policy. Report funded by the Greenwall Foundation. California Polytechnic State University, San Luis Obispo. 1 January 2013. http://ethics.calpoly.edu/Greenwall_report.pdf. Last Referenced 22nd October 2017.
- [8] Nick Bostrom, *Superintelligence*. (Oxford University Press, 2014)
- [9] Abney, K., Lin, P., and Mehlman, M. “Military Neuroenhancement and Risk Assessment” in James Giordano (ed.), *Neuroscience and Neurotechnology in National Security and Defense: Practical Considerations, Ethical Concerns* (Taylor & Francis Group, 2014)
- [10] Keith Abney, “Robots and Space Ethics,” ch 23 in *Robot Ethics 2.0*, eds. Lin, P., Jenkins, R., and Abney, K. (Oxford University Press, 2017)
- [11] M. Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver - The Case for Really Useful Machine Learning," in *Computer Science & Information Technology (CS & IT)*, 2017.
- [12] M Ciupa, N Tedesco, and M Ghobadi, “Automating Automation: Master Mentoring Process” 5th International Conference on Artificial Intelligence and Applications (AIAP-2018), Jan 2018, Zurich, Switzerland
- [13] M Ciupa and K Abney, “Conceptualizing AI risk” (AIFU-2018), Melbourne, Australia February, 2018.

Martin Ciupa

Martin Ciupa is the CTO of calvIO Inc., a company (associated with the Calvary Robotics group of companies) focused on simplifying the cybernetic interaction between man and machine in the industrial setting. Martin has had a career in both technology, general management and commercial roles at senior levels in North America, Europe and Asia. He has an academic background in Physics and Cybernetics. He has applied AI and Machine learning systems to applications in decision support for Telco, Manufacturing and Financial services sectors and published technical articles in Software, Robotics, AI and related disciplines.



Keith Abney

Keith Abney is senior fellow in the Ethics + Emerging Sciences Group and senior lecturer in the Philosophy Department of California Polytechnic State University - San Luis Obispo. He is co-editor of *Robot Ethics* (MIT Press) and the newly published *Robot Ethics 2.0* (Oxford UP).



3 Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed

Keith Abney

What is robot ethics? The term may cause perplexity; on some ethical views, it seems to be a field of study without an object to study (as some gibe at astrobiology or theology). In the emerging literature devoted to robot ethics, however, the term has at least three distinct meanings, the first two of which clearly refer to something real. First, it can refer to the professional ethics of roboticists (often termed “roboethics” [Verrugio 2007]); second, it can refer to a moral code programmed into the robots themselves—the moral code the robots, not the roboticists, follow; or third (the possibly nonexistent one), “robot ethics” could refer to the self-conscious ability to do ethical reasoning by robots—to a robot’s own, self-chosen moral code. The epilogue of this volume describes these three senses in more detail.

The term “ethics” also needs disambiguation. “Ethics” is sometimes used synonymously with “morality,” but sometimes refers to “the study of morality.” Some robot ethicists, like Rafael Capurro [2009], prefer to distinguish these by

calling the second sense above (a programmed-in robotic moral code) a robot *morality*, whereas only the third sense of a self-conscious, voluntary adoption of a particular code would be called robot *ethics*. Others use “robot ethics” or “machine morality” in both the second and third senses, or even for issues in the discipline philosophers call “metaethics,” and so may leave unclear the meaning of terms like “artificial moral agents” and “machine ethics.” Accordingly, this essay aims to examine some common confusions, misunderstandings, equivocations, and other problems in understanding these three senses of robot ethics, and to introduce the ethical and metaethical issues concerning robots discussed later in this volume.

3.1 Four Questions

I begin with four crucial, but often misunderstood, questions for doing ethics, all of them relevant to robotics:

- (1) What is morality or ethics: the *right*, or the *good*?
- (2) What are moral rights? What is their relationship to moral duties? And who or what can be rights-holders?
- (3) What are the major contemporary moral theories? How do they bear on

robot ethics?

(4) What is a person, in the moral sense? Can a robot be a person?

3.1.1 What is Morality or Ethics? The Right, or the Good?

So what is morality? Morality always involves an “ought (not)” —it is about the way the world ought (or ought not) to be, as opposed to the way it actually is. The “ought” of morality has been understood in two primary ways: as doing the right, or as being good—i.e., the content of morality is understood either as what rules make for right action, or as how one ought to live in order to have a good life. These two approaches are practically equivalent, if living a good life means following some set (the right set) of rules; if not, there is a potential chasm between these two conceptions of morality.

Top-down, rule based approaches, like Asimov’s Three Laws of Robotics, understand ethics as the investigation of right action—what are the rules to follow in order to be morally right, to perform the morally correct (or at least morally permissible) action? The analogy with the legal system is instructive: if one obeys the rules, one is moral; if one disobeys or breaks the rules, one acts immorally. The investigations of ethics are fundamentally, then, an inquiry into what the rules ought to be, for any particular society. Robot ethics then concerns

following (in senses one and three) or programming (sense two) the correct set of rules.

The usual divide within rule-based approaches is between those who say one must intend to obey the rules, no matter what—even if the consequences will be bad (deontologists, associated with Kant), versus those who say the main or only rule is always to make the future consequences as good as possible—*the ends justify the means* (consequentialists, most commonly represented by utilitarians, who tend to measure the ends or results in terms of happiness gained or lost).

There is another historically influential approach that understands ethics as the art/science of living a good life, not as being bound by some set of rules that may not apply to one's unique circumstances. For programming robots, this view represents a “bottom-up” or “hybrid” approach that involves trial and error machine learning of what constitutes (un)acceptable behavior, or a “good” or “bad” robot, that goes beyond mere obedience to a set of rules.

One justification for this understanding of morality as the good, not the right, is the observation that all rule-based approaches have assumed: (a) the rule(s) would amount to a decision procedure for determining what the right action was in any particular case; and (b) the rule(s) would be stated in such terms that any non-virtuous person could understand and apply it (them)

correctly [Hursthouse 2009]. But despite centuries of work by moral philosophers, no (plausible) such set of rules has been found. Moral particularism [Dancy 2004] is one, perhaps unhelpful, purported solution to this quandary: there are no moral rules, only moral facts, and acts can only be judged according to the unique particulars of each case. But if each moral situation is *sui generis*, how then could we ever program robots to be moral?

A more helpful approach for robots is virtue ethics, which asserts the problem with rule-based morality is that it has the wrong object of evaluation. Morality is asserted to be about the character of persons, not the rightness or wrongness of individual acts. Top-down moral theories are concerned with action, and attempt to answer the question, “What should I *do*?” with some set of rules. Virtue ethics, by contrast, attempts to answer the question, “What should I *be*?” Virtue ethics consists not in following moral rules that stipulate right actions, but in striving to be a particular kind of person (or robot)—a virtuous one.

As such, virtue ethicists usually deny that mere actions are meaningfully good or evil—it may be morally wrong (betray a defective character, a “vice”) for me to begin to carve your chest with a knife, but someone else performing exactly the same action in the same circumstances may be perfectly moral (“virtuous”)—if you are lying on an operating table, and she is a surgeon,

whereas I am not! She evinces a perfectly virtuous character in cutting you open because of her skills and her role in the situation; because my skills and my role are different—because I am not a licensed surgeon—performing the same act would reveal my character is flawed, even if my intentions were good; indeed, even if (miraculously) the surgery turned out well—that is, even if the consequences of my act were good. For robots, this same proper functioning approach to evaluation appears natural: is the surgical robot operating properly in carving one's chest, or is my new robotic bandsaw dysfunctionally attempting to do the same thing?

Virtue ethicists thus claim what counts is one's moral character—moral evaluation is of persons, not of actions. The virtues are understood as dispositions to act in a certain way (would-be habits); ideally, to know by practical wisdom the right thing to do, in the right way, at the right time. Context sensitivity means virtues do not act as rules and may conflict; in a difficult situation, one should not ask what abstract rule to follow, but instead should ask: what would a role model do in my situation? Or—if I do X, would it start a bad habit? Will I become dysfunctional in my proper role(s)?

The implications of this divide for robot ethics (in all three senses) are potentially profound. For the first sense, is roboethics simply the search for a list of rules that any and all roboticists must follow in their work, such that all who

adhere to the rules are automatically moral, and those who break them automatically immoral? Or is it perhaps the search for the rules that will produce the best future net consequences for society (rule-utilitarianism)?

Or, following the second approach, should roboethics search instead for distinctive principles that roboticists of good character evince in their work (i.e., virtues of doing robotics), as well as character traits that lead to dysfunction in their work (i.e., vices of doing robotics)? For a roboticist, a claim that *"I'm not responsible because I followed the rules"* would be indefensible from a virtue ethics perspective. Instead, one should emulate a role model of professionalism. One example would be "The Roboticist's Oath" [McCauley 2007], understood as a statement of principles that any professional roboticist should evince. Bill Joy also asserted the need for such an oath as a means of setting up a professional exemplar and standards; he wrote: "scientists and engineers [need to] adopt a strong code of ethical conduct, resembling the Hippocratic oath" [Joy 2000]. Further, if robots themselves are proper objects of moral assessment, then robot virtue ethics would become the search for the virtues a good (properly functioning) robot would evince, given its appropriate roles.

So, is ethics the study of the right, or the good? Despite the arguments above for ethics as the study of the good, the case for the rule-based approach has practical import in another social tendency: to equate moral and legal,

immoral and illegal—that is, to construe any action that avoids legal sanction as morally permissible, and to insist on redress (in the form of legal rights) when such laws have been broken by others, or to insist such actions were permissible when others wish to cast moral blame, by saying “but I had a right!”

The relationship between virtues and rights begins with an observation: when all parties in a given social context are acting virtuously, no one mentions their rights; in fact, such appeals would appear unseemly when no vices exist. Rights claims inevitably arise *only* when something has gone amiss. That is, appeals to rights inevitably occur only when moral conflict already exists, and rights-based approaches based on rules/laws are always an attempt to fix something that is already broken—or to prevent it from getting worse. And rules invariably have unintended consequences, as the attitude that “whatever is within the rules is permissible” leads to the unscrupulous finding malicious means to bend the rules to their advantage, without (quite) breaking them. So, in a moral utopia, there would be no need for moral rights. And many moral theorists, running the gamut from utilitarians, like Bentham, to virtue ethicists, like MacIntyre, to various existentialists, have denied their existence.

But despite such views, rights-claims may be a necessary feature of the ethics of any large, complex society. When groups are relatively small, with common social mores reinforced by shared moral education and acceptance of

one's proper roles, the virtues may be largely taken for granted and enforced by purely social sanctions—as the opprobrium of those with whom one has substantial relationships is a powerful tool for enforcing social moral consensus. Our behavior is usually far more affected by the (dis)approval of those around us than by an abstract, remote threat of law enforcement, in “ordinary” contexts.

For roboethics, moral education (in the virtues of the profession) and other social means of enforcing shared mores (such as causing a bad reputation, or denying conference participation, publication, grants, tenure, or even employment for those who violate shared virtues) may be effective, at least for a while. But as the group of those dealing with robots becomes larger and more variegated, social sanctions and shared virtues gradually become less effective at minimizing harm.

At such a point, outside regulation and institutions, with clear procedures, rights, and duties, usually become necessary in order to keep the smaller group's practices acceptable within the larger society. So, although rights-claims may be a “second-best” form of morality, appealed to only when immorality is already rampant or at least expected; nonetheless, in the real world, in which vices are all too common, they may remain a necessary evil. Accordingly, I next attempt to clarify the concept of a moral right, whether for humans or for robots.

3.1.2 What are Moral Rights? What Constitutes their Relationship to Moral Duties? And Who or What can be Rights-Holders?

There are two main competing theories of rights—the “will” theory and the “interest” (or “welfare”) theory [Wenar 2010]. The “interest theory” maintains that rights correlate with interests (or welfare)—everything that has interests (or a “welfare”) has rights. All persons have a duty to respect the rights of everything that has interests (including, potentially, robots?). But the “will theory” of rights disagrees: it asserts the right to liberty is the foundation of all other rights claims, and a rights claim is understood as the entitlement to a particular kind of choice—a rights claim entitles me to claim or perform something, or not—*it is up to me* (and nobody else). A rights claim entails no duty upon the rights holder, but only a freedom—to perform/ claim something, or not. But the correlativity thesis makes clear that rights-claims do entail duties, not for the rights-holder, but for all other persons—if I have a right, then you (and everyone else) have a correlative duty.

The correlativity thesis is essential to rights theory, in conceptualizing the relationship between rights and duties. It has a slogan form: “no rights without

responsibilities”—rights do not exist unless others have duties. Rights are guaranteed freedoms, which then guarantee duties on everyone else.

But this has an additional implication, relevant here—who is “everyone else?” In this context, it refers to moral agents, beings capable of moral responsibility. It makes no sense to claim that trees or dogs or the environment have a moral responsibility to respect my freedom of speech; given “ought implies can,” they are incapable of it. If a tree falls on my head and silences me, we cannot hold it morally responsible! So, “no rights without responsibilities” carries an additional implication: on the “will theory,” only morally responsible agents can have moral rights. If I am incapable of agency, of the exercise of liberty, of rational free will, then I am incapable of being a rights-holder. If there were no moral agents, there would be no moral rights—because there are no rights without responsibilities.

But then, on the will theory, anyone and anything incapable of being held responsible for their actions would thereby have no moral rights. This would explain why current robots have no rights, but its implications cause unease for many, not least because much reasoning in applied ethics takes the following form: first, assess all the rights claims in a situation; if no rights have been violated, then an action is morally permissible. So if moral agents are the only rights-holders, then on such reasoning, agents appear morally free to act

however they wish towards non-agents—so torturing pets or destroying robots is ok?

Such reasoning usually commits the fallacy of assuming a statement and its converse are equivalent—in particular, the correlativity thesis and its converse. And it mistakes the true nature of the relationship between rights and duties. The correlativity thesis: if I have a right, then all other agents have a correlative duty. The converse correlativity thesis: if I have a duty, then someone else has a correlative right. Upon a moment's reflection, the latter is absurd. Suppose I have a moral duty to give some of my disposable income to charity; which charity thereby has a right to my donation? The correct answer is: none. Some charity will receive my donation, but none of them were entitled to it—no one has a right to my charity, although I have a duty to give it.

Despite the prominence of rights claims in much applied ethics, the failure of the converse correlativity thesis means that we all have duties that correspond to no rights at all; and the impulse that supported the “interest theory” of rights disappears. Many non-agents (such as animals or the environment) have no rights, because they are not moral agents. But they plausibly are *moral patients*, to whom we agents owe duties; this possibility becomes clear once we realize we have many duties that correlate to no specific right. We merely equivocate when we call those duties “rights,” as the interest

theory does. Hence, we can safely say that, for the foreseeable future, robots will have no rights—at least until robot ethics approaches the third sense above, of robots as fully autonomous moral agents. But that realization leaves unresolved our moral duties concerning senses one and two—how roboticists ought to behave, and what moral code roboticists should install in their creations.

So, in robot ethics, we should not reason that if no rights have been violated, then an action is automatically morally permissible—because every moral duty cannot correspond to a discrete, identifiable right. We need a more encompassing moral approach than mere rights theory in order to fully discuss our moral duties in at least senses one and two of robot ethics. What other ethical theories are widely considered plausible candidates to specify our duties?

3.1.3 What are the Major Contemporary Moral Theories? How do they Bear on Robot Ethics?

We already discussed virtue ethics in section 3.1.1 above, as one major moral theory based on the good. Let us now turn to two more influential "top-down" rule-based approaches that can be applied to robot ethics—deontological and consequentialist theories.

Deontological (duty-based) approaches to robot ethics would simply see roboticists (sense one) or the robots themselves (sense two) acting in accord with some finite set of (presumably algorithmic, programmable) rules, and moral decision-making would thus consist simply in computing the proper outcome of the (programmable) rules, in accordance with a monotonic first-order logic. There are concerns that such a basic logic could not capture ethical insights, but work on deontic logics that would have programmable rules is well advanced [e.g., Arkin 2009, Bringsjord and Taylor, this volume]. Hence, deontological approaches that see ethics as merely a set of (programmable) rules to follow are, in principle, a natural approach to creating sense two of an ethical robot, and making sure it conforms to any (programmable) set of ethical standards.

Asimov's Three Laws and Kant's Categorical Imperative (CI) are influential examples of such an approach in robot ethics; Kant's [1785] theory has two primary formulations:

CI(1)—or the formula of universal law (FUL): "Act only in accordance with that maxim through which you can at the same time will that it become a universal law."

A maxim is a (true) statement of one's intent or rationale: why one did what was done. So, Kant asserts that the only intentions that are moral are those that could be universally held; partiality has no place in moral thought. Kant also asserts that when we treat other people as a mere means to our ends, such action must be immoral; after all, we ourselves don't wish to be treated that way. Hence, when applying the CI in any social interaction, Kant provides a second formulation as a purported corollary:

CI(2)—or the Means-Ends Principle: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means."

One could never universalize the treatment of another as a mere means to some other ends, claims Kant, in his explanation that CI(2) directly follows from CI(1). This formulation is credited with introducing the idea of "respect" for persons; that is, respect for whatever collective attributes are required for human dignity, as ends in ourselves.

A Kantian deontologist thus believes that acts such as stealing and lying are always immoral, because the intent to universalize them creates a paradox.

For instance, one cannot universalize stealing property (that which is rightfully owned) without undermining the very concept of property. Kant's approach is widely influential, but has problems of applicability and disregard for consequences; e.g., a robot that could never lie would certainly not be an asset if the enemy captured it.

Further, CI(1) is too permissive, and potentially permits horrors by allowing any action that can have a universalizable maxim; this can also cause a conflict with CI(2). For instance, CI(1) might sanction voluntary slavery, a topic discussed by Petersen (this volume) for robots. Worse yet for programming deontological ethics into robots, using CI(1) could produce a *conflict of duties*—when two maxims both appear universalizable on their own, but come into conflict jointly.

Next, CI(2) is too stringent—interpreted literally, it forbids all war, or any other action in which I affect someone without their consent (and thereby treat them as a “mere means”). This would render most human-robot interaction, most especially military action, impossible. Not only do enemy civilians (as “collateral damage”) not give consent to being harmed as a means to victory, there are also innumerable other human activities in which a minority who object are nonetheless treated as a means for the good of the majority—or do you consent to everything that the government does? In practice, this creates a

reductio ad absurdum of this deontological constraint. To accomplish much of anything, a robot will sometimes have to engage in actions that affect humans without their consent; the key is for it to make the correct decisions about how, when, and why that should be.

Finally, differences in *roles and capacities* problematize universalization—so a robot may be able to universalize “never shoot children” on a normal battlefield, but if insurgents become aware of this, child soldiers could wreak havoc as the robot stands passively by. Or, the laws of war deem it appropriate to target enemy soldiers with a gun pointed at you—but not if they are severely wounded and incapable of firing. Would a robot be able to discriminate the degree of wounding and retaliatory (in)capacity, and do the right thing?

Another deontological approach that has engendered much discussion in robot ethics is Asimov’s *Three Laws of Robotics* [1942] (he later added a fourth, “Zeroth Law”); they are as follows: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey orders given to it by human beings, except where such orders would conflict with the First Law; (3); a robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The laws are prioritized to minimize conflicts. Thus, doing no harm to humans takes precedence over obeying a human, and obeying trumps self-

preservation. However, in story after story, Asimov demonstrated that three simple hierarchically-arranged rules could lead to deadlocks when, for example, the robot received conflicting instructions from two people, or when protecting one person might cause harm to others. It became clear that the first law was incomplete, as stated, due to the problem of ignorance: a robot was fully capable of harming a human being as long as it did not know that its actions would result in (a risk of) harm, i.e., the harm was unintended. For example, a robot, in response to a request for water, could serve water teeming with parasites, or drown a human in a pool, or crush someone with ice, *ad infinitum*, as long as the robot was unaware of the risk of harm.

One attempted solution is to rewrite the first and subsequent laws with an explicit knowledge-qualifier: “A robot may do nothing that, to its knowledge, will harm a human being; nor, through inaction, knowingly allow a human being to come to harm” [Asimov 1957]. But the cleverly immoral could divide a task among multiple robots, so that no one robot could know that its actions would lead to harm; suppose one disposal robot places nuclear medical waste in a package, another places a wire, another attaches the timer, and so on until the “dirty bomb” detonates. Of course, this simply illustrates the problem with deontological, top-down approaches, that one may follow the rules perfectly but still produce terrible consequences.

An additional difficulty is determining the degree of acceptable risk. The “through inaction” clause of Asimov’s first law apparently implies a robot would have to constantly intervene to minimize all sorts of risks to humans, possibly rendering it incapable of performing its primary mission. A modified First Law attempts a fix: (1′) A robot may not harm a human being.

But removing the First Law’s “inaction” clause solves one problem only to create a greater one: a robot could initiate an action that would harm a human. For example, suppose a military robot initiates an automatic firing sequence, then watches a noncombatant wander into the firing line. The robot knows it is capable of preventing the harm (by ceasing the automatic firing), but it may nevertheless fail to do so, since it is now not strictly required to act.

Asimov later added a Zeroth Law [1985]—so named to continue the pattern of lower-numbered laws superseding in importance the higher-numbered laws—so that the Zeroth Law had highest priority and must not be broken: (0) a robot may not harm all humanity or, through inaction, allow humanity to come to harm. This would allow a robot to harm individual humans, if so doing prevented an “existential threat” to all humanity. But how could a robot determine when such a threat exists (or how serious it is), so that harming individual humans to prevent it is permitted? Would the Zeroth Law permit robots to force human guinea pigs into medical experiments, to create a vaccine

against a virus that *might* cause a pandemic? How strong is this version of the “precautionary principle?”

Such problems raise a central criticism of all deontological approaches—they fail to take the likely consequences into account. So, consequentialist ethics explicitly addresses this; utilitarianism—the primary consequentialist theory—proposes the goal of morality is to maximize utility, and utility is defined as the sum of the good consequences of an action, minus the sum of the bad consequences of the act. The work of Jeremy Bentham and JS Mill stands as the *locus classicus* of utilitarianism; their view asserted a single rule of right action, the *Greatest Happiness Principle (GHP)*: one ought always to act so as to maximize the greatest amount of net happiness (utility) for the largest number of people.

Like the deontologists, classical utilitarians emphasized *egalitarianism* (everyone’s happiness counts equally), *impartiality* (I care no more for my happiness than for yours, in deciding what’s right), and *universal scope*—so the moral rightness of an act depends on the consequences for all people (as opposed to only the individual agent, present people, or any other limited group).

However, this approach fails to be computationally tractable. So, the *calculational* objection: it is an impossible demand to calculate the utility of

every alternative course of action; thus, utilitarianism makes moral evaluation impossible, as even the short-term consequences of most actions are impossible to accurately forecast and weigh, much less the long-term consequences. One response to this objection is *cost-benefit analysis*: translate good and bad consequences into economic value (benefits and costs), and then calculate which outcome maximizes expected profit/utility. Ethics becomes a branch of economics. But there are serious reasons to believe that moral values cannot systematically be reduced to economic values—for instance, the claim that the values of love, devotion, and honor do not have a price. The ethicist Mark Sagoff [1982] claims it betrays a fundamental moral confusion to conflate our *economic* values as consumers with our *moral* values as citizens—and the attempt to place a price on everything important is morally debilitating.

Can robots, with their potentially enormous computing power, solve this calculational problem? Unlikely—even if Sagoff is wrong. For robots, the calculational difficulties include how utility is represented within a computational system, how long-run the consequences are to be computed, how much data must be input, and scope—whose consequences (welfare) should be included in the calculation. Given limitations of available information and the sheer multitude of variables needed for any plausible decision-making, such a calculation poses a tremendous computation load on even the fastest

systems. A utilitarian robot may either fail to determine which course of action is most acceptable within the time allotted, or use grossly insufficient information in order to shoehorn its calculations into the time available. But if utility is (in practice) incalculable, and one's obligation is to maximize utility, what is left of utilitarianism?

Even if the calculational problem is solvable, there are other objections to utilitarianism: e.g., the justice or *scapegoating* objection would point out that maximizing utility may demand injustice, such as executing an innocent person to prevent a riot that would have resulted in deaths and economic damage. This is to say that utilitarianism, at least in its basic form, cannot readily account for the notion of rights and duties nor moral distinctions between, e.g., killing versus letting die, or intended versus merely foreseen deaths, or other harms (assuming we think such notions and distinctions exist).

Whether deontological or utilitarian, for robots there is an additional, fatal flaw in each of the top-down theories, connected to the calculational objection: they all suffer from a version of the *frame problem*—that is, knowing what information is (ir)relevant to moral decision-making. In order to decide anything, does a robot have to know everything? How can a robot be sure to take into account all the information that is relevant to moral decisions

(especially in novel situations), without being swamped by considering terabytes of irrelevancies?

The frame problem reinforces the worry that top-down theories require an impossible computational load for robot decision-making, due to the requirements for representing knowledge of the relevant effects of action in the world, the difficulty of estimating the sufficiency of the initial information, and knowledge about the psychology of agents and their causal consequences. Human agents also have such problems, but at least sometimes appear able to apply rough and ready top-down evaluations in their selection of courses of action. Evolutionary psychologists such as Tooby and Cosmides [1997] suggest that human minds do so by having special-purpose modules, rather than by being general computing machines. So, perhaps, limited-domain robotic systems might solve the frame problem, too—particularly if the goal is not to create a perfect system, but only one that makes as good (or better) decisions than humans do, in specific contexts.

Even so, would such robots be moral persons? For Kantians, only fully autonomous agents—rational beings who can self-consciously choose their own life goals, rather than serving as a mere means to the ends of others—can be full moral persons. So, can robots become fully autonomous moral agents? And should they? That is, if it is possible, should (human) moral agents build robotic

moral agents? Or should humanity retain full agency only for itself? In short, can (and should) robots become persons?

3.1.4 What is a Person, in the Moral Sense? Does it Require an Emotional “Inner” Life?

Some theorists claim that robots cannot become fully-fledged moral persons until (and unless) they can have an inner moral sense, with a full emotional “inner” life. Perhaps robots will one day have emotions; but our legal system assumes that moral agency does *not* require a normal, properly functioning emotional “inner” life. Psychopaths/sociopaths, rational agents with dysfunctional or missing emotional affect, are still morally and legally responsible for their crimes; whereas those who have emotional responses, but cannot exercise rational control (like the severely mentally disabled or infants) are not. But psychopaths, while emotionally dysfunctional, plausibly still have emotions. Would an emotionless robot possibly be a person?

The existence of two types of decision-making systems in human psychology may help explain some of the confusion over this claim in the history of ethics. Numerous philosophers have defended theories of the moral sentiments, or emotivism—the claim that ethics is ultimately nothing but an

expression of our emotional attitudes—despite the clear uniqueness of ethics in our species, and the clear sharing of emotions with other species. Such views, in addition to being unable to explain why nonhuman animals lack morality, also have struggled to explain the apparent cognitive meaningfulness of ethical claims and especially ethical disagreement. (They also naturally have severe difficulties accounting for the ethics of emotionless robots.)

A better ethics involves the proper understanding of the implications of evolution for morality. Even primate researcher Frans de Waal [2010] writes: “I am reluctant to call a chimpanzee a ‘moral being.’ This is because sentiments do not suffice. ... This is what sets human morality apart: a move towards universal standards combined with an elaborate system of justification, monitoring, and punishment.” So why are humans uniquely (for now, anyway) moral beings? Evolutionary psychologists [Marcus 2008] claim there are not one but two types of decision-making systems within most humans—an instinctual, emotionally laden system which serves as the default for much of human activity, particularly when stressed or under pressure. Many other animals share this noncognitive decision-making system, in which (quite literally) we “know not what we do”—or quite why we do it. Research by Libet [1985] indicates that this system can have, e.g., our arm begin to move *before* we are conscious of deciding to do so! But this “ghost in the machine” does not exhaust human agency; Libet and others

found we also have a “veto” ability that can, after its subconscious initiation, still alter our action, in accord with a conscious decision. So, the uniqueness of current humans lies in a second, cognitive, decision-making system, called the “deliberative system,” which can also cause us to act due to deliberative agency.

In humans, this deliberative system overlays the ancestral instinctual, emotional (and faster) decision-making system, and so reason is quite often trumped by our instinctual drives; all too often, I “instinctively” do what I (upon reflection, using the slower deliberative system) later regret. We humans stereotype, harbor irrational prejudices, exhibit superstitious behavior—all the unconscious work of our emotionally laden ancestral system. (We, too often, also use our deliberative system to rationalize or “justify” such biases after the fact.) We also know that many other Earthly animals share such an ancestral, emotional system—indeed, it is sometimes called the “reptilian brain”—but lack the deliberative, and we realize they lack morality. That is, we do not hold them morally responsible for what they do. They are not moral persons.

The deliberative system involves our ability to structure alternative possible futures as mental representations, and then to choose our actions based on which of the representations that we wish to become our experienced reality. In other words, the deliberative system incorporates moral agency. Without it, morality simply cannot exist; your dog makes decisions about

urinating on the carpet, but it cannot fully understand and cogitate upon those decisions, and decide in a rational manner. It uses the “emotional” ancestral system, because it has no fully developed deliberative system. That is why it makes no sense to hold dogs morally responsible for their actions, or to have them incur moral or legal guilt for their trespasses. Likewise for human non-agents—babies and the severely cognitively disabled simply do not *know* what they are doing, albeit they constantly make decisions. And neither common morality nor the legal system thus holds them responsible for their actions, whatever their consequences.

3.2 The Requirements of Moral Personhood: Robots and their Implications

Hence, a deliberative system capable of agency appears necessary for the existence of morality, and so for moral personhood. But is the ancestral emotional system needed as well? What of hypothetical creatures that could rationally deliberate, yet lack emotions? Would they have morality? In other words—could (emotionless) robots be moral persons?

Yes, they could. And realizing this problematizes all systems of non-cognitive ethics, whether based merely upon the “moral sentiments,” or any other basis that takes our ancestral, emotion, and instinct-laden systems as

crucial to ethics. As argued, that flies directly in the face of our moral practice, in which we only hold those beings with fully functioning deliberative systems morally responsible for their actions, and take defects or temporary breakdowns or lulls in that deliberative system to be morally exculpatory. My cat is not put on trial for arson when it knocks over a candle and burns down the house—nor is a baby, or someone asleep in the midst of a nightmare. But we could imagine an intelligent alien without emotions (Mr. Spock?) who would be held responsible. Or—a robot, who deliberately did the same thing.

And so the key to moral responsibility and personhood is the possession of moral agency, which requires the capacity for rational deliberation—but not the capacity for functional emotional states, per psychopaths—and so robots may well qualify. The essays by Petersen, Sparrow, and the epilogue (this volume) examine some of the implications of artificial personhood.

But what of freedom? Another objection to robotic morality and personhood is not their lack of emotions, but rather, their presumed lack of a free will—of the freedom to do otherwise, that is required for the proper assignation of moral responsibility. A robot, it is argued, must follow a deterministic algorithm—its computer program. Even if it appears to be making a choice, that is but an illusion borne of our ignorance of the underlying program, or the external input, that together determine the robot's every

behavior. A robot cannot do other than as it is programmed to do. Unlike (it is supposed) rational human agents, the robot has no free will—so while it may have the reasoning capacity required for morality, it lacks the freedom required to be a true moral agent.

Well, perhaps. First, it is not clear that humans actually have the type of freedom the argument alleges is required for morality (as Lokhorst and van den Hoven argue, this volume); debates on free will between compatibilists and libertarians have simmered for centuries. And even if humans do have such libertarian freedom, is it really true that robots cannot? The answer might plausibly be no—robots could have libertarian freedom, if anything can.

The short version of this speculative argument goes as follows: first, the “hard problem” of consciousness, according to David Chalmers, is subjectivity, or subjective experience—i.e., there is something it is like to be me—and all current explanations of information processing leave that unexplained. Chalmers [1995] writes: “perhaps the most popular ‘extra ingredient’ of all is quantum mechanics [e.g. Hameroff 1994]. The attractiveness of quantum theories of consciousness may stem from a Law of Minimization of Mystery: consciousness is mysterious and quantum mechanics is mysterious, so maybe the two mysteries have a common source.”

Second, consider David Deutsch's [1997] argument for reality of parallel universes given the reality of quantum computing. Deutsch notes we have already built quantum computers, and computation always requires a substrate—something on which to compute. But quantum computers are nonlocal—they cannot have a causally closed substrate in four-dimensional space-time. Hence, on Deutsch's view, they can only sensibly be said to be computing across multiple parallel four-dimensional space-times—i.e., “parallel universes.”

So quantum computing—which is already being done—proves the existence of parallel universes, Deutsch asserts. He interprets these multiple universes via Hugh Everett's “Many Worlds Interpretation” of quantum mechanics: every possible probability distribution is actualized in a separate universe, so there's a universe in which you read this essay to the end, another in which you quit reading now, another in which you ceased existing 5 seconds ago, another... and so on. And all are equally real; but you are only aware of this one, because the information carried by the rest of the quantum wave(s) is now invisible to you—the act of observation guarantees it is in another universe.

Now, return to the problem of rational free will/agency—the problem is, what is it? Our commonsense conception of it appears incompatible with determinism (despite the valiant efforts of compatibilists): to have freedom, it

cannot be the case that one could not do otherwise. To be an agent is to have at least two logically, physically possible futures open to me right now: one in which I choose to do X, and one in which I do not.

But our understanding of agency is also incompatible with causal indeterminism—uncaused events are simply not the same as an act due to agency. If my hand begins flopping around for no apparent reason, I do not believe that proves my agency—instead, it makes me call the doctor. To be an agent, I must be in rational control of which of those possible futures comes into existence. There are (at least) two possible futures, and “it is up to me” (not randomness) which occurs.

Thus, commonsense (libertarian) agency seems to be a causal power, but not one that is determined by antecedent events. So agency, in conception, is a nonphysical causal power in addition to the typical physical causal nexus. But what exactly is this mysterious causal power? Does it really exist, or is libertarian agency merely a massive, species-wide delusion, borne of our ignorance of the fine-scale causal structure of our brains and bodies and the world?

Recall Chalmers’s Law of Minimization of Mystery: consciousness is mysterious and quantum mechanics is mysterious, so perhaps the two mysteries have a common source. Perhaps the collapse of the wave function in QM, as several interpretations insist, is associated with the consciousness of a physical

state. Perhaps the solution of the collapse of the wave function has to do with mind/agency?

Suppose the following: agency consists in the rational examination of (deliberation upon) nearby possible worlds/parallel universes, and then deciding between them in terms of which one to bring about as an object of subjective experience. To make sense of this, agents would need a mental causal power of accessing and deciding between parallel universes, to determine which one your self-consciousness inhabits after one's choice. Some such account could make sense of why there is no causal closure of the (4D) physical, but nonetheless there is always causal closure when agency is included.

So, on this hypothesis, libertarian agency is an ability to access and decide between various possible worlds, understood as parallel universes, in order to single out one to experience. Is this additional causal power to access parallel universes only possible for biology (as emergentist approaches to agency like Searle's seem to imply)? The implication of Deutsch's argument is: no, computers already do it. So, if libertarian agency is possible in this way, then robots with libertarian agency are possible, if they can do quantum computing. Such quantum computing would be needed to move from simulated agency to real agency.

In summary, without attempting here to clearly argue for the truth of either compatibilism or libertarianism, let me finally indicate why it is unlikely to make a difference to robot ethics: if compatibilism is true, then the kind of freedom humans have—a freedom compatible with deterministic physical processes—seems obviously possible for robots. If libertarianism is true and intelligible, the quantum computing argument claims that the necessary and sufficient conditions for human libertarian freedom could also be met by robots. So, no matter which type of freedom you believe is required for morality, we have good reason to think that robots could have it, too.

3.3 Conclusion: On Robots and Ethics, and Combining the Two

If I am right, one day robots could become moral agents, and, so, full moral persons. It seems possible that cyborgization will render the issue moot, by gradually merging biological and mechanical persons until no one seriously doubts that robots are fully fledged persons, as former biologicals retain their personal identity while gradually gaining an ever-increasing mechanical body [e.g. Warwick, this volume, and the epilogue]. Assuming robot personhood is possible, humans will eventually have a momentous decision to make: will we enlarge the moral community to include our fellow (artificial) persons, or will we

deny robots the right to become our newest kind of children—ones born, not biologically, but through manufacturing techniques? Their robotic nature and ethics, previously selected by designers (not by natural selection) to serve humans, would then become their own choice. Robots would be “emancipated.”

But for the foreseeable future, robotic morality will necessarily involve the ethics of humans creating robots to follow rules or evince a good character, and not the rules or character robots choose for themselves. Near-term robots will require moral character/rules that are programmable or machine-learnable, and hence not dependent solely on incalculable, uncontrollable consequences or on emotions or moral sentiments. As such, deontology and virtue ethics appear the only plausible candidates for robot morality among the major ethical approaches, and the problems of a strict deontological approach to programming ethics are detailed elsewhere in this text, not least in considering the “frame problem.”

Hence, although deontological approaches involving categorical, universal rights and duties may be possible, I believe that hypothetical imperatives (within a deliberately restricted, not universal, frame) coming from virtue ethics appear the best bet for near-term robotic morals (in sense two). The emphasis on being able to perform excellently in a particular role, and the corresponding specificity of the hypothetical imperatives of virtue ethics to the

programming goals, restricted contexts, and learning capabilities of non-Kantian autonomous robots, makes virtue ethics a natural choice as the best approach to robot ethics—at a minimum, until and unless robots ever acquire something approaching full autonomy in sense three, choosing their own life goals. If and when that happens, robots will do ethics (in the third sense) alongside us—or replace us biologically-instantiated ethicists!

References

Arkin, R.C., 2009. *Governing Lethal Behavior in Autonomous Systems*, Chapman and Hall Imprint, Taylor and Francis Group.

Asimov, Isaac. 1942. "Runaround," *Astounding Science Fiction*, March.

Asimov, Isaac. 1957. *The Naked Sun*. Garden City, NY: Doubleday & Company.

Asimov, Isaac. 1985. *Robots and Empire*. Garden City, NY: Doubleday & Company.

Capurro, Rafael. 2009. "Ethics and Robotics," in *Ethics and Robotics*, Capurro and Nagemborg (Eds.), AKA IOS Press.

Chalmers, David, 1995. "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2(3): 200-219.

Dancy, J., 2004. *Ethics without Principles*, Oxford: Clarendon Press.

Deutsch, David, 1997. *The Fabric of Reality*, New York: Viking Adult.

de Waal, Frans. 2010. "Morals Without God", *New York Times*, October 17, 2010.

Available at <http://opinionator.blogs.nytimes.com/2010/10/17/morals-without-god/?scp=1&sq=Frans%20de%20Waal%20&st=cse>. Accessed November 18, 2010.

Hursthouse, Rosalind, 2009. "Virtue Ethics," *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition)*, Edward N. Zalta (ed.). Available at <http://plato.stanford.edu/archives/spr2009/entries/ethics-virtue>. Accessed November 18, 2010.

Joy, Bill, 2000. "Why the Future Doesn't Need Us", *Wired* 8.04: 238-262.

Kant, Immanuel. 1785. *Groundwork of the Metaphysic of Morals* (several editions).

Libet, B., 1985. "Unconscious cerebral initiative and the role of conscious will in voluntary action", *Behavioral and Brain Sciences*, 8:529-566.

Marcus, Gary, 2008. *Kluge*, New York: Houghton Mifflin.

McCauley, Lee, 2007. "AI Armageddon and the Three Laws of Robotics," *Ethics and Information Technology* 9:153–164.

Sagoff, Mark, 1982. "At the Shrine of Our Lady of Fatima or Why Political Questions Are Not All Economic," *Arizona Law Review* 23: 1281-1298.

Veruggio, Gianmarco, 2007. "The EURON Roboethics Roadmap", European Robotics Research Network, Atelier on Roboethics, 2005-2007. Available at <http://www.roboethics.org>. Accessed November 18, 2010.

Wenar, Leif, 2010. "Rights," *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, Edward N. Zalta (ed.). Available at <http://plato.stanford.edu/archives/fall2010/entries/rights/>. Accessed November 18, 2010.



Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Robot ethics: Mapping the issues for a mechanized world

Patrick Lin^{a,*}, Keith Abney^{b,c,2}, George Bekey^{a,3}^a California Polytechnic State University, Philosophy Department, 1 Grand Avenue, San Luis Obispo, CA 93407, USA^b California Polytechnic State University, College of Engineering, 1 Grand Avenue, San Luis Obispo, CA 93407, USA^c University of Southern California, Viterbi School of Engineering, Los Angeles, CA 90089-0781, USA

ARTICLE INFO

Article history:

Received 13 September 2010

Accepted 28 November 2010

Available online xxxx

Keywords:

Robot

Robotics

Ethics

Society

Philosophy

Psychology

Law

Policy

Safety

Error

ABSTRACT

As with other emerging technologies, advanced robotics brings with it new ethical and policy challenges. This paper will describe the flourishing role of robots in society—from security to sex—and survey the numerous ethical and social issues, which we locate in three broad categories: safety & errors, law & ethics, and social impact. We discuss many of these issues in greater detail in our forthcoming edited volume on robot ethics from MIT Press.

© 2011 Elsevier B.V. All rights reserved.

Bill Gates recently observed that “the emergence of the robotics industry ... is developing in much the same way that the computer business did 30 years ago” [18]. As a key architect of the computer industry, his prediction has special weight. In a few decades—or sooner, given exponential progress forecasted by Moore’s Law—robots in society will be as ubiquitous as computers are today, he believes; and we would be hard-pressed to find an expert who disagrees.

But consider just a few of the challenges linked to computers in the last 30 years: They have displaced or severely threatened entire industries, for instance, typewriter manufacturing and sales by word-processing software, accountants by spreadsheets, artists by graphic-design programs, and many local businesses by Internet retailers. Customer-tracking websites, street-view maps, and the free and anonymous flow of information online still raise privacy concerns. The digital medium enables sharing that may infringe on copyright claims, and a largely unregulated process of registering domain names has led to charges of cybersquatting or trademark disputes. The effects of social networking and virtual reality on real-world relationships are still unclear, and cyberbullying is a new worry for parents. Internet addiction, especially to online gaming and pornography, continues to ruin real lives. Security efforts to protect corporate networks and personal computers require a massive educational campaign, not unlike safe-sex programs in the physical world. And so on.

* Corresponding author. Tel.: +1 (805) 756 2041.

E-mail addresses: palin@calpoly.edu (P. Lin), kabney@calpoly.edu (K. Abney), gbekey@calpoly.edu (G. Bekey).

¹ Patrick Lin is the director of the Ethics + Emerging Sciences Group—<http://ethics.calpoly.edu>—and an assistant philosophy professor at California Polytechnic State University, San Luis Obispo. He is also an affiliated scholar at Stanford Law School’s Center for Internet and Society and a visiting research fellow at Australia’s Centre for Applied Philosophy and Public Ethics (CAPPE).

² Keith Abney is a senior philosophy lecturer and bioethicist at California Polytechnic State University, San Luis Obispo. Tel.: +1 (805) 756 2041.

³ George Bekey is a professor emeritus of computer science, electrical engineering, and biomedical engineering at University of Southern California as well as founder of its robotics lab. He is also a distinguished adjunct professor of engineering at California Polytechnic State University, San Luis Obispo. Tel.: +1 (805) 756 2131.

To be clear, these are not arguments that the computer industry should never have been developed, but only that its benefits need to be weighed against its negative effects. However, we are not interested in making such a cost-benefit evaluation here but would like to focus on an important lesson: If the evolution of the robotics industry is analogous to that of computers, then we can expect important social and ethical challenges to rise from robotics as well, and attending to them sooner rather than later will likely help mitigate those negative consequences.

Society has long been concerned with the impact of robotics, even before the technology was viable. Beginning with the first time the word ‘robot’ was coined [13], most literary works about robots are cautionary tales about insufficient programming, emergent behavior, errors, and other issues that make robots unpredictable and potentially dangerous (e.g., [5,6,16,45]). In popular culture, films continue to dramatize and demonize robots, such as *Metropolis*, *Star Wars*, *Blade Runner*, *Terminator*, *AI*, and *I, Robot*, to name just a few. Headlines today also stoke fears about robots wreaking havoc on the battlefield as well as financial trading markets, perhaps justifiably so (e.g., [19]).

A loose band of scholars worldwide has been researching issues in robot ethics for some time (e.g., [42]). And a few reports and books are trickling into the marketplace (e.g., [43,26,38]). But there has not yet been a single, accessible resource that draws together such thinking on a wide range of issues, e.g., programming design, military affairs, law, privacy, religion, healthcare, sex, psychology, robot rights, and more. To fill that need, the authors of this paper are in the process of editing a collection of robot-ethics papers [28] for MIT Press, a leading publisher in robotics as well as ethics. In this journal paper, we will briefly introduce the major issues in robot ethics.

1. What is a robot?

Let us start with a basic issue: What is a robot? Given society’s long fascination with robotics, it seems hardly worth asking the question, as the answer surely must be obvious. On the contrary, there is still a lack of consensus among roboticists on how they define the object of their craft. For instance, an intuitive definition could be that a robot is merely a computer with sensors and actuators that allow it to interact with the external world; however, any computer that is connected to a printer or can eject a CD might qualify as a robot under that definition, yet few roboticists would defend that implication.

Certainly, artificial intelligence (AI) by itself can raise interesting issues, such as whether we ought to have humans in the loop more in critical systems, e.g., those controlling energy grids and making financial trades, lest we risk widespread blackouts and stock-market crashes [14,29]. But robots or embodied AI that can directly exert influence on the world seem to pose additional or special risks and ethical quandaries we want to distinguish here. A plausible definition, therefore, needs to be more precise and distinguish robots from mere computers and other devices.

We do not presume we can resolve this great debate here, but it is important that we offer a working definition prior to laying out the landscape of current and predicted applications of robotics. In its most basic sense, we define “robot” as *an engineered machine that senses, thinks, and acts*: “Thus a robot must have sensors, processing ability that emulates some aspects of cognition, and actuators. Sensors are needed to obtain information from the environment. Reactive behaviors (like the stretch reflex in humans) do not require any deep cognitive ability, but on-board intelligence is necessary if the robot is to perform significant tasks autonomously, and actuation is needed to enable the robot to exert forces upon the environment. Generally, these forces will result in motion of the entire robot or one of its elements (such as an arm, a leg, or a wheel)” [9].

This definition does not imply that a robot must be electromechanical; it leaves open the possibility of biological robots, as well as virtual or software ones. But it does rule out as robots any *fully* remote-controlled machines, since those devices do not “think”, e.g., many animatronics and children’s toys. That is, most of these toys do not make decisions for themselves; they depend on human input or an outside actor. Rather, the generally accepted idea of a robot depends critically on the notion that it exhibits some degree of autonomy or can “think” for itself, making its own decisions to act upon the environment. Thus, the US Air Force’s Predator unmanned aerial vehicle (UAV), though mostly tele-operated by humans, makes some navigational decisions on its own and therefore would count as a robot. By the same definition, the following things are not robots: conventional landmines, toasters, adding machines, coffee makers, and other ordinary devices.

As should be clear by now, the definition of “robot” also trades on the notion of “think”, another source of contention which we cannot fully engage here. By “think”, what we mean is that the machine is able to process information from sensors and other sources, such as an internal set of rules either programmed or learned, and to make some decisions autonomously. Of course, this definition merely postpones our task and invites another question: What does it mean for machines to have autonomy? If we may simply stipulate it here, we define “autonomy” in robots as *the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time* [9].

Thus again, *fully* remote- or tele-operated machines would not count as autonomous, since they depend on external control; they cannot “think” and therefore cannot act for themselves. But what about the everyday desktop or laptop computers: Are they autonomous? Doesn’t their programming count as human inputs or external control in some important sense? If so, how can robots ever be said to be free from external control, if all robots are computers (electromechanical or otherwise) at their core?

These are all good questions that demand answers, for a complete discussion of what it means to be a robot. Many will engage other difficult issues from technical to philosophical, such as complexity, unpredictability, determinism, responsibility,

and free will. As such, we cannot offer a complete discussion given space limitations of this paper, and we will have to content ourselves with the working definitions stipulated above—which should be enough to understand why we include some machines and not others in the following section.

2. Robots today and tomorrow

Robots are often tasked to perform the “three Ds”, that is, jobs that are dull, dirty, or dangerous. For instance, automobile factory robots execute the same, repetitive assemblies over and over, with precision and without complaint; military surveillance UAVs patrol the skies for far more hours than a human pilot can endure at a time. Robots crawl around in dark sewers, inspecting pipes for leaks and cracks, as well as do the dirty work in our homes, such as vacuuming floors. Not afraid of danger, they also explore volcanoes and clean up contaminated sites, in addition to more popular service in defusing bombs and mediating hostage crises.

We can also think of robots more simply and broadly—as human replacements. More than mere tools which cannot think and act independently, robots are able to serve in many old and new roles in society that are often handicapped, or made impossible, by human frailties and limitations; that is, semi- and fully-autonomous machines could carry out those jobs more optimally. Beyond the usual “three D’s”, robots perform delicate and difficult surgeries, which are risky with shaky human hands. They can navigate inaccessible places, such as the ocean floor or Mars. As the embodiment of AI, they are more suited for jobs that demand information processing and action too quick for a human, such as the US Navy’s Phalanx CIWS that detects, identifies, and shoots down enemy missiles rapidly closing in on a ship. Some argue that robots could replace humans in situations where emotions are liabilities, such as battlefield robots that do not feel anger, hatred, cowardice, or fear—human weaknesses that often cause wartime abuses and crimes by human soldiers [4]. Given such capabilities, we find robots already in society or under development in a wide range of roles, such as:

Labor and services: Nearly half of the world’s 7-million-plus service robots are Roomba vacuum cleaners [21], but others exist that mow lawns, wash floors, iron clothes, move objects from room to room, and other chores around the home. Robots have been employed in manufacturing for decades, particularly in auto factories, but they are also used in warehouses, movie sets, electronics manufacturing, food production, printing, fabrication, and many other industries.

Military and security: Grabbing headlines are war robots with fierce names such as Predator, Reaper, Big Dog, Crusher, Harpy, BEAR, Global Hawk, Dragon Runner, and more. They perform a range of duties, such as spying or surveillance (air, land, underwater, space), defusing bombs, assisting the wounded, inspecting hideouts, and attacking targets. Police and security robots today perform similar functions, in addition to guarding borders and buildings, scanning for pedophiles and criminals, dispensing helpful information, reciting warnings, and more. There is also a growing market for home-security robots, which can shoot pepper spray or paintball pellets and transmit pictures of suspicious activities to their owners’ mobile phones.

Research and education: Scientists are using robots in laboratory experiments and in the field, such as collecting ocean surface and marine-life data over extended periods (e.g., Rutgers University’s Scarlet Knight) and exploring new planets (e.g., NASA’s Mars Exploration Rovers). In classrooms, robots are delivering lectures, teaching subjects (e.g., foreign languages, vocabulary, and counting), checking attendance, and interacting with students.

Entertainment: Related to the above is the field of “edutainment” or education-entertainment robots, which include ASIMO, Nao, iCub, and others. Though they may lack a clear use, such as for military or manufacturing, they aid researchers in the study of cognition (both human and artificial), motion, and other areas related to the advancement of robotics. Robotic toys, such as AIBO, Pleo, and RoboSapien, also serve as discovery and entertainment platforms.

Medical and healthcare: Some toy-like robots, such as PARO which looks like a baby seal, are designed for therapeutic purposes, such as reducing stress, stimulating cognitive activity, and improving socialization. Similarly, University of Southern California’s socially assistive robots help coach physical-therapy and other patients. Medical robots, such as da Vinci Surgical System and ARES ingestible robots, are assisting with or conducting difficult medical procedures on their own. RIBA, IWARD, ERNIE, and other robots perform some the functions of nurses and pharmacists.

Personal care and companions: Robots are increasingly used to care for the elderly and children, such as RI-MAN, PaPeRo, and CareBot. PALRO, QRIO, and other edutainment robots mentioned above can also provide companionship. Surprisingly, relationships of a more intimate nature are not quite satisfied by robots yet, considering the sex industry’s reputation as an early adopter of new technologies. Introduced in 2010, Roxxy is billed as “the world’s first sex robot” [17], but its lack of autonomy or capacity to “think” for itself, as opposed to merely respond to sensors, suggests that it is not in fact a robot, per the definition above.

Environment: Not quite as handy as WALL-E, robots today still perform important functions in environmental remediation, such as collect trash, mop up after nuclear power plant disasters, remove asbestos, cap oil geysers, sniff out toxins, identify polluted areas, and gather data on climate warming.

In the future: As AI advances, we can expect robots to play more complex and a wider range of roles in society: For instance, police robots equipped with biometrics capabilities and sensors could detect weapons, drugs, and faces at a distance. Military robots could make attack decisions on their own; in most cases today, there is a human triggerman behind those robots. Driverless trains today and DARPA's Grand Challenges are proof-of-concepts that robotic transportation is possible, and even commercial airplanes today are controlled autonomously for a significant portion of their flight today, never mind military UAVs. A general-purpose robot, if achievable, could service many of our domestic labor needs, as opposed to a team of robots each with its own job.

We can also expect robots to scale down as well as up: Some robots are miniature today and ever shrinking, perhaps bringing to life the idea of a “nano-bot”, swarms of which might work inside our bodies or in the atmosphere or cleaning up oil spills. Even rooms or entire buildings might be considered as robots—beyond the “smart homes” of today—if they can manipulate the environment in ways more significant than turning on lights and air conditioning. With synthetic biology, cognitive science, and nanoelectronics, future robots could be biologically based. And man-machine integrations, i.e., cyborgs, may be much more prevalent than they are today, which are mostly limited to patients with artificial body parts, such as limbs and joints that are controlled to some degree by robotics. Again, much of this speaks to the fuzziness of the definition of robot: What we intuitively consider as robots today may change given different form-factors and materials of tomorrow.

In some countries, robots are quite literally replacements for humans, such as Japan, where a growing elderly population and declining birthrates mean a shrinking workforce [35]. Robots are built to specifically fill that labor gap. And given the nation's storied love of technology, it is therefore unsurprising that approximately one out of 25 workers in Japan is a robot [32]. While the US currently dominates the market in military robotics, nations such as Japan and South Korea lead in the market for social robotics, such as elderly-care robots. Other nations with similar demographics, such as Italy, are expected to introduce more robotics into their societies, as a way to shore up a decreasing workforce [19]; and nations without such concerns can drive productivity, efficiency, and effectiveness to new heights with robotics.

3. Ethical and social issues

The Robotics Revolution promises a host of benefits that are compelling and imaginative, but as with other emerging technologies, they also come with risks and new questions that society must confront. This is not unexpected, given the disruptive nature of technology revolutions. In the following, we map the myriad issues into three broad (and interrelated) areas of ethical and social concern and provide representative questions for each area:

3.1. Safety and errors

We have learned by now that new technologies, first and foremost, need to be safe. Asbestos, DDT, and fen-phen are among the usual examples of technology gone wrong (e.g., [41,20,24]), having been introduced into the marketplace before sufficient health and safety testing. A similar debate is occurring with nanomaterials now (e.g., [3]).

With robotics, the safety issue is with their software and design. Computer scientists, as fallible human beings, understandably struggle to create a perfect piece of complex software: somewhere in the millions of lines of code, typically written by teams of programmers, errors and vulnerabilities likely exist. While this usually does not result in significant harm with, say, office applications—just lost data if users do not periodically save their work (which arguably is their own fault)—even a tiny software flaw in machinery, such as a car or a robot, could lead to fatal results.

For instance, in August 2010, the US military lost control of a helicopter drone during a test flight for more than 30 minutes and 23 miles, as it veered towards Washington DC, violating airspace restrictions meant to protect the White House and other governmental assets [11]. In October 2007, a semi-autonomous robotic cannon deployed by the South African army malfunctioned, killing nine “friendly” soldiers and wounding 14 others (e.g., [34]). Experts continue to worry about whether it is humanly possible to create software sophisticated enough for armed military robots to discriminate combatants from noncombatants, as well as threatening behavior from nonthreatening (e.g., [26]).

Never mind the scores of other military-robot accidents and failures [46], human deaths can and have occurred in civilian society: The first human to be killed by a robot was widely believed to be in 1979, in an auto factory accident in the US [23]. And it does not take much to imagine that a mobile city-robot—a heavy piece of machinery—could accidentally run over a small child.

Hacking is an associated concern, given how much attention is paid to computer security today. What makes a robot useful—its strength, ability to access and operate in difficult environments, expendability, and so on—could also be turned against us, either by criminals or simply mischievous persons. This issue will become more important as robots become networked and more indispensable to everyday life, as computers and smart phones are today. Indeed, the fundamentals

of robotics technology are not terribly difficult to master: as formidable and fearsome as military robots are today, already more than 40 nations have developed those capabilities, including Iran [39,15].

Thus, some of the questions in this area include: Is it even possible for us to create machine intelligence that can make nuanced distinctions, such as between a gun and an ice-cream cone pointed at it, or understand human speech that is often heavily based on context? What are the tradeoffs between non-programming solutions for safety—e.g., weak actuators, soft robotic limbs or bodies, using only non-lethal weapons, or using robots in only specific situations such as a “kill box” in which all humans are presumed to be enemy targets—and the limitations they create? How safe ought robots be prior to their introduction into the marketplace or society, i.e., should a precautionary principle apply here? How would we balance the need to safeguard robots from running amok (e.g., with a kill-switch) with the need to protect it from hacking or capture? How can the “frame problem” be solved in practice, in such a way as to ensure robots only take salient, relevant safety information into account?

3.2. Law and ethics

Linked to the risk of robotic errors, it may be unclear who is responsible for any resulting harm. Product liability laws are largely untested in robotics and, anyway, continue to evolve in a direction that releases manufacturers from responsibility, e.g., end-user license agreements in software. With military robots, for instance, there is a list of characters throughout the supply chain that may be held accountable: the programmer, the manufacturer, the weapons legal review team, the military procurement officer, the field commander, the robot’s handler, and even the President of the United States, as the commander-in-chief.

As robots become more autonomous, it may be plausible to assign responsibility to the *robot itself*, e.g., if it is able to exhibit enough of the features that typically define personhood. If this seems too far-fetched, consider that synthetic biology is forcing society to reconsider the definition of life, blurring the line between living and non-living agents (e.g., [10]). Also consider that there is ongoing work in integrating computers and robotics with biological brains (e.g., [44]). A conscious human brain (and its body) presumably has human rights, and replacing parts of the brain with something else, while not impairing its function, would seem to preserve at least some of those rights and responsibilities (and possibly even add novel rights as new capacities emerge, given “ought implies can”). We may come to a point at which more than half of the brain or body is artificial, making the organism more robotic than human; seeing such a continuum between humans and robots may make the issue of robot rights and duties more plausible. And if some (future) robots or cyborgs meet the necessary requirements to have rights, which ones should they have, and how does one manage such portfolios of rights, which may be unevenly distributed given a range of biological and technological capabilities?

In the near term, one natural way to think about minimizing risk of harm from robots is to program them to obey our laws or follow a code of ethics. Of course, this is much easier said than done, since laws can be vague and context-sensitive, which robots may not be sophisticated enough to understand, at least in the foreseeable future. Even the three (or four) law of robotics in Asimov’s stories, as elegant and sufficient as they appear to be, create loopholes that result in harm (e.g., [6–8]).

Programming aside, the use of robots must also comply with law and ethics, and again those rules and norms may be unclear or untested on such issues. For instance, landmines are an effective but horrific weapon that indiscriminately kills, whether soldiers or children; they have existed for hundreds of years, but it was only in 1983—after their heavy use in 20th century wars—that certain uses of landmines were banned, e.g., planting them without means to identify and remove them later [40]; and only in 1999 did an international treaty ban the production and use of landmines [1]. Likewise, the use of military robots may raise legal and ethical questions that we have yet to fully consider (e.g., [26,27]) and, later in retrospect, may seem obviously unethical or unlawful.

Another relevant area of law concerns privacy. Several forces are driving this concern, including: the shrinking size of digital cameras and other recording devices, an increasing emphasis on security at the expense of privacy (e.g., expanded wiretap laws, a blanket of surveillance cameras in some cities to monitor and prevent crimes), advancing biometrics capabilities and sensors, and database integrations. Besides robotic spy planes, we previously mentioned (future) police robots that could conduct intimate surveillance at a distance, such as detecting hidden drugs or weapons and identifying faces unobtrusively; if linked to databases, they could also run background checks on one’s driving, medical, banking, shopping, or other records to determine if the person should be apprehended [36]. Domestic robots too can be easily equipped with surveillance devices—as home security robots already are—that may be monitored or accessed by third parties [12].

Of course, ethical and cultural norms, and therefore law, vary around the world, so it is unclear *whose* ethics and law ought to be the standard in robotics; and if there is no standard, which jurisdictions would gain advantages or cause policy challenges internationally? Such challenges may require international policies, treaties, and perhaps even international laws and enforcement bodies. This kind of political-cultural schism is not merely theoretical, but one we have already seen in military law: the US, for instance, has refused to sign or accede the aforementioned landmine ban, also known as the Ottawa Treaty. The relationship of robotic aircraft (drones) to international law is likewise a vexed issue: the US assumes its Predator attacks in Pakistan are legal, whereas many other countries disagree. And Japanese and Americans may well have different moral sensibilities in leaving care of the elderly solely in the hands (or extender arms) of robots. From these global variations in ethics and law, it may be reasonable to expect an uneven trajectory for robot development worldwide, which affects the proliferation of associated benefits and pragmatic challenges.

Other questions in this area include: If we could program a code of ethics to regulate robotic behavior, which ethical theory should we use? Are there unique legal or moral hazards in designing machines that can autonomously kill people? Or should robots merely be considered tools, such as guns and computers, and regulated accordingly? Is it ethically permissible to abrogate responsibility for our elderly and children to machines that seem to be a poor substitute for human companionship (but perhaps better than no—or abusive—companionship)? Will robotic companionship (that could replace human or animal companionship) for other purposes, such as drinking buddies, pets, other forms of entertainment, or sex, be morally problematic? At what point should we consider a robot to be a “person”, thus affording it some rights and responsibilities, and if that point is reached, will we need to emancipate our robot “slaves”? Do we have any other distinctive moral duties towards robots? As they develop enhanced capacities, should cyborgs have a different legal status than ordinary humans? At what point does a technology-mediated surveillance count as a “search”, which would generally require a judicial warrant? Are there particular moral qualms with placing robots in positions of authority, such as police, prison or security guards, teachers, or any other government roles or offices in which humans would be expected to obey robots?

3.3. Social impact

How might society change with the Robotics Revolution? As with the Industrial and Internet Revolutions, one key concern is about a loss of jobs. Factories had replaced legions of workers who used to perform the same work by hand, giving way to the faster, more efficient processes of automation. Internet ventures such as Amazon.com, eBay, and even smaller “e-tailers” are still edging out brick-and-mortar retailers who have much higher overhead and operating expenses. Likewise, as potential replacements for humans—performing certain jobs better, faster, and so on—robots may displace human jobs, regardless of whether the workforce is growing or declining.

The standard response is that human workers, whether replaced by other humans or machines, would then be free to focus their energies where they can make a greater impact, i.e., at jobs in which they have a greater competitive advantage [33]; to resist this change is to support inefficiency. For instance, by outsourcing call-center jobs to other nations where the pay is less, displaced workers (in theory) can perform “higher-value” jobs, whatever those may be. Further, the demand for robots itself creates additional jobs. Yet, theory and efficiency provide little consolation for the human worker who needs a job to feed her or his family, and cost-benefits may be negated by unintended effects, e.g., a negative customer experience with call-center representatives whose first language is not that of the customers.

Connected to labor, some experts are concerned about technology dependency (e.g., [42]). For example, as robots prove themselves to be better than humans in difficult surgeries, the resulting loss of those jobs may also mean the gradual loss of that medical skill or knowledge, to the extent that there would be fewer human practitioners. This is not the same worry with labor and service robots that perform dull and dirty tasks, in that we care less about the loss of those skills; but there is a similar issue of becoming overly reliant on technology for basic work. For one thing, this dependency seems to cause society to be more fragile: for instance, the Y2K problem caused significant panic, since so many critical systems—such as air-traffic control and banking—were dependent on computers whose ability to correctly advance their internal clock to January 1, 2000 (as opposed to resetting it to January 1, 1900) was uncertain; and similar situations exist today with malicious computer *viruses du jour*.

Like the social networking and email capabilities of the Internet Revolution, robotics may profoundly impact human relationships. Already, robots are taking care of our elderly and children, though there are not many studies on the effects of such care, especially in the long term. Some soldiers have emotionally bonded with the bomb-disposing PackBots that have saved their lives, sobbing when the robot meets its end (e.g., [38,22]). And robots are predicted to soon become our lovers and companions [25]: they will always listen and never cheat on us. Given the lack of research studies in these areas, it is unclear whether psychological harm might arise from replacing human relationships with robotic ones.

Harm also need not be directly to persons, e.g., it could also be to the environment. In the computer industry, “e-waste” is a growing and urgent problem (e.g., [31]), given the disposal of heavy metals and toxic materials in the devices at the end of their product lifecycle. Robots as embodied computers will likely exacerbate the problem, as well as increase pressure on rare-earth elements needed today to build computing devices and energy resources needed to power them. Networked robots would also increase the amount of ambient radiofrequency radiation, like that created by mobile phones—which have been blamed, fairly or not, for a decline of honeybees necessary for pollination and agriculture [37], in addition to human health problems (e.g., [2]).

Thus, some of the questions in this area include: What is the predicted economic impact of robotics, all things considered? How do we estimate the expected costs and benefits? Are some jobs too important or too dangerous for machines to take over? What do we do with the workers displaced by robots? How do we mitigate disruption to a society dependent on robotics, if those robots were to become inoperable or corrupted, e.g., through an electromagnetic pulse or network virus? Is there a danger with emotional attachments to robots? Are we engaging in deception by creating anthropomorphized machines that may lead to such attachments, and is that bad? Is there anything essential in human companionship and relationships that robots cannot replace? What is the environmental impact of a much larger robotics industry than we have today? Could we possibly face any truly cataclysmic consequences from the widespread adoption of social robotics, and if so, should a precautionary principle apply?

4. Engaging the issues now

These are only some of the questions which the emerging field of robot ethics is concerned with, and many of these questions lead to the doorsteps of other areas of ethics and philosophy, e.g., computer ethics and philosophy of mind, in addition to the disciplines of psychology, sociology, economics, politics, and more. Note also that we have not even considered the more popular “Terminator” scenarios in which robots—through super artificial intelligence—subjugate humanity, which are highly speculative scenarios that continually overshadow more urgent and plausible issues.

The robotics industry is rapidly advancing, and robots in society today are already raising many of these questions. This points to the need to attend to robot ethics now, particularly as consensus on ethical issues is usually slow to catch up with technology, which can lead to a “policy vacuum” [30]. As an example, the Human Genome Project was started in 1990, but it took 18 years after that for Congress to finally pass a bill to protect Americans from discrimination based on their genetic information. Right now, society is still fumbling through privacy, copyright, and other intellectual property issues in the Digital Age, nearly 10 years since Napster was first shut down.

As researchers and educators, we hope that our forthcoming volume of robot-ethics papers will provide and motivate greater discussion—both in and outside the classroom—across the broad continuum of issues, such as described above. The contributors to our edited book include many respected and well-known experts in robotics and technology ethics today, including: Colin Allen, Peter Asaro, Anthony Beavers, Selmer Bringsjord, Marcello Guarini, James Hughes, Gert-Jan Lokhorst, Matthias Scheutz, Noel Sharkey, Rob Sparrow, Jeroen van den Hoven, Gianmarco Veruggio, Wendell Wallach, Kevin Warwick, and others.

Sometimes to deaf ears, history lectures us on the importance of foresight: While the invention of such things as the printing press, gunpowder, automobiles, computers, vaccines, and so on, has profoundly changed the world (for the better, we hope), they have also led to unforeseen consequences, or perhaps consequences that might have been foreseen and addressed had we bothered to investigate them. Least of all, they have disrupted the *status quo*, which is not necessarily a terrible thing in and of itself; but unnecessary and dramatic disruptions, such as mass displacements of workers or industries, have real human costs to them. Given lessons from the past, society is beginning to think more about ethics and policy in advance of, or at least in parallel to, the development of new game-changing technologies, such as genetically-modified foods, nanotechnology, neuroscience, and human enhancement—and now we add robotics to that syllabus.

At the same time, we recognize that these technologies seem to jump out of the pages of science fiction, and the ethical dilemmas they raise also seem too distant to consider, if not altogether unreal. But as Isaac Asimov foretold: “It is change, continuing change, inevitable change, that is the dominant factor in society today. No sensible decision can be made any longer without taking into account not only the world as it is, but the world as it will be . . . This, in turn, means that our statesmen, our businessmen, our everyman must take on a science fictional way of thinking” [7]. With human ingenuity, what was once fiction is becoming fact, and the new challenges it brings are all too real.

Acknowledgements

Some of this paper is excerpted from our book *Robot Ethics: The Ethical and Social Implications of Robotics*, to be published by MIT Press in late 2011 [28]. We would like to thank Kyle Campbell for his editorial assistance with this paper.

References

- [1] Jeff Abramson, Arms Control Association: The Ottawa Convention at a Glance, June 2008. Accessible at <http://www.armscontrol.org/factsheets/ottawa>. Last accessed on September 12, 2010.
- [2] Ashok Agarwal, Nisarg R. Desai, Kartikeya Makker, Alex Varghese, Rand Mouradi, Edmund Sabanegh, Rakesh Sharma, Effects of radiofrequency electromagnetic waves (RF-EMW) from cellular phones on human ejaculated semen: an in vitro pilot study, *Fertility and Sterility* 92 (4) (2009) 1318–1325.
- [3] Fritz Allhoff, Patrick Lin, Daniel Moore, What Is Nanotechnology and Why Does It Matter?: From Science to Ethics, Wiley-Blackwell, Hoboken, NJ, 2010.
- [4] Ronald C. Arkin, Governing lethal behavior: embedding ethics in a hybrid deliberative/hybrid robot architecture, Report GIT-GVU-07-11, Georgia Institute of Technology's GVV Center, Atlanta, GA, 2007. Accessible at <http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf>. Last accessed on September 12, 2010.
- [5] Isaac Asimov, I, Robot, 2004 edition, Bantam Dell, New York, NY, 1950.
- [6] Isaac Asimov, The Naked Sun, Doubleday, New York, NY, 1957.
- [7] Isaac Asimov, My own view, in: Robert Holdstock (Ed.), *The Encyclopedia of Science Fiction*, St. Martin's Press, New York, NY, 1978.
- [8] Isaac Asimov, Robots and Empire, Doubleday, New York, NY, 1985.
- [9] George Bekey, Autonomous Robots: From Biological Inspiration to Implementation and Control, MIT Press, Cambridge, MA, 2005.
- [10] Steven A. Benner, Q&A: life, synthetic biology, and risk, *BMC Biology* 8 (2010) 77.
- [11] Elisabeth Bumiller, The New York Times: Navy Drone Violated Washington Airspace (Aug. 25, 2010).
- [12] M. Ryan Calo, Robots and privacy, in: Patrick Lin, George Bekey, Keith Abney (Eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA, 2010 (forthcoming). Accessible at <http://ssrn.com/abstract=1599189>. Last accessed on September 12, 2010.
- [13] Karel Čapek, Rossum's Universal Robots, 2004 edition, Penguin Group, New York, NY, 1921 (trans. Claudia Novack).
- [14] Keay Davidson, How a butterfly's wing can bring down Goliath, *San Francisco Chronicle* (August 15, 2003).
- [15] Defense Update, Karrar-Iran's new jet-powered recce and attack drone, 2010. Accessible at http://defense-update.com/products/k/karrar_jet_powered_drone_24082010.html. Last accessed on September 12, 2010.
- [16] Philip K. Dick, Do Androids Dream of Electric Sheep?, Del Rey Books, New York, NY, 1968.
- [17] Yvonne Fulbright, Meet Roxxy, the ‘woman’ of your dreams, 2010. Accessible at <http://www.foxnews.com/story/0,2933,583314,00.html>. Last accessed on September 12, 2010.

- [18] Bill Gates, A robot in every home, *Scientific American* (January 2007) 58–65.
- [19] Gary Geipel, Global aging and the global workforce, a Hudson Institute article, 2003. Accessible at http://www.hudson.org/index.cfm?fuseaction=publication_details&id=2740. Last accessed on September 12, 2010.
- [20] Christine Gorman, Danger in the diet pills?, *Time Magazine* (July 21, 1997). Accessible at <http://www.time.com/time/magazine/article/0,9171,986725,00.html>. Last accessed on September 12, 2010.
- [21] Erico Guizzo, IEEE Spectrum: World Robot Population Reaches 8.6 Million, April 14, 2010. Accessible at <http://spectrum.ieee.org/autotom/robotics/industrial-robots/041410-world-robot-population>. Last accessed on September 12, 2010.
- [22] Jeremy Hsu, Real soldiers love their robot brethren, *Live Science* (May 21, 2009). Accessible at <http://www.livescience.com/technology/090521-terminator-war.html>. Last accessed on September 12, 2010.
- [23] Tim Kiska, Death on the job: Jury awards \$10 million to heirs of man killed by robot at auto plant, *Philadelphia Inquirer* (August 11, 1983) A-10.
- [24] Linda Lear, *Rachel Carson: Witness for Nature*, Henry Hoyten, New York, NY, 1997.
- [25] David Levy, *Love and Sex with Robots: The Evolution of Human–Robot Relationships*, Harper Collins Publishers, New York, NY, 2007.
- [26] Patrick Lin, George Bekey, Keith Abney, Autonomous military robots: risk, ethics, and design. A report commissioned by US Department of Navy/Office of Naval Research, 2008. Accessible at http://ethics.calpoly.edu/ONR_report.pdf. Last accessed on September 12, 2010.
- [27] Patrick Lin, George Bekey, Keith Abney, Robots in war: issues of risk and ethics, in: Rafael Capurro, Michael Nagenborg (Eds.), *Ethics and Robotics*, AKA Verlag/IOS Press, Heidelberg, Germany, 2009.
- [28] Patrick Lin, Keith Abney, George Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA (forthcoming).
- [29] Alexis Madrigal, Market data firm spots the tracks of bizarre robot traders, *The Atlantic* (August 4, 2010).
- [30] James H. Moor, Metaphilosophy: What is computer ethics? 16 (4) (1985) 266–275.
- [31] Amy Joi O'Donoghue, E-waste is a growing issue for states, *Deseret News* (August 22, 2010). Accessible at <http://www.deseretnews.com/article/70059360/E-waste-is-a-growing-issue-for-states.html?pg=1>. Last accessed on September 12, 2010.
- [32] RedOrbit, Japan hopes to employ robots by 2025, April 8, 2008. Accessible at http://www.redorbit.com/news/technology/1332274/japan_hopes_to_employ_robots_by_2025/. Last accessed on September 12, 2010.
- [33] Mitch Rosenberg, The surprising benefits of robots in the DC, Supply & Demand Chain Executive (June/July 2009).
- [34] Noah Schactman, Robot cannon kills 9, wounds 14, *Wired* (October 18, 2007). Accessible at <http://blog.wired.com/defense/2007/10/robot-cannon-ki.html>. Last accessed on September 12, 2010.
- [35] Chana Schoenberger, Japan's shrinking workforce, *Forbes* (May 25, 2008). Accessible at http://www.forbes.com/2008/05/25/immigration-labor-visa-oped-cx_crs_outsourcing08_0529japan.html. Last accessed on September 12, 2010.
- [36] Noel Sharkey, 2084: big robot is watching you, A Commissioned Report, 2008. Accessible at <http://www.dcs.shef.ac.uk/~noel/>. Last accessed on September 12, 2010.
- [37] Ved Parkash Sharma, Neelima R. Kumar, Changes in honeybee behaviour and biology under the influence of cellphone radiations, *Current Science* 98 (10) (2010) 1376–1378.
- [38] Peter W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century*, Penguin Press, New York, NY, 2009.
- [39] Peter W. Singer, Robots at war: the new battlefield, *Wilson Quarterly* (Winter 2009). Accessible at <http://www.wilsonquarterly.com/article.cfm?aid=1313>. Last accessed on September 12, 2010.
- [40] United Nations, The Convention on Certain Conventional Weapons, Entered into force on December 2, 1983. Accessible at <http://www.armscontrol.org/factsheets/CCW>. Last accessed on September 12, 2010.
- [41] US Environmental Protection Agency, Asbestos ban and phase out, April 25, 2007. Accessible at <http://www.epa.gov/asbestos/pubs/ban.html>. Last accessed on September 12, 2010.
- [42] Gianmarco Veruggio (Ed.), EURON Roboethics Roadmap, EURON Roboethics Atelier, EURON, Genoa, Italy, 2006. Accessible at <http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf>. Last accessed on September 12, 2010.
- [43] Wendell Wallach, Colin Allen, *Moral Machines: Teaching Robots Right From Wrong*, Oxford University Press, New York, 2009.
- [44] Kevin Warwick, Implications and consequences of robots with biological brains, in: Anthony Beavers (Ed.), Special Issue on Robot Ethics and Human Ethics, *Ethics & Information Technology* 12 (1) (2010).
- [45] Daniel H. Wilson, *How to Survive a Robot Uprising: Tips on Defending Yourself Against the Coming Rebellion*, Bloomsbury Publishing, New York, NY, 2005.
- [46] David Zucchino, War zone drone crashes add up, *Los Angeles Times* (July 6, 2010).

23

ROBOTS AND SPACE ETHICS

Keith Abney

Space: the final frontier . . . of nonhuman exploration. To boldly go where no robot has gone before . . . these are the travels of the Starship Von Neumann probe.

Not the way the Trekkies remember the taglines. But should this be our future, instead of one in which oversensitive, emotional, violent, fragile meatbags constantly get in trouble because their vulnerabilities cause endless troubles in the exploration of outer space? As Mr. Spock might say, that seems . . . illogical. And thinking critically about the reasons for using robots, instead of humans, in space can help dispel the often overemotional, militaristic, and patriotic veneer that has accompanied discussion of human spaceflight; and it can help inform discussion of robots replacing humans in other contexts.

To start, why should humans go into space at all? If the goal is exploration, robotic probes can travel and work in environments far too inhospitable for humans. Robots, after all, excel in the traditional “three Ds”—environments too dull, dirty, or dangerous for humans (e.g., Lin 2012). Space exemplifies such an environment, with its terrifying radiation and temperature extremes, lack of air or water, and long stretches of dull reconnaissance necessary during exploratory missions—and increasingly, ever longer flight times to get to our destination. And the robotic fourth “D,” being dispassionate, may well be crucial (Veruggio and Abney 2012); robots will not disobey mission commands out of jealousy, revenge, lust, or any of the other emotional drivers that so often cause humans—even well-trained humans—to commit egregious tactical blunders.

So robots, which need no sleep or healthcare, no food or water, and don’t get tired or hungry or angry, will inevitably do a better job of exploring the cosmos than humans replete with all our biological limitations. And the gap in performance will only widen as technology improves and robotic capabilities become ever more advanced.

Already robots have explored the farthest reaches of our solar system, and even beyond, to interstellar space. Human astronauts have gotten only as far as the moon, and no farther, in almost fifty years.

As for the purpose of human colonization, why believe that humans should go to other planets to live? The difficulties will be stupendous; even the most clement nearby world, Mars, looks like a suicide mission for any human colonists with near-term technology (Do et al. 2016). Even supposing human astronauts could survive a six-month voyage to Mars in a cramped capsule, with hazards from cosmic rays, solar flares, meteorites, and innumerable other travel risks (including each other!), the challenges would hardly dissipate upon arrival, as cinephiles saw Matt Damon demonstrate in *The Martian*.

And successful colonization demands that humans not merely be able to achieve some type of rudimentary, temporary survival upon arrival at Mars (or other destination), but also make that survival sustainable indefinitely, with the associated requirement that, sooner or later, the colonists must successfully reproduce and raise children. But bringing up babies would cause yet more travails, both ethical and existential. Should we allow or demand abortions of the less fit, or demand enhancements, even genetic engineering or species hybridization, in order to up the odds of success (Lin and Abney 2014)?

Even assuming success at colonization is possible, why is that a good thing? Won't humans simply repeat the mistakes of the past? Suppose we find new worlds with living creatures, containing heretofore unknown rich ecosystems; why suppose that we wouldn't simply re-enact our legacy of ecological destruction (including the introduction of both diseases and newly invasive species), as has happened in every virgin territory our forebears settled? Perhaps such considerations leave one unmoved, if one thinks nonhuman life has no moral value. But even so, we should be wary, for perhaps the contamination could go both ways—perhaps human astronauts will return to Earth unwitting bearers of alien microbes that mutate and result in mass illness, or even the death of humanity. If we ever find alien life, we may need off-world quarantine or else risk catastrophe (Abney and Lin 2015). Wouldn't it be better if our first discoverers of, and emissaries to, alien life were our machines?

Even aside from such biomedical and ecological concerns, there remain other serious objections to sending humans into space. One is cost, and particularly the concept of "opportunity cost." It seems plausible to claim that combatting hunger, ecological disruption, climate change, terrorism, or any number of other problems here on Earth deserves higher priority in rationing our limited funds (and other resources) than sending a handful of humans into space. A straightforward application of, say, John Rawls's two principles of justice (1971) would plausibly require that the resources spent on space travel, by robots or (more expensively) humans, would be justified only if they somehow redounded to

the benefit of the poor; and arguing that they produced Tang and astronaut ice cream won't cut it.

And even if the expense of space travel can somehow be morally justified, surely ethics would bid us to do it by the most cost-efficient means, with the highest odds of success, available. And as NASA and other space programs around the world have learned all too well from often painful, even disastrous, experience, unmanned robotic missions offer far more bang for the buck. Just ask the U.S. Air Force: when they wanted a space plane, did they push for the continuation of the shuttle program, with its human crew and disastrous failures? No, the USAF opted for the robotic X-37.

Human pilots are fast becoming a legacy boondoggle among atmospheric craft, as even U.S. Secretary of the Navy Ray Mabus admits (Myers 2015). And the legacy aircraft too costly to retrofit into drones are instead being fitted with a PIBOT—a pilot robot that renders human pilots superfluous for the purposes of flying (*Economist* 2016)—though PIBOT as yet isn't programmed to tell you to look out your window at the Grand Canyon. As Todd Harrison (2015) notes, “[P]hysics, physiology, and fiscal facts” are why drones will soon render manned fighter jets obsolete; the far greater hurdles human pilots face in space, from merely staying alive to the physiological burdens on human performance in harsh zero-G conditions, mean that “physics, physiology, and fiscal facts” have already made human astronauts an extravagant, morally indefensible indulgence.

So what purposes do we serve by going into space? Which of these are unethical? And of those that could be ethical, what role should robots play in fulfilling them? In particular, are there compelling reasons to think our ethical goals in space would require robots to assist humans—or supplant them? Does ethics require that the only spacefaring representatives of our civilization be our machines?

23.1 Purposes: Exploration

23.1.1 *To Go Boldly . . . but Robots Do It Better?*

So why go into space? The most commonly cited purpose is exploration—to boldly go where no human has gone before. But exploration itself can have many (sub)goals. To gain knowledge of a heretofore uncharted realm is one obvious such goal; we want to make a map of *terra incognita*, to specify the physical details of what was formerly unknown. The limits of remote sensing are often used as a justification for the need for human astronauts; the claim is that a human astronaut on, say, Mars could do experiments that the *Spirit* or *Opportunity* rovers could never perform.

But, while true for now, that is largely a function of cost and opportunity; more sophisticated (and costly) robots than the current rovers could perform most if not all of the experiments that a human astronaut could, and still at a fraction of the cost of sending human astronauts on a round-trip mission to Mars. And this does not apply merely to futuristic robots capable of planning and carrying out an entire mission without human guidance from Earth; even short of such full autonomy for robots, the financial considerations remain the same.

After all, paying for a Mars robot and its Earth-based teleoperator (say, a scientist at Mission Control in Houston) is far cheaper than sending a human astronaut to Mars with equivalent tools. And of course, as artificial intelligence improves and space robots become ever more autonomous, the savings will become ever greater. If cost is truly an issue, claiming that we should send human astronauts in order to do it better is essentially always a false economy.

But of course, the human itch to explore the cosmos is not purely for the sake of increasing public scientific knowledge. After all, tourists don't usually go to national parks in order to add to the corpus of scientific knowledge. Instead, we go because they're there; one might say that "wanderlust is in our DNA" (Lin 2006). Space tourism is already part of the business plan for Virgin Galactic, Bigelow Aerospace, and other companies, and the demand is likely to increase.

For at least some humans, the desire to push one's physical, intellectual, emotional, and other boundaries is part of what makes life worth living. The idea that such striving plays a large role in human flourishing is reflected by inspirational memes such as the idea that the value and meaning of life are found in one's experiences, not one's possessions. This innate desire for a kind of private, experiential, personally meaningful knowledge of oneself and one's place in the larger scheme of things could conceivably remain an ethically defensible driver of human space travel, even if human astronauts no longer play any significant role in adding to civilization's store of collective knowledge.

To be clear: insofar as scientifically verifiable, public knowledge is the goal of exploration, robots are already simply better, cheaper, and faster than humans, and their advantage will only increase over time. Their sensors are more reliable, their memories have far greater capacity and are less prone to error, and they can operate in environmental conditions humans would never dare venture into, with far fewer ethical constraints regarding self- and other-preservation when they do.

And our knowledge of the farther reaches of space makes ever more clear the robotic imperative; given how hard it has been to get a human to our closest planetary neighbors, it may require a propulsion breakthrough, or else be a long, long time, before any human will swing into orbit around Jupiter or admire the rings of Saturn from up close. For the near future, space tourism will not leave the Earth-moon system.

So despite these technical difficulties, can a case be made for space tourism as the key to an ethical defense of human travel in space? Not so fast. Even if one could overcome the distributive justice concerns raised by the idea that luxury space travel by billionaire adrenaline junkies is a defensible use of limited resources, we may find that having experiences remotely by robot is better than being there in person! For even the quest for private, experiential knowledge may yet be overcome by the advance of robotic technology: as virtual reality (VR) and/or brain-machine interface (BMI) technology advances, it may be that we could, via remote linkages, have apparently “direct,” unmediated experience of what a robot could see and discover.

After all, a “direct” experience of seeing a new vista is already mediated by your senses and processed by your brain. If a remote robotic connection enables the same immediacy, it could be effectively the same experience (Nozick 1974). In the future, when hooking your brain up to the feed from the Mars rover can enable you to feel the Martian sand squishing beneath your feet, or mainstreaming the feed from the European submarine enables you to see, feel, even smell the alien sharks through the robotic sensors, what possible advantage could you have by actually being there? And so why should we ever bother going there in person?

23.1.2 *Other Purposes: Why We Need Satellites and Autonomous Vehicles*

Another purpose of exploring space is to reduce dangers to humans here on Earth. For instance, satellite technology has served to reduce risks in all sorts of ways: weather satellites allow the tracking of developing hurricanes, enabling meteorologists to project the path and strength of the storm several days to even a week in advance, and so giving people time to evacuate and the authorities time to plan for the aftermath.

Communication satellites allow us near-instant access to information from around the earth; they enable us not only to download TV shows from Netflix or order next-day service from Amazon, but they also enable intelligence agencies to foil terrorist plots. And GPS satellites, in addition to having military and business uses, allow us to track transport vessels on land or sea, as well as people who may be lost or otherwise in need of help (Plait 2007).

Other spacefaring robots protect us in more recondite ways. Satellites such as *SOHO* point their sensors at the sun, so scientists using its data can better understand and predict huge solar eruptions that can damage satellites and cause power blackouts. In 1989, a solar flare “provoked geomagnetic storms that disrupted electric power transmission from the Hydro Québec generating station in Canada, blacking out most of the province and plunging 6 million people into

darkness for 9 hours; aurora-induced power surges even melted power transformers in New Jersey” (Bell and Phillips 2008).

The fearsome Carrington event of 1859 was much larger; people were able to read their newspapers at night from the glow of the aurora, and telegraph wires sparked so much they caught on fire across the United States (Lovett 2011). If a solar storm the size of the Carrington event happened today, much of our communications infrastructure would be wiped out in a literal flash, as GPS and communications satellites would have their circuits fried. Worse yet, if hundreds of giant transformers were destroyed at once, large swaths of the electrical grid could be out for months.

Satellite data are also crucial for studying the effects of the sun’s interaction with Earth itself, from the details of glacial melt in Antarctica and Greenland and other aspects of climate change to pollution emissions. And the risks from space are not only from our sun. Asteroid protection has become a Hollywood blockbuster-level part of our understanding of how we should protect Earth. The growing awareness of impact events ranges from a recent awareness of how the moon formed (when a Mars-sized object struck Earth) to the “dinosaur killer” Chicxulub impactor—an asteroid or, quite possibly, a comet (Rincon 2013)—of 66 million years ago.

More recently, people around the globe enjoyed the use of the Hubble Space Telescope and satellite TV transmissions to witness the depredations of the Shoemaker-Levy 9 comet, whose fragment G struck Jupiter in 1994 with an estimated energy equivalent to 6 million megatons of TNT (about 600 times the estimated nuclear arsenal of the world), leaving a larger-than-Earth-sized scar on the surface of Jupiter (Bruton 1996). All these events and more have enhanced public awareness of the possibility of death from the skies.

In the hope of avoiding such a calamity, the B612 Foundation wants to track Earth-crossing asteroids and is soliciting funding for something called the Sentinel Mission (B612 Foundation 2016), a space-based infrared survey mission to find and catalog 90% of the asteroids larger than 140 meters in Earth’s region of the solar system. If it turns up a killer comet or asteroid that threatens life on Earth, then that presumably would be money well spent (Abney and Lin 2015).

Space missions also help scientists put Earth into a larger context, to understand what developments both on and off the planet are true threats to our way of life. For example, why is Mars dry, cold, nearly airless, and dead, when it apparently had running water, even oceans, billions of years ago? Could that happen to Earth? Could we inadvertently cause it to happen—or deliberately stop it? Or why is Venus covered in thick clouds of carbon dioxide and sulfuric acid (literal brimstone!), with a surface temperature of well over 800 degrees Fahrenheit, thanks to a runaway greenhouse effect? Or why has the hurricane on Jupiter called the “Great Red Spot” lasted for more than three centuries?

Answering such questions can be crucial for understanding our own planet and how human actions are affecting it. And, of course, such understanding is crucial to the moral question, what should we do to try to keep Earth as it is? Or how should we change it? Should we, e.g., attempt geoengineering to combat climate change? And if so, what methods should we use? Perhaps we should try space-based solar power (SBSP), in which robots install huge solar panels in space to block solar radiation from reaching Earth, and produce copious amounts of electricity to boot.

One commonality to all of these uses of space for planetary protection: they do not require a human presence above the atmosphere. Whatever the ethically justified uses of satellites are, none of them require a human to be along for the ride. Having humans in space is simply not required for GPS or communications satellites or solar observatories or space telescopes to work, or even to be repaired. Robots do it better.

And even on the surfaces of the other bodies in the solar system, humans are not required; e.g., autonomous vehicles are already functioning in space in lieu of human astronauts, as the *Spirit* and *Opportunity* rovers make clear. They have already made fundamental discoveries on Mars, and such robotic vehicles will be key to further long-term exploration and colonization on at least the moon and Mars. For equivalent cost, their capabilities will always be far superior to those of space buggies requiring a human driver, like the Apollo rovers on the moon.

Autonomous vehicles would be a major boon even for crewed missions; they could take an incapacitated, wounded, or ill astronaut back to safety without needing another astronaut as a driver. They would serve as the ambulances, fire trucks, shippers, mining transports, and all the other things needed to traverse a lunar, Martian, or other surface far better if we were not along for the ride—and underwater autonomous vehicles would obviously not require humans in exploring Europa's or Ganymede's subsurface seas or Titan's liquid ethane lakes.

For instance, take perhaps the most consequential discovery a space program could make: the discovery of life not native to Earth. If a robot car equipped with drills and sensors does not make the first such discovery of extraterrestrial life by digging on Mars, an unmanned underwater vehicle may actually be the first to discover life when it dives into Europa's ocean (or Ganymede's or Titan's).

But perhaps there are other ethically pressing space issues that would require a human presence. Or would it be too risky?

23.2 Safety and Risk

23.2.1 *The Perils of Space and Risk Assessment*

Scenario: Imagine you are an astronaut on the initial SpaceX mission to Mars. The spaceship has suffered a strike from a small asteroid and is losing fuel. Internal

shutdowns have slowed the loss by sealing off most egress points, but a tiny leak remains that apparently only a spacewalk can repair. It may be that the repair could be done from inside the ship, but to determine that means discussions with Mission Control, involving a time delay—as the fuel continues to escape, imperiling the entire crew and mission. But a spacewalk is hazardous, as the field of micrometeors that caused the leak may still pose a risk to an astronaut outside the ship in a thin suit. Also, space weather monitors indicate a solar flare has occurred that may blast an astronaut outside the ship, wearing only a spacesuit for protection, with an unusually high dose of radiation. A spacewalk repair, even if successful, may be a suicide mission for the astronaut performing it. Should a spacewalk repair be attempted immediately, or should the crew delay, awaiting consultations with experts at Mission Control? Who (and how) will determine which risky activity should be undertaken? (Of course, having a robot along that could do the job would be much safer for the human crew. Even safer would be no human crew at all.)

The types of risks astronauts face run from the mundane but potentially lethal (space radiation, both from the sun and cosmic rays) to the unusual, even bizarre, e.g., will Islamic extremists disrupt a space launch to enforce a fatwa (Mars One 2015)? NASA knows that risk mitigation is a crucial issue—after all, it has lifeboats. Could it be that the risks for humans in space are always too great, so only robots should ever go?

23.2.2 Bioethics, Robots, and Risk-Benefit Analysis

Space-based medical concerns often focus on the health hazards posed by space radiation and bone and muscle loss in microgravity. NASA also studies psychological dangers, including even suicidal intentional sabotage of the mission; this is far less a worry for robots. Despite fictional depictions like *2001: A Space Odyssey*, it's unlikely a future HAL will ever intentionally decide to kill its human crew.

The dangers are far more likely to come from space itself: the sobering reality is that we still don't know if it's even possible for humans to survive indefinitely in space. NASA's Human Research Program has just finished one of the first semi-controlled trials, termed the "One-Year Mission" (NASA 2015), in which astronaut Scott Kelly spent a year in space on the International Space Station (ISS), while his identical twin brother was being monitored back on Earth. Kelly's official remarks to the House Committee on Science, Space and Technology asserted the "loss of bone and muscle, vision impairment and effects on our immune system," among other problems (Walker 2016).

And there's more bad news: for instance, while the sample sizes remain small, the existing evidence implies that even relatively brief travel outside Earth's magnetosphere is extraordinarily deleterious to one's long-term health. The Apollo astronauts were in a high-radiation environment for merely a matter of days. Yet

they now suffer approximately 478% greater mortality from cardiovascular disease than astronauts that never flew and 391% greater mortality than astronauts who only ascended to low-Earth orbit (like the ISS) and remained protected by Earth's magnetic field (Seemangal 2016).

So given the alternative of sending only robots, how shall we assess what constitutes an acceptable risk for using human astronauts instead? All of the following have been bruited as methods for determining (un)acceptable risk (Abney and Lin 2015):

Good-faith subjective standard: Under this standard, it would be left to each individual astronaut to determine whether an unacceptable risk exists. But the idiosyncrasies of human risk aversion make this standard problematic; perhaps the most thrill-seeking or suicidal risk-takers will become the first deep-spacefaring astronauts, with little regard for any rational assessment of risks and benefits. Further, what happens in a crew when some members disagree about the acceptability of the risk of some operation (say, a spacewalk to repair a communications antenna during a radiation event)?

The reasonable-person standard: An unacceptable risk might be simply what a fair, informed member of a relevant community believes to be an unacceptable risk. But what is the relevant community—NASA? Will NASA regulations really suffice for the difficult-to-predict vagaries of risk assessment during a manned spaceflight that may last six months or longer?

Objective standard: Assessing a risk as (un)acceptable requires evidence and/or expert testimony. But the “first-generation problem” remains: How do we have evidence for an unacceptable risk unless some first generation has already suffered from it? For the first spacefarers to deep space or Mars, such a standard appears impossible to implement.

And even if we decide which standard to use, there remain other crucial issues that befuddle risk managers. For example, should we believe that we should handle a first-party risk (to myself) in the same way that we handle a third-party risk (that I pose to someone else)? This partially (if not entirely) maps onto the distinction between voluntary, involuntary, and nonvoluntary risks: a first-party risk I consciously choose is voluntary; a third-party risk I force onto someone against his or her will is involuntary. But an involuntary risk may be either first- or third-party if neither I nor the other person knows of the risk, yet undergoes it anyway.

And there are numerous other risk issues, e.g., statistical versus identifiable victims, risk to future generations, and the “non-identity problem” (Parfit 1987),

and related issues concerning the ethics of reproduction in space (Lin and Abney 2014). Should astronauts be sterilized before liftoff? Or should we demand unisex crews?

All of these risks disappear if there are no humans, only robots, in space. But one additional consideration may justify attempts to send humans, and not just robots, into space: colonization. But why?

23.2.3 *A Wild Card: Existential Risk?*

The concept of existential risk refers to a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential (Bostrom 2002). For instance, suppose we discover a killer asteroid too late to stop its impact on Earth or fall prey to any number of other potential calamities that could imperil civilization. If a permanent, sustainable colony off-world already existed (as remote as that prospect currently appears), then humanity could survive such a cataclysm on Earth. Accordingly, utilitarian existential-risk theorists hold that colonization efforts might be worthwhile, even if the odds of success are minuscule.

And for deontologists, the idea that “one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation” is at least a plausible *prima facie* (and perhaps absolute) duty; such a survival principle appears to be required for any viable ethics (Abney 2004). So sending humans, not just robots, into space may be crucial for decreasing existential risk. Is there any hurry, though? My final argument suggests that there is.

23.2.4 *The Interstellar Doomsday Argument*

To understand the Interstellar Doomsday Argument, we first need some background in probability theory. First, let’s introduce the Self-Sampling Assumption (SSA): “One should reason as if one were a random sample from the set of all observers in one’s reference class” (Bostrom and Cirkovic 2011, 130). Using this SSA, I can now explain how robots, space colonization, and human extinction are linked.

First, a data point: our first robotic envoy to the stars, *Voyager 1*, entered interstellar space in August 2012 (Cook and Brown 2013). Next, apply the SSA: Assume that you are a random observer as a member of a species that has achieved interstellar travel. As of the publication of this text, it has been about five years since we became an interstellar civilization. If one reasons on this basis, there is a 95% chance that our future as a species with interstellar probes will last only between 47 more days and 195 more years. How did I come up with that (presumably alarming) answer?

Here's the calculation: Call L the time already passed for the phenomenon in question (for interstellar robots at time of publication, it's been five years). How much longer will there be such robots? Gott's delta t argument asserts that if there's nothing special about one's observation of a phenomenon, one should expect a 95% probability that the phenomenon will continue for between $1/39$ and 39 times its present age, as there's only a 5% possibility your random observation comes in the first 2.5% of its lifetime, or the last 2.5% (Gott 1993).

So as of the publication of this book in 2017, it's 95% probable that our time left as a civilization that sends interstellar robots is between $L/39$ (for $L = 5$ years, that's 47 more days) and $39L$ (195 more years). If we're still around and the reader makes this observation in 2021 or 2024, adjust the math accordingly. And understood correctly, the math offers even gloomier news: it also means a 75% chance we will go extinct (as robotic interstellar signalers) within $3L$ —the next 15 years.

The reader might find this absurdly pessimistic. Could a simple random observation about the length of time we've had robots leaving the solar system really imply Doom Soon? Actually, yes; and connecting Fermi's paradox, the Great Silence, the Great Filter, and our interstellar robots may shed light on this reasoning. Fermi's paradox refers to the question that Enrico Fermi first raised about aliens—namely, "Where is everybody?" This connects to the problem David Brin (1983) termed the "Great Silence": if aliens exist, why don't we see clear evidence of their presence in the cosmos?

We can understand Fermi's paradox and the Great Silence by using the Drake equation," understood as follows: the number of detectable alien civilizations currently in the galaxy, N , is given as $N = R^* \cdot f_p \cdot n_e \cdot f_i \cdot f_c \cdot L$ (SETI 1961).

Our understanding of the first three variables is improving thanks to recent discoveries in astronomy about the rate of star formation and the discovery of Earth-like exoplanets (e.g., Seager 2016). But the last four, biological factors in the Drake equation are as yet merely guesswork; we have only one life-bearing planet in our sample thus far, Earth.

Robin Hanson's (1998) explanation of the Great Silence is called the "Great Filter." It implies that one (or more) of the as yet unknown variables must have a value very close to zero. Recent discoveries indicate that none of the first three factors are close enough to zero to explain the Great Silence, so the Great Filter must lie among the biological factors. Perhaps either the fraction of planets upon which life evolves (f_i), or the fraction with intelligent life (f_c), or the fraction with detectable communications (f_e) is nearly zero. If so, this is good news for humankind: the Great Filter is in our past. We may be a complete evolutionary fluke, but that would have no direct implications for the duration of our future.

But if many past civilizations have arisen in our galaxy and developed technology that could be detected across interstellar space, then L is the factor that drives the Great Filter, and L must be very, very small—meaning that whenever previous

alien civilizations ascended to our level of technology, they became undetectable very, very quickly. That is, the Great Filter is in our future. The most plausible way to render our civilization undetectable very soon is, of course, human extinction.

Most thinkers, contemplating the Great Silence and *L*, consider how long we will send our radio signals into space or how long in-person journeys to the stars will endure. But sending humans across interstellar space is at best extraordinarily difficult, and may be impossible; and both our TV and radio signals are intermittent (in any particular direction), and those that spread in every direction are particularly weak. Hence, aliens with similar technology that are more than a few tens of light years away, using our test of a multiply confirmed and clearly non-random signal, would have trouble (for now) making an unambiguous detection of our civilization.

But in considering the Great Silence, the Great Filter, and our future, there's another form of contact that's much more practical and longer-lasting—our robotic spacecraft. So our test for understanding *L* should actually be our interstellar robots, which have escaped the surly bonds of our sun and ventured out into the galaxy. For if other past civilizations in the Milky Way's 13-billion-year history did send probes, just as we now have, it seems (overwhelmingly) likely that some of those probes would be here by now.

That becomes a near certainty when we contemplate von Neumann probes, i.e., probes capable of self-reproduction. A space-borne von Neumann probe could be programmed, upon arrival at its target destination, to use materials found *in situ* to produce copies of themselves and then send those copies on to other stars. Plausibly, such probes would have emissaries throughout the galaxy in a cosmically brief interval: sending out one von Neumann probe to a single alien solar system (say, Alpha Centauri) could result in its making two copies of itself and sending those probes to two new systems; those daughter probes could then go on to visit four additional systems; and so on.

Even assuming a relatively slow rate of probe reproduction and slow flight times, we run into a mathematical near certainty: within a few million to a few hundred million years at most, every solar system in the galaxy would have at least one von Neumann probe within it. With a technically feasible average speed of $c/40$, the process of investigating every stellar system in the galaxy would take about 4 million years (Webb 2002, 82). Such a timeframe is less than 1/3,000th of the time the Milky Way has been in existence. If the galaxy is not saturated by robotic probes from ancient alien civilizations, it remains implausible to think the reason is that they need more time to get here.

Now, one might protest that we know more—for there is not just one, but already five spacecraft on escape trajectories from the solar system, of which *Voyager 1* was merely the first. Plus, there are three used rocket motors on interstellar trajectories, which if encountered would also be convincing evidence to

aliens of our civilization (Johnston 2015). But in fact that reinforces the point; there is no reason to believe this moment in time is privileged with respect to our robotic interstellar probes. We reached the threshold five years ago, and there's no sign we're going to stop. But, clearly, it has stopped (or never started) everywhere else in the galaxy; that is the message of the Great Silence.

And let's be clear about the technology: sending robots to the stars is vastly easier than having humans colonize Mars, much less any other planet or moon. If we become incapable of sending probes to other stars very, very soon, then presumably we humans will be unable to escape the earth. And if we cannot escape the earth, then sooner or later we will go extinct. (My bet is sooner.) So this argument should reinforce a sense of urgency: if humans are to escape becoming just another species to go extinct on the earth, whether through external calamity or self-inflicted wounds, we had better get humans, and not just our robots, off-planet.

23.3 Conclusion

If my argument is successful, then we have reasons to use robots, not humans, for almost all the purposes that people have used to justify spaceflight. Accordingly, it may in fact be immoral to use humans instead of robots for exploration, construction, mining, communications, security, and all the other morally legitimate uses of space. Well, with one exception: having human astronauts can be justified for the purpose of colonization, in order to reduce existential risk.

Unfortunately, in the short to medium term, colonization efforts are almost guaranteed to fail. That is in part because our space policies are more concerned with vanity projects like having humans (re-)plant a flag on the moon or Mars than with building a sustainable off-world colony. The solution, it turns out, still involves our robots: in addition to enabling our global communications and commerce, they also need to build a second home for us.

Until spacefaring robots can create a permanently habitable ecosystem for us by terraforming the moon or Mars or Venus (or build a suitable habitat elsewhere in deep space), and until human spaceflight technology advances enough for it to be highly probable that such missions will end not with dead or dying astronauts eking out a few days on Mars, but with thriving colonists able to set up a sustainable civilization, it seems morally clear that we should let our robots do our exploring, mining, fighting, and all other deep-space work for us.

Works Cited

- Abney, Keith. 2004. "Sustainability, Morality and Future Rights." *Moebius* 2 (2). <http://digitalcommons.calpoly.edu/moebius/vol2/iss2/7/>

- Abney, Keith. 2012. "Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed." In *Robot Ethics*, edited by Patrick Lin, Keith Abney, and George Bekey, ch. 3. Cambridge, MA: MIT Press.
- Abney, Keith and Patrick Lin. 2015. "Enhancing Astronauts: The Ethical, Legal and Social Implications." In *Commercial Space Exploration: Ethics, Policy and Governance*, edited by Jai Galliot, ch.17. New York: Ashgate.
- B612 Foundation. 2016. <https://b612foundation.org/sentinel/>.
- Bell, Trudy and Tony Phillips. 2008. "A Super Solar Flare." *NASA Science News*, May 6. http://science.nasa.gov/science-news/science-at-nasa/2008/06may_car-ringtonflare/.
- Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9 (1). <http://www.nick-bostrom.com/existential/risks.html>.
- Bostrom, Nick and Milan M. Cirkovic. 2011. *Global Catastrophic Risks*. New York: Oxford University Press.
- Brin, G. David. 1983. "The 'Great Silence': The Controversy Concerning Extraterrestrial Intelligent Life." *Quarterly Journal of the Royal Astronomical Society* 24 (3): 283–309.
- Bruton, Dan. 1996. "Frequently Asked Questions about the Collision of Comet Shoemaker-Levy 9 with Jupiter." <http://www.physics.sfasu.edu/astro/sl9/comet-faq2.html#Q3.1>.
- Cook, Jia-Rui C. and Dwayne Brown. 2013. "NASA Spacecraft Embarks on Historic Journey into Interstellar Space." NASA, September 12. https://www.nasa.gov/mission_pages/voyager/voyager20130912.html.
- Do, Sydney, Koki Ho, Samuel Schreiner, Andrew Owens, and Olivier de Weck. 2014. "An Independent Assessment of the Technical Feasibility of the Mars One Mission Plan—Updated Analysis." *Acta Astronautica* 120 (March–April): 192–228.
- Economist*. 2016. "Flight Fantastic," August 20. <http://www.economist.com/news/science-and-technology/21705295-instead-rewiring-planes-fly-themselves-why-not-give-them-android>.
- Gott, J. Richard III. 1993. "Implications of the Copernican Principle for our Future Prospects." *Nature* 363 (6427): 315–19.
- Hanson, Robin. 1998. "The Great Filter—Are We Almost Past It?" George Mason University, September 15. <https://mason.gmu.edu/~rhanson/greatfilter.html>.
- Harrison, Todd. 2015. "Will the F-35 Be the Last Manned Fighter Jet? Physics, Physiology, and Fiscal Facts Suggest Yes." *Forbes*, April 29. <http://www.forbes.com/sites/toddharrison/2015/04/29/will-the-f-35-be-the-last-manned-fighter-jet-physics-physiology-and-fiscal-facts-suggest-yes/#13242e871912>.
- Johnston, Robert, comp. 2015. "Deep Space Probes and Other Manmade Objects Beyond Near-Earth Space." Robert Johnston website. <http://www.johnstonsarchive.net/astro/awrjp493.html>.
- Lin, Patrick. 2006. "Space Ethics: Look Before Taking Another Leap for Mankind." *Astropolitics* 4 (3): 281–94.

- Lin, Patrick. 2012. "Introduction to Robot Ethics." In *Robot Ethics*, edited by Patrick Lin, Keith Abney, and George Bekey, ch. 1. Cambridge, MA: MIT Press.
- Lin, Patrick and Keith Abney. 2014. "Introduction to Astronaut Bioethics." *Slate.com*. http://www.slate.com/articles/technology/future_tense/2014/10/astronaut_bioethics_would_it_be_unethical_to_give_birth_on_mars.html.
- Lovett, Richard. 2011. "What If the Biggest Solar Storm on Record Happened Today?" *National Geographic*, March 4. <http://news.nationalgeographic.com/news/2011/03/110302-solar-flares-sun-storms-earth-danger-carrington-event-science/>.
- Mars One. 2015. "Mars One's Response to the Fatwa Issued by the General Authority of Islamic Affairs and Endowment." <http://www.mars-one.com/news/press-releases/mars-ones-response-to-the-fatwa-issued-by-the-general-authority-of-islamic>.
- Myers, Meghann. 2015. "SECNAV: F-35C Should Be Navy's Last Manned Strike Jet." *Navy Times*, April 16. <https://www.navytimes.com/story/military/2015/04/16/navy-secretary-ray-mabus-joint-strike-fighter-f-35-unmanned/25832745/>.
- NASA. 2015. "One-Year Mission | The Research." NASA website, May 28. <https://www.nasa.gov/twins-study/about>.
- Nozick, Robert (1974). *Anarchy, State, and Utopia*. New York: Basic Books.
- Parfit, Derek. 1987. *Reasons and Persons*. Oxford: Clarendon Press.
- Plait, Phil. 2007. "Why Explore Space?" *Bad Astronomy*, November 28. <http://blogs.discovermagazine.com/badastronomy/2007/11/28/why-explore-space/>.
- Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press.
- Rincon, Paul. 2013. "Dinosaur-Killing Space Rock 'Was a Comet.'" *BBC News*, March 22. <http://www.bbc.com/news/science-environment-21709229>.
- Seager, Sara. 2016. "Research." <http://seagerexoplanets.mit.edu/research.htm>.
- Seemangal, Robin. 2016. "Space Radiation Devastated the Lives of Apollo Astronauts." *Observer*, July 28. <http://observer.com/2016/07/space-radiation-devastated-the-lives-of-apollo-astronauts/>.
- SETI Institute. 1961. "The Drake Equation." <http://www.seti.org/drakeequation>.
- Veruggio, Gianmarco and Keith Abney. 2012. "Roboethics: The Applied Ethics for a New Science." In *Robot Ethics*, edited by Patrick Lin, Keith Abney, and George Bekey, ch. 22. Cambridge, MA: MIT Press.
- Walker, Hayler. 2016. "'Space Travel Has 'Permanent Effects,' Astronaut Scott Kelly Says." *ABC News*, June 15. <http://abcnews.go.com/US/space-travel-permanent-effects-astronaut-scott-kelly/story?id=39884104>.
- Webb, Stephen. 2002. *If the Universe Is Teeming with Aliens . . . Where is Everybody?* New York: Copernicus Books.

Book chapter:

Space War and AI

Abstract: New technologies, including AI, have helped us begin to take our first steps off Earth and into outer space. But conflicts inevitably will arise and, in the absence of settled governance, may be resolved by force, as is typical for new frontiers. But the terrestrial assumptions behind the ethics of war will need to be rethought when the context radically changes, and both the environment of space and the advent of robotic warfighters with superhuman capabilities will constitute such a radical change. This essay examines how new autonomous technologies, especially dual-use technologies, and the challenges to human existence in space will force us to rethink the ethics of war, both from space to Earth, and in space itself.

Introduction

The ongoing debate over lethal autonomous weapons systems (LAWS) has for now made distinctly terrestrial assumptions; the robots doing the fighting and killing, whether on land, under the sea, or in the air, have always been assumed to be part of terrestrial forces. But as military planners have long understood, space is 'the ultimate high ground'; and despite the Outer Space Treaty, there is an increasing push towards weaponizing space. Inevitably, that push will involve LAWS, as Lt. General David Thompson, the Vice Commander of Air Force Space Command, made clear in his keynote address at the International Society of Military Ethics (Thompson 2019). For it turns out that all the reasons for adopting LAWS in terrestrial contexts, including problems of latency and accuracy, as well as other ways in which human control is suboptimal, apply to an even greater degree in space. Such weapons systems could easily also have non-military purposes in space, making the ethics of their dual-use deployment even more complicated. This essay investigates these issues and introduces some basic guidelines as to how to think about the ethical issues involved.

1. What is AI?

What is AI? The most famous attempt to connect it to human intelligence was Alan Turing's (1950) "imitation game", better known as the Turing Test, which involves a machine demonstrating human-level linguistic performance that can deceive a judge into thinking they are conversing with another human.

But of course, we use what we term 'AI' for far more than deceptive chatbots.

1. Recent, influential definitions of AI and its goals

So, AI is now variously defined; Marr (2018) lists 6 different definitions currently used by leading AI companies. For example, the Sony Corporation (2018) ethics guidelines offer the following definition: "'AI' means any functionality or its enabling technology that performs information processing for various purposes that people perceive as

intelligent, and that is embodied by machine learning based on data, or by rules or knowledge extracted in some methods.”

A common thread can be discerned through the various definitions, though: Vallor and Bekey (2017) posit the real goal of most AI research is “systems that can emulate, augment, or compete with the performance of intelligent humans in well-defined tasks.” Russell and Norvig’s (2002) textbook *Artificial Intelligence: A Modern Approach* further disambiguates this approach of seeing AI as something that attempts to “emulate, augment, or compete with the performance of intelligent humans” by providing different ways of defining AI. These possibilities arise from defining AI in terms of its *telos* along two axes; is an AI trying to match or surpass human abilities in rational thinking, or in the sophistication of its behavior? Accordingly, should we assess AI according to how it thinks, or by what it does?

1.1 Artificial versus natural intelligence?

For our purposes in thinking about AI and space war, a useful distinction is between ‘natural’ and artificial intelligence. Per Russell and Norvig, the idea of intelligence seems to involve goal-oriented, problem-solving behavior that, even if not ideal, can at least match some human capabilities in thinking or acting. Presumably, natural intelligence evolved to solve the problems of evolution, most generally the need to survive and reproduce, by having (to paraphrase, *mutatis mutandis*, Popper on the scientific method) ideas about survival strategies die in our stead. Those general goals then resulted in various sub-goals as various evolutionary mechanisms themselves evolved, until natural human intelligence became capable of being unconstrained from those original goals, free to seek new goals of its own, even ones antithetical to survival and reproduction. As we began to conceive ways of testing our ideas abstractly, without risk to our actual lives, the idea of an abstract intelligence running on something other than our human bodies began to make sense.

Following this line of reasoning, we can then define an AI as a goal-oriented, problem-solving thinking process, with at least some human-level (or better) capabilities, that arose artificially, not naturally. An analogy may help: the ability to fly naturally evolved in response to certain selection pressures, and created in birds a process that involves muscles, light bones, and wings flapping. Humans have created artificial flying machines, but did not create them by simply copying nature; they achieved the same goal of flight without using bones, muscles, or even flapping wings. So, an AI is an artificially constructed attempt to achieve cognitive goals, but may do so through means that are entirely different from our understanding of natural human cognitive processes. For example, current AIs may use propositional symbolic logic, evolutionary algorithms, Bayesian inference, or other non-biological approaches. Or, an AI can use an approach designed to mimic the human brain, such as a neural network approach. Typically, this uses artificial “neurons” that continuously compare their output calculations to a desired outcome, and then update the strength of the connections between “neurons” to “reinforce” those that seem useful. The goal of producing an artificial general intelligence by ‘whole brain emulation’ would take such an approach to its logical extreme.

1.2 AI: questions of moral status

A crucial issue in space ethics is moral status: what kind of value should non-human life or technology have? Denying speciesism (Singer 1974) does not solve the issue. Nonhuman alien persons, like the Star Trek characters Worf or Spock, would clearly deserve intrinsic moral consideration. But what of alien bacteria, or even fishlike creatures, in the oceans of Europa? Numerous options on the scope of intrinsic moral value have been defended (Lupisella 2010) including all life (biocentrism), or, indeed, everything in the universe (cosmocentrism). Other common views include rationality/sapience as crucial (ratiocentrism), or sentience—the capacity for conscious experience, of pleasure and pain—termed ‘sentientism’.

Two arguments for ratiocentrism over sentientism directly involve technology: first, the ‘Zombie/ robot argument’ (Abney 2019): suppose philosophical zombies could exist (beings with human-level complexity of behavior, but no consciousness). They reason capably enough to demonstrate moral behavior and be deemed worthy of continued existence, passing one kind of moral Turing test (Sparrow 2012), yet experience no pleasure/ pain.

Second, the Wireheading argument (Abney 2019): suppose a superintelligent AI believes sentientism, and so refashions Earth into “hedonium”: all remaining material objects merely experience unending, unknowing pleasure, until they cease to exist. This eventuality constitutes a ‘malignant failure mode’ for superintelligence (Bostrom 2014). Why a failure? Because a world with only sentience is not a moral paradise, but instead a world in which morality has disappeared, as all beings capable of it have ended. This seems a *reductio ad absurdum* argument against sentientism (Abney 2004).

The belief in sentientism may stem from a common confusion between an entity having intrinsic value (which by definition requires no extrinsic, external relationship) and final value (valuable as an end in itself, not as a mere means to some further end) (cf. Korsgaard 1983). Intrinsic value requires final value, but the converse is false. If the existence of morality requires moral responsibility (Abney 2019), then only morally responsible agents have intrinsic value, but many other things could have final value. Hence, all other moral value depends on what agents should value—whether it is purely instrumental (like an asteroid valued for its mineral wealth), or a final value, like taking pleasure simply in the view of Saturn’s rings, regardless of whatever other use they may be.

AIs, by definition, can certainly reason; they are defined as an attempt to achieve cognitive goals. Do they thus have intrinsic moral status? My answer: not yet, for they as yet lack the crucial requirement of agency and full autonomy (for more detail, see (Abney 2012, Verrugio and Abney 2012)). There is an immediate implication: AIs and their embodiment in robots (that lack agency) have a practical moral superiority over humans for almost all our near-term legitimate purposes in space. As long as AI and robots have no intrinsic moral status, there is no inherent wrong in using them as a means to explore, mine, and even help humans colonize alien locales; and of course, to fight any just wars in space. It may even be crucial that they have no sense of their own self-importance, no drive to preserve themselves at the expense of harm to others. For obeying the dictates of just war theory (JWT), it will help if AIs/ robots are willing to

sacrifice themselves to avoid harm to civilians. And further, suppose we encounter alien persons, or have as a final value unspoiled alien landscapes, flora, or fauna; it's best if the first representatives of humanity to encounter them are creatures which are morally dispensable—our AI/ robotic scouts and ambassadors and warriors, our proxies and advocates for human civilization. It may turn out to be very important indeed to make a good first impression, to be willing to sacrifice our creations for our would-be friends; as opposed to making them our enemies. In fact, this willingness could become a matter of existential importance; an argument to be continued in the final section.

1.3 Narrow AI versus AGI

Given that an AI is an artificially constructed attempt to achieve cognitive goals, another distinction is needed: between narrow AI and general AI (AGI). A narrow AI matches or exceeds human capabilities at attaining a specific cognitive goal; say, winning at chess, or calculating the product of two 17 digit primes. But while current AI chess programs can beat the human world champion, they are helpless at telling you where the closest gas station is. So, an AGI is an AI that has human-level or better capabilities generally – ideally, for all possible cognitive goals. Narrow AI, as seen, is already commonplace, whereas AGI remains for now a dream (or nightmare) of AI research. We could imagine such an AGI being better than human at speed – it thinks and attains all cognitive goals as fast or faster than humans; or, a mere collective superintelligence would think like humans, but could in parallel have trillions of human-equivalent minds working simultaneously, with accompanying superperformance. Or, there may be a 'quality superintelligence' (Bostrom 2014) that transcends mere human rationality and attains cognitive goals in ways we cannot yet imagine. Narrow AIs are commonly speed AIs, and with advances in parallel programming, are increasingly collective AIs. But narrow AIs are like *idiot savants*: brilliant in their specialization, but useless outside it. So, as humans anthropomorphize AIs and tend to (over)trust it once shown reliable in certain settings, they may well be oblivious to its limitations when asked to move outside its narrow specialization, leading at least to frustration, and perhaps disaster.

But some would think an AI 'oracle', of which we merely ask questions, poses no threat on its own; only stupid or evil human actions that result are the problem. But embodied AIs will be able to act autonomously. Clarifying those issues are next.

1.4 Robots as embodied AI and LAWS

A classic definition of a robot (Bekey 2012) is a machine, situated in the world, that senses, thinks, and acts. Like humans, robots have sensors that detect aspects of its environment and actuators that enable it to move and affect that environment. And like humans, a robot can think: that is, it has an AI.

Lethal autonomous weapons systems (LAWS) are robots designed for the purpose of being able to choose to target and kill humans. Humans can be 'in the loop', meaning a robot will never fire a weapon at a human on its own; a human must always make the ultimate firing decision. Or a LAWS can have a human 'on the loop', meaning that robot

may target autonomously, but a human can override its decision to fire; or the robot may have humans entirely ‘out of the loop’, in which both targeting and firing are autonomous, and all a human could do is deactivate the robot after the fact.

The problem of ‘latency’ helps explain the push towards LAWS with humans out of the loop: there is a lag time between a robot targeting an enemy and the human authorizing the firing, and then the robot carrying out the order. This latency may cause the robot to miss, or even worse, hit an unintended target (‘collateral damage’), or reduce its fighting efficiency and be more vulnerable itself. As the tempo and complexity of warfare increase, there will be increasing pressure to allow the AI to make the targeting and firing decisions itself. That will be particularly true in contexts in which the latency is great, or in which it is difficult or impossible for human controllers to participate. Space is perhaps the most obvious such arena for such considerations: the distances are vast, the needed electromechanical linkages between operator and robot are difficult to achieve and particularly fragile even when achieved; and it ranges from unrealistic to simply impossible for human operators to be in the battlespace or easily intercede during the warfighting. Accordingly, the pressure will be immense to deploy LAWS in space war. Could that be ethically permissible? To answer that, we need to examine LAWS in the context of just war theory.

2 What is traditional JWT?

2.1 *Jus ad bellum*, *jus in bello*: the independence thesis?

Just war theory traditionally treats the relationship between the justice of declaring war, *jus ad bellum*, and the justice of how the war is actually fought, *jus in bello*, as one of complete *independence*: the justice of a (civilian) government’s decision to wage a war is one issue, and the questions of the moral boundaries of the ways in which the professional military of that nation wages war is an entirely separate issue. But recent developments in the use of autonomous weapons systems problematizes such views, and their foreseeable use in space will exacerbate the issues.

Indeed, some of the most common objections to the use of autonomous weapons systems by militaries (especially by the U.S.) involve the claim that they will make war more alluring, so that decreasing the cost of war would make nations more likely to go to war unjustly. Such arguments effectively deny the independence assumption, as they use *jus ad bellum* worries to justify a *jus in bello* restriction or ban on LAWS.

An alternative view (term it the ‘mutual dependence’ or ‘interdependence’ view) allows such a connection between *jus ad bellum* and *jus in bello* concerns, and instead reasons as follows: the more morally justifiable a war is, the fewer the restrictions on how it may ethically be fought; and vice-versa, the less morally justifiable declaring a war is, the greater the restrictions should be on how it may be prosecuted. That is, the more morally justifiable waging and winning a war is, the more that can be morally justified to make sure the correct side wins it. And as the moral case for beginning or waging a war lessens, the greater the moral restrictions on how it may be waged. For instance, on this view, it might be permissible by the Allies (whose cause was clearly just) to drop nuclear weapons on Japan or Germany to stop the horror of WW2, despite such weapons clearly

violating several traditional *jus in bello* principles; whereas it might be morally impermissible to use nuclear weapons to wage, say, a preventive war against Iran (in the mere expectation of a possible future nuclear capability).

2.2 What matters for the independence vs mutual dependence (interdependence) thesis?

So, in assessing the independence vs mutual dependence (interdependence) thesis, what matters? Is it:

- The type of wars? Symmetric, asymmetric, large vs small?
- Or - types of combatants/ wars? Large/ small nation-states? Civil wars? Anti-terrorism campaigns? Peacekeeping?
- Or - perhaps the types of weapons? Lethal vs non-lethal? Ships vs planes? Ballistic missiles vs bullets? 'Rods from God' vs nukes?
- Or, who's doing the fighting? Humans, 'human in the loop' robots, or LAWS?
- And crucially, **location**: is war in space ethically bound by the same rules and reasoning as terrestrial war?

2.2.1 Relevance of LAWS to issue

LAWS affect and seemingly support a common interdependence argument: LAWS cause unjust wars, because they so greatly decrease the cost of war - the 'blood sacrifice'. But take cyberconflicts: do they reinforce such interdependence arguments? Was Stuxnet an act of war by the US and Israel against Iran?

Should we allow a permanent autonomous cyberdefense – i.e., permanent cyberwar? Is this a slippery slope to mutually assured destruction (MAD)? Such considerations for space warfare and LAWS raise directly the **Autonomy worry**: for war (in space), should human commanders be in, on, or out of the loop? 'Latency', as mentioned, refers to the time lag between a commander giving an order (to fire) and that order being executed. With a human in or on the loop, latency is always a problem as compared to fully automated warfare; and the greater the distance (and fragility of telecommunication connections), the greater the problem. Because of latency issues, current Space Command assumes that commanders will be out of the loop; space battles may well already be over before human commanders on the ground are even aware they are happening. So, for any chance of success in a war in space, JWT may plausibly require AI/ LAWS, due to the doctrine of military necessity (Thompson 2019).

2.3 'Ought implies can' issues and moral luck

Latency considerations are used to justify LAWS in space because success in a just war would be impossible without them. This brings up a common principle in ethical theorizing, one that entails that ethics cannot demand what is irrational for agents to do or be. In particular, it cannot expect us to perform an action or evince a character that is practically impossible for us. This is termed the '*Ought implies can*' principle – which logically requires that the morally permissible act is not only one we are capable of

doing, but the action is also one we can avoid; an agent must have an actual choice in order for moral responsibility to exist.

Accordingly, moral responsibility demands full agency – the capacity for the rational exercise of free will. The result: a proper understanding of ethics has no role for what Thomas Nagel (1979) termed ‘moral luck’. *Moral luck* is defined as: when you are held morally responsible for things over which you had no rational control; i.e., when the ‘ought implies can’ principle is violated in assigning moral responsibility. Hence, for a proper moral theory, there should be no moral luck.

Does this constitute proof of the independence thesis? Inasmuch as political leadership makes the decision to go to war and the active military does not, the ‘ought implies can’ principle would apparently imply that the active military, in determining how the war ought to be conducted, should not be held responsible for the decision to go to war, and so the independence of the two considerations would remain a moral necessity.

For humans in traditional warfare this understanding of moral luck appears to support, if not prove, the independence thesis. One does not choose what country one is born into, and so one’s birth nationality and initial citizenship cannot be other than a matter of luck. Nazi soldiers did not deserve to be born in Germany, any more than American soldiers deserved to be born in the U.S.A. They accordingly cannot be held responsible for their government’s decisions to participate in WWII, whether just or unjust. Their only moral responsibility was to obey the moral rules of *jus in bello*, with *jus ad bellum* playing no role; to insist otherwise seems to insist on moral luck.

Generalizing, typical native-born soldiers in any country cannot be held responsible for their civilian government’s decisions to go to war, and hence considerations of *jus ad bellum* cannot pertain to their moral responsibilities in the conduct of their duties as warfighters, regardless of the (in)justice of their cause.

But the reality is not so simple. First off, political leadership rarely if ever declares war ignorant of the likely strategies and tactics of its military. Secondly, in democratic countries, soldiers are typically also voters, and voters may share in the collective responsibility of going to war determined by their elected leaders. One may protest that they did not vote for the war, or even for the elected leader; but in a society in which one is free to emigrate, citizens have made the voluntary choice to remain a part of a society that wages this war.

Result: in both *jus ad bellum* and *jus in bello*, the ‘ought implies can’ principle, when taken seriously, does mean moral luck can play no role in correct attributions of moral responsibility. Accordingly, any version of act-consequentialism appears doomed as a defensible principle – to hold either soldiers or political leaders or citizens, combatants or noncombatants, responsible for what happens due to luck is to engage in moral error.

Accordingly, a better understanding of JWT must seek to minimize the role of moral luck in formulating its correct principles. That also implies that moral evaluation would ideally be agent-based, not act-based; moral evaluation is properly of the character of an

agent, in terms of the rational opportunities actually available to one, rather than the act that they did or did not do, without sensitive consideration of what capabilities – what liberties – were actually open to them.

And introducing LAWS will complicate this argument. First off, moral luck is not directly an issue for robots unless/ until they attain Kantian autonomy. Decisions to go to war are always a function of perceived odds of success, as a military option versus continued negotiation/ surrender. Hence military leadership may inform civilian leadership that using/ not using LAWS affects odds of military success (if not, there's no reason to use), and hence *jus ad bellum* considerations.

This appears to support the independence thesis in JWT, then – soldiers cannot decide whether to start wars; they only have a real choice over their conduct within them. One has no control over the country in which one was born, or whether one was drafted to serve in that country's armed forces.

But is it really true that soldiers have no control over *jus ad bellum*? Was there really such a person as the good Nazi concentration camp guard? (This type of argument is also known as the 'Reductio ad Hitlerum').

After all, if one believes their country's actions in war are unethical, then they can always vote or emigrate; or in serving in such a war, refuse to follow orders, or even surrender to the other side. Belief otherwise reinforces worries over collective responsibility problems and commits the fallacy known as the paradox of the heap – what if one person changing their mind changes the war not at all, but if many do... there is always some threshold at which additional individual acts will make a collective difference – the 'straw that broke the camel's back'.

Taking seriously the 'ought implies can' principle and notions of collective responsibility, it seems there should be a linkage between *jus ad bellum* and *jus in bello*.

2.4 Can LAWS have moral responsibility?

At a minimum, while the AI/ LAWS fighting in space cannot (yet) have moral responsibility, the officers choosing to use LAWS (or not) do have moral responsibility – they can make a meaningful choice as to the means by which they fight, even if they cannot control whether or not their government decides to prosecute the war. So, is this support for independence thesis after all, then? Because space LAWS (which command assures are a military necessity) can as yet have no moral responsibility?

To address this, we need more detail on the crucial issue: do LAWS in space or AI-enhanced astronaut warfighters undermine or reinforce the independence thesis? Dual-use issues in space exacerbate the conceptual problems. Dual-use concerns, of course, are built into almost all space-based technologies. For example, private corporations wish to build space hotels; but of course any large space base that hosted a hotel could do double duty as a launch platform for space-based missiles that could kill far faster than ground based-attacks, and may be undetectable until too late. Even the ISS only

orbits at a height of 249 miles, a distance that on the ground would be nearby enough to seriously alarm an enemy – and that does not even count the fact that, unlike Earth-based missiles, gravity would be helping, not hurting, the weapon's takeoff and impact velocity, and hence time to target. Indeed, even unarmed missiles from space could easily pack a wallop equivalent to a small nuclear weapon – the military calls them 'Rods from God' (Shainin 2006). Aggressive military planners have long termed space 'the ultimate high ground' and dreamed of using it to launch devastating preemptive attacks. Indeed, serious military strategists propose that the U.S. formalize space warfare into a separate branch of the military, like the Army or Navy, in claiming that "America Needs a U.S. Space Corps" (Smith 2017).

To be clear, space-based attacks would not even have to use kinetic force; for example, a cyberattack on satellites could disrupt an enemy's communications and weapons guidance systems; and a cyberattack on a large space-based solar power (SBSP; e.g., Mankins 2014) array could cut off power to one's enemy, or suddenly darken (or illuminate) an unexpected area of Earth's surface, or even modulate the microwaves that beam the power down to Earth into an intensity that could harm humans. Both the USA and China have worked on microwave weapon systems as part of their research on 'directed energy' weapons (Fingas 2014).

Prominent politicians have already floated such plans. For instance, in his 2012 presidential campaign, Newt Gingrich wanted the USA to have a manned base on the Moon by 2020. Later NASA feasibility studies endorsed the idea, and the Constellation program had it as one of its goals. (Whittington 2015). Dual-use issues were clear in Gingrich's pitch. He broached the idea that the Moon colony would be the '51st state' and made clear the potential dual use aspects of such a colony when he said "We will have commercial near-Earth activities that include science, tourism, and manufacturing, and are designed to create a robust industry precisely on the model of the development of the airlines in the 1930s, because it is in our interest to acquire so much experience in space that we clearly have a capacity that the Chinese and the Russians will never come anywhere close to matching." (Gingrich 2012).

But it would be costly to keep humans permanently on a lunar colony: not even counting the initial costs of getting there and construction, Phil Plait (2012) estimated simply maintaining even a small colony would take at least \$7.4 billion per year, over 1/3 of NASA's budget. Moreover, any military activities on a lunar base would explicitly violate the Outer Space Treaty (1967).

So, we need a ***red line analysis***: what constitutes war in space, as opposed to a merely civilian use? What is an (il)legitimate *dual use in space*? Is it the same as on Earth? It certainly seems not, as the simple physics of space mean that acts that would be relatively innocuous on Earth (like launching an javelin-shaped tungsten rod) pack a far different wallop when descending on an enemy from orbit at 17,000 mph: the so-called 'Rods from God'.

2.5 Scenarios, including ultimate scenario

A primary goal of this essay is to help us avoid the **First Generation** problem in space, a problem more generally for technology ethics (Abney 2019). The problem: does a first generation of users of a new technology always need to suffer or even die in order to accrue objective evidence that the risk of its use is real and unacceptable? This problem is especially pronounced in space, in which many near-term scenarios will be occurring for the first time in the near future, and the perils are still underexplored. We clearly have a need to do what is called 'anticipatory ethics'. The following scenarios are introduced as thought experiments in the beginnings of an attempt to do such anticipatory ethics for space war.

Scenario 1: A private company (e.g., Planetary Resources) proposes to bring a very large, heavily metallic asteroid in to Earth orbit, in order to lower the cost of mining it. For example, a single platinum-rich 500 meter wide asteroid would contain approximately 174 times the yearly world output of platinum, and 1.5 times the known world-reserves of platinum group metals (also including ruthenium, rhodium, palladium, osmium, and iridium). This amount could fill a basketball court to four times the height of the rim. By contrast, all of the platinum group metals mined to date in history would not reach waist-high on that same basketball court. (Planetary Resources 2016)

To make such asteroid capture, movement, and mining in Earth orbit remotely plausible and cost-effective, it likely will not have astronauts involved; instead, an AI-powered robotic tow spacecraft would capture the asteroid and then use its motors to create a delta-V so as to re-position it in Earth orbit, and then robotic miners would do the actual digging and robotic delivery vehicles would take it down to Earth (or wherever the precious metals were destined to be used).

Should a private company be allowed to give Earth another moon? Or moons? And then gradually disassemble it, with associated risks of having pieces of the asteroid plunge to Earth? What if maneuvering/ mining the asteroid used nuclear fission for propulsion? Or used fission (A-bombs) for blasting or other mining operations? What about nuclear fusion (H-bombs)?

To be clear, an expansive interpretation of the Outer Space Treaty may give private companies some leeway here. In 2015, President Obama signed the U.S. Commercial Space Launch Competitiveness Act (H.R. 2262), which recognizes the right of U.S. citizens to own asteroid resources they obtain and encourages the commercial exploration and utilization of resources from asteroids. Planetary Resources (2016) trumpets the law as a key milestone in their roadmap to access asteroid resources for commercial use in space.

Scenario 2: Bigelow Aerospace and Virgin Galactic both plan to offer space tourism, and suppose in 2020 both are ready to begin work on flights to low-Earth orbit to a commodious new private satellite, meant to be their new space hotel. Suppose their studies have led both companies to want the same orbital slot for their new space hotel, in geosynchronous orbit over Washington, D.C., with a view of the entire USA's East

Coast. How should it be decided which, if either, should occupy that orbital slot? Are there unacceptable dual-use or security concerns in occupying a slot directly over a nation-state's capital city? What if it was directly over Minot, North Dakota, home of nuclear ICBMs?

Scenario 3: Already contracting with NASA, SpaceX and Blue Origin both also want to contract their rockets to help private companies haul goods to and from orbit. What regulations should they have to abide by? What cargos should they be (not) permitted to carry?

Scenario 4: SpaceX and Mars One both have plans for Mars settlement and colonization. Should both be allowed to go forward unimpeded? Suppose they both decide on the same area on Mars for their colony – how shall it be decided who gets the prime spot? Is it simply whomever gets there first? How would their claims to ownership of land of Mars be adjudicated? Would Earth governments have any say over the government of their colonies? Should they?

Let us examine a few scenarios in more detail, including a final one that involves existential risk.

2.5.1 Asteroid mining and exploration – is it clearly civilian?

Or could asteroid mining and exploration be dual-use? The asteroid 101955 Bennu has a mean diameter of approximately 492 m. It has an Earth-crossing orbit, and has an estimated 1-in-2,700 chance of impacting Earth between 2175 and 2199, which would result in immense devastation. It will approach within 460,000 miles of Earth on 23 September 2060 (NASA 2019).

Currently NASA has a probe, the OSIRIS-Rex, which is investigating the asteroid. On 18 June 2019, NASA announced that OSIRIS-REx managed to capture a picture shot at a distance of a mere 0.4 miles (0.64 km) above Bennu's surface. This mission further plans to land on Bennu and take a sample of the asteroid regolith. The sample return to Earth is planned to land in 2023.

It would take only a small bump in Bennu's trajectory to turn its near miss in 2060 (or even earlier) into a collision with Earth, or with any satellite or other space platform in near-Earth orbit. NASA has no public plans to weaponize Bennu – but they could. Should they be allowed to make such plans? And of course, America and NASA are not the only space organizations that could engage in such asteroid retargeting. Can we be sure the space programs of China, Russia, or even India will not make such plans? Could a space conflict between Russia and a NATO nation, or China and Taiwan, escalate from a commercial to a military engagement?

2.5.2 Could tourist spaceflight be dual-use?

Could Virgin Galactic or another company dedicated to giving joyrides to space for extremely rich tourists be a dual-use technology? Some folks at the Pentagon certainly think so; they reportedly planned for the same space planes that take civilians joyriding to also transport UAVs or even human troops to a distant battlefield quickly.

In theory, SUSTAIN (an acronym for "Small Unit Space Transport and Insertion" (Axe 2014)) could deploy forces from the USA to anywhere in the world within two hours. Flying at sub-orbital altitudes, SUSTAIN theoretically would be invulnerable to enemy air defenses, and could avoid violating the national airspace of countries bordering the war zone.

SUSTAIN was supposed to be incognito: disguised as part of a venture for lifting tourists into space, like Virgin Galactic. The dual-use would be simple: to change from space tourism to war, simply switch out the passengers and retarget the coordinates. But (officially, at least) SUSTAIN is not being developed, so shock troops descending from space is on hold outside of movies like *Starship Troopers*. However, the USA does have the X-37 robotic space plane. It stays aloft for weeks to months, and officially carries no weapons, but if it did, it could launch 'rods from God' or any other weapons system that would fit in its hold almost anywhere in the world in a matter of hours. The Russians are trying to build a similar space robot with the capacity to fire nuclear weapons anywhere around the world within 2 hours—an ability they believe the X-37 may already have (Axe 2016).

Nonetheless, even the X-37 is not destabilizing in the way the SUSTAIN program would be; having AI-enhanced human astronaut or robotic troops ready to swoop down on any battlefield from space at any moment raises fundamentally different strategic concerns from those associated with mere (sub-)orbital flying robots. Generally, dual-use concerns are exacerbated by a human presence in space. After all, the civilian astronauts or tourists may secretly be spies or soldiers preparing an attack from the ultimate high ground. In addition to personally causing an attack, humans may accomplish nefarious ends by stealth: they may be able to override whatever safety measures are in place, either by an in-person cyberattack, or even by physically overriding or destroying security features of spacecraft. Humans could also engage in other kinds of subterfuge undetectable from the ground; they could reorient satellites, or change their orbit to encounter debris, and so on. These concerns would be alleviated somewhat by only having robots in space; civilian robots could still be hacked and repurposed for military attacks, but short of that, dual-use problems with civilian spacecraft are minimized when no humans, only robots, are allowed into space.

2.6 The ultimate scenario - Existential risk

Suppose alien robots attack, clearly intending to wipe out humanity (as in many sci-fi stories and movies, such as *Independence Day* (1996)) – would there be any *jus in bello* restrictions when the survival of humanity is at stake? Michael Walzer's version of traditional just war theory seems to make exceptions to *jus in bello* constraints in cases of 'Supreme Emergency' (Walzer 2006).

Many authors (e.g. Toner 2005; Cook 2007) think Walzer is wrong about this. But I believe Walzer is correct, at least in cases of existential risk, and hence the independence thesis must be false. A utilitarian argument for the conclusion may be obvious, but traditional JWT is deontological, not utilitarian. But I believe an absolute deontological principle also supports the conclusion. Here is the argument:

Deontologists routinely distinguish between *prima facie* (sometimes termed *pro tanto*) duties, which hold unless they are overridden by some other, competing duty; versus absolute duties, which hold no matter what. It can sometimes be ethical to violate a *prima facie* duty, if upholding it would violate some other, equally or even more important duty. But it is always unethical to violate an absolute duty; it takes precedence over every other obligation one could have.

So, understanding an absolute duty is crucial to ethics—if any exist. Various moral theories claim they do, but differ as to what they are. The most plausible way of justifying that a duty is absolute is to argue that it is required for morality itself to exist (Abney 2019). That is, any duty that conflicted with such an absolute duty could not be ethically required, because it would do away with ethical requirements! What kind of duty could itself be morally required for morality itself to exist? Well, both I (e.g. Abney 2017, Abney 2019) and Brian Green (2018), following the work of Hans Jonas, have argued that humanity’s continued existence is such an absolute duty.

If so, we can formulate as a corollary a plausible absolute duty: the **Extinction Principle** (Abney 2019): “one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation.” It then is an absolute duty to keep things capable of obeying absolute duties in existence. Accordingly, mitigating existential risk is an absolute duty, which wins any conflict it has with any other duty. If some foreseeable space war minimizes existential risk, then it is our highest duty, and trumps any conflicting obligation rooted in *jus in bello*.

Virtue ethics may yield a different emphasis than deontological or utilitarian approaches; it’s plausible that a virtue ethicist might insist that an obsession with decreasing existential risk to the detriment of other aspects of human flourishing betrays a flawed, even vicious character. But for the deontological *Extinction Principle* or a standard version of expected utility, decreasing existential risk trumps all other considerations.

Without invoking alien, robotic attackers, we can wonder about other humans that could pose a doomsday risk – for example, a religious cult that believes bringing about Armageddon will result in their adherents going to heaven, and has access to space-based weapon systems? Should we have any *jus in bello* restrictions on stopping/defeating them?

3 Conclusions

For AI and space war, how can we update traditional JWT to meet the novel challenges?

3.1 New principles

For now, the principal challenges coming for using LAWS from just war theory lie in the principle of distinction or discrimination, which seems largely beyond the technical abilities of AI-enabled autonomous weapons for now. To address such concerns, first, I advocate a new principle, given the military necessity (given latency issues) of LAWS in space, which I call the ‘**Discrimination Command principle**’: Identify the military official with control over LAWS (typically the CO who chooses to use LAWS in battle). Unless that military official making the decision to use LAWS (or someone of equivalent or higher rank) is willing to participate in a realistic simulation of LAWS as a noncombatant before deployment, then the actual deployment of LAWS in war is immoral. So, unless the officer in charge (up to and including generals!) is willing to assume the role of a civilian in a realistic simulation, then the military should not use LAWS.

This principle seems amply justified from a Kantian perspective, based on the categorical imperative principle of universalization and not making oneself an exception to a rule that everyone should follow.

Second, I also advocate the ‘**Metaethical principle of moral luck**’: any ethical theory that incorporates moral luck is wrong. The implications for space-based JWT are that the independence thesis is strictly false (though sometimes approximately true) – the question is the degree of (in)dependence of any factor on the others; and that may change, depending on the details of the warfighting scenario and upon one’s own (il)legitimate role in the conflict.

Accordingly, moral luck is a necessary but not sufficient condition to guide moral assessments. The existence of AI/LAWS in space does complicate interdependence, as (for now) no moral agency/ responsibility. To address this, we need a third principle: our concept of ‘military necessity’ should be widened to assess the degree of moral necessity of the just side winning (for a lasting peace and a flourishing post-war society). Merely requiring the condition of a lasting peace allows scenarios that could result from a stable but horrible totalitarian result. As an example, take WW2 or the Cold War: what was justified to ensure Nazis/ Stalinist communism/ other totalitarian regimes did not take over the world?

So, with proper safeguards, a just state could use LAWS to enforce a lasting peace and a flourishing post-war society. The problem, of course is that the military superiority of AI/ LAWS means that an unjust state could use LAWS to enforce its tyranny without possibility of revolt. Citizens armed with guns cannot defeat e.g., poisonous drone swarms or space-based weapons platforms.

3.2 Final conclusion

To stop indefinite tyranny with new tech by an unjust state – there should be no *jus in bello* restrictions on LAWS or any other technology to avoid this worst-case outcome.

In near future, just states must be willing to do whatever it takes to ensure new tech does not allow a permanent tyranny to take root. A just state could use LAWS to enforce a lasting peace and a flourishing post-war society, re Kant's dream of perpetual peace, (Abney 2012); an unjust state could use LAWS to enforce its tyranny without possibility of revolt – what good would shotguns and pistols be against an army of missile-firing drones, or swarms of tiny poisonous drones, etc.? There should be no *jus in bello* restrictions to avoid the latter outcome, including the possible necessity of AI-directed war in space. In the not so distant future, such technologies will be extant, indeed widespread; and just states must be willing to do whatever it takes to ensure such technologies do not allow a permanent tyranny to take root