The "Deepfake Defense": An Evidentiary Conundrum

By Judge Herbert B. Dixon Jr.

A long, long time ago, in the summer 2019 issue of *The Judges' Journal*, I penned a technology article entitled "Deepfakes: More Frightening than Photoshop on Steroids."[1] In that article, I described a deepfake as a video created or altered with the aid of artificial intelligence (AI) in which a person appears to do or say things that did not happen. Because deepfakes are designed to gaslight the observer, I mused that any truism associated with the ancient statement "seeing is believing" might disappear from our ethos because of the influence of deepfakes.

My most significant observation in that article was that judges and litigators were not thinking enough about how to address deepfake evidentiary issues when they show up in a few years. Well, a few years have passed, and although thought leaders have expressed opinions on how to address deepfakes in the courtroom, there is still no universally accepted solution.

Consider, for example, the hypothetical offered in that earlier article of a judge being presented with the following situation at a pretrial conference: (1) a party offers an exhibit of a cell phone video disclosed during discovery that supports the proponent's position of an agreement reached by the two parties, (2) the proponent will testify affirmatively concerning the authenticity and accuracy of the video, and (3) the opposing party will testify that he never said the words portrayed in the video. The fact that the pretrial conference provided advance notice of the evidentiary issue to the judge is fortunate, but that does not solve the problem. Indeed, in many circumstances, similar evidentiary disputes can raise their ugly heads for the first time during trial, which may require judges to call on their knowledge of the rules of evidence to solve the problem quickly.

In what has become known as the "deepfake defense," attorneys for some of the individuals charged with storming the Capitol on January 6, 2021, argued that the jury could not trust the videos because there was no assurance they were not fake or had not been altered. As of the submission of this article for publication, that defense has not been successful in any of the January 6 cases.

The problem caused by deepfakes is not far-fetched. Take, for example, a recent real-world event that made national news. In January 2024, an audio recording went viral with the voice of a high school principal making racist and antisemitic comments about students and faculty at the school.[2] Public outrage in response to the recording was immediate. The principal denied making the remarks. He received overwhelming condemnation and threats of violence. The local police stationed officers outside of the principal's home to provide security. There was no shortage of individuals in the community expressing their opinions either that they were not surprised by the recording of the principal's voice or that they believed the recording was fake. The principal was placed on administrative leave pending investigation.

Three months later, law enforcement sought and obtained an arrest warrant for the school's athletic director, whose employment contract with the school was pending termination by the principal. In addition, the police issued a subpoena to Google and ultimately traced the recording back to an email account and recovery telephone number associated with the athletic director and an IP address associated with one of his relatives. Also, the police consulted with two forensic analysts. One said the recording had traces of

AI-generated content with human editing after the fact. The other analyst said the recording was manipulated, and multiple recordings were spliced together.

The high school principal was fortunate that law enforcement obtained the expert services of forensic analysts. Imagine a similar scenario occurring in a courtroom where the proponent claims he recorded or was present when the statements were made, and his opponent denies making the statements. In most instances, neither the proponent nor the opponent will have an expert witness to testify that the evidence is real or fake. If a judge receives sworn testimony from the proponent that the evidence is a true and accurate representation of what the person said and sworn testimony from the opponent that the evidence is fake, the likely result is that the evidence will be admitted, after which the decision whether the evidence is real or fake will be left to the fact finder (judge or jury) based on the credibility of the witnesses.

Government and Industry Efforts to Identify Deepfakes

Last year, President Biden issued a comprehensive Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[3] The White House issued a fact sheet accompanying the executive order explaining that the Department of Commerce would develop guidance for content authentication and watermarking to clearly label AI-generated content. The fact sheet further explained that federal agencies will use the tools to (1) make it easier to detect deepfake content and (2) set an example for the private sector and governments around the world and make it easy for Americans to know that the communications they receive from their government are authentic. The fact sheet further reported that the president had convened at the White House seven leading AI companies—Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI—to announce that his administration had secured voluntary commitments from those companies to help move toward safe, secure, and transparent development of AI technology.[4]

Although the big tech companies that met with President Biden committed to developing technology to clearly watermark AI-generated content, there is a long way to go before watermarking features will be generally available and trusted. Some AI generators place visible labels on images made by their products. However, the labels can easily be removed. Google has a new watermarking tool that digitally embeds invisible watermarks into AI-generated images. Still, academic researchers have shown ways to compromise Google's system and others using similar approaches to watermark AI images.[5]

In the interim, we should expect that other companies and individuals will continue to develop and distribute AI applications that do not watermark or otherwise distinguish their AI-generated products. In other words, for now, the public cannot count on the equivalent of a flashing red light to warn people that they are viewing AI-created or modified content.

Proposed Evidence Rules Regarding Alleged Deepfakes

Experts and commentators have proposed at least three amendments to the Federal Rules of Evidence to guide judges in handling issues related to alleged deepfake evidence. Three of the proposed evidence rules are below. Although I leave the evaluation of each proposal to the reader, I strongly urge a review of each author's complete written explanation in support of their proposed rule.

LaMonaga's Proposed Rule

John P. LaMonaga wrote a law review article, "A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes," in which he proposed a new Federal Rule of Evidence (Fed. R. Evid.) 901(b)(11).[6] He urges a higher standard to prove authenticity than merely a witness with knowledge testifying that the exhibit fairly and accurately portrays the events or scene at issue. His proposed new Fed. R. Evid. 901(b)(11) is as follows:

> Before a court admits photographic evidence under this rule, a party may request a hearing requiring the proponent to corroborate the source of information by additional sources.

To support this stringent standard of authenticity regarding an alleged deepfake, LaMonaga contends that the traditional means of authentication (a person with knowledge attesting that the evidence is what it is claimed to be) will no longer work with deepfakes because a witness cannot reliably testify that the video accurately represents reality. Because witnesses will no longer be able to meet the standard of a knowledgeable witness by attesting that a video is a fair and accurate portrayal, LaMonaga argues that courts will need to look elsewhere to make a sufficient finding that photographic evidence is what its proponent claims it is.

Delfino's Proposed Rule

Professor Rebecca Delfino wrote a law review article titled "Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery" in which she

Judge Herbert B. Dixon Jr. is a senior judge with the Superior Court of the District of Columbia. He is chair of the *ABA Journal* Board of Editors, a former chair of both the National Conference of State Trial Judges and the ABA Standing Committee on the American Judicial System, and a former member of the ABA TECHSHOW Planning Board. You can reach him at Jhbdixon@gmail.com. Follow Judge Dixon on X (formerly known as Twitter) @Jhbdixon.

proposes that because of the danger of deepfakes, the judge (not the jury) should decide authenticity.[7] She concludes that jurors cannot be trusted to fairly analyze whether a video is a deepfake because deepfakes appear to be genuine and that a new Federal Rule of Evidence should be created to expand the court's gatekeeping function by assigning the responsibility of deciding authenticity issues solely to the judge. Delfino's proposed new Fed. R. Evid. 901(c) is as follows:

> Notwithstanding subdivision (a), to satisfy the requirement of authenticating or identifying an item of audiovisual evidence, the proponent must produce evidence that the item is what the proponent claims it is in accordance with subdivision (b). The court must decide any question about whether the evidence is admissible.

According to Delfino, if, after a hearing to determine the authenticity of the evidence, the court finds that the item is more likely than not authentic, the court should admit the evidence. The court would instruct the jury that it must accept as authentic the evidence the court has determined to be genuine. The court would also instruct the jury not to doubt the authenticity simply because of the existence of deepfakes. According to Delfino, this new rule would take the jury out of deciding authenticity and avoid the problems invited by juror distrust and doubt. Delfino says the court would address the threat of counsel exploiting juror doubts over the authenticity of evidence using the deepfake defense by ordering counsel not to make such arguments.

Grimm and Grossman's Proposed Rule

At the April 2024 meeting of the Advisory Committee on Evidence Rules, Judge Paul Grimm (Ret.) and Dr. Maura Grossman made a presentation about the evidentiary problems caused by deepfakes and proposed a new Fed. R. Evid. 901(c).[8] The proposed rule provides:

> *Potentially Fabricated or Altered Electronic Evidence*. If a party challenging the authenticity of computer-generated or other electronic evidence demonstrates to the court that it is more likely than not either fabricated, or altered in whole or in part, the evidence is admissible only if the proponent demonstrates that its probative value outweighs its prejudicial effect on the party challenging the evidence.

Grimm and Grossman contend that their proposed new Fed. R. Evid. 901(c) puts the initial burden on the party challenging the authenticity of computer-generated or electronic evidence as AI-generated fakery to make a showing to the court that it is more likely than not either fabricated or altered in whole or part. It requires the challenging party to produce evidence to support the claim that the proffered exhibit is fabricated or altered. According to Grimm and Grossman, if the challenging party makes the required showing, then the burden shifts to the proponent of the challenged evidence to show that its probative value outweighs its prejudicial effect on the party challenging the evidence.

The Advisory Committee took no action to adopt the Grimm-Grossman proposed new Fed. R. Evid. 901(c) at the April 2024 meeting. Some committee members expressed the opinion that the current rules are adequate to address the issue. Other members suggested that more judicial experience with the issue is needed as there have been few instances of judges being asked to exclude AI-generated evidence. The committee expects Grimm and Grossman to rework their proposal for future committee consideration based on the discussions at the meeting.

Final Thoughts

As technology advances, deepfakes will improve and become more difficult to detect. Presently, the general population is not able to identify a deepfake created with current technology. AI technology has reached the stage where the technology needed to detect a deepfake must be more sophisticated than the technology that created the deepfake. So, in the absence of a uniform approach in the courtroom for the admission or exclusion of audio or video evidence where there are credible arguments on both sides that the evidence is fake or authentic, the default position, unfortunately, may be to let the jury decide. ■

Endnotes

1. Herbert B. Dixon Jr., *Deepfakes: More Frightening than Photoshop on Steroids*, 58 Judges' J., no. 3, Summer 2019, https://bit.ly/3GnpNgg.

2. Paul Schwartzman & Pranshu Verma, *Baltimore Principal's Racist Rant Was an AI Fake. His Colleague Was Arrested*, Wash. Post (Apr. 26, 2024), https://bit.ly/4aV5h4h; Thomas Lake, *A School Principal Faced Threats After Being Accused of Offensive Language on a Recording. Now Police Say It Was a Deepfake*, CNN (Apr. 26, 2024), https://bit.ly/4bkySnz.

3. White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, Briefing Room (Oct. 30, 2023), https://bit.ly/4dm7LtZ.

4. Press Release, White House, Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (Oct. 30, 2023), https://bit.ly/3svukKj.

5. Gerrit De Vynck, *The AI Deepfake Apocalypse Is Here. These Are the Ideas for Fighting It.*, Wash. Post (Apr. 5, 2024), https://bit.ly/49VJtUJ.

6. John P. LaMonaga, *A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes*, 69 Am. U. L. Rev. 1945, 1984 (2020), https://bit.ly/3Qt0nmW.

7. Rebecca A. Delfino, *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery*, 74 Hastings L.J. 293 (2023), https://bit.ly/4bls3Ty.

8. Advisory Comm. on Evidence Rules, Agenda, Proposed New Rule 901(c) to Address "Deepfakes," at 18 (Apr. 19, 2024), https://bit.ly/3JFAL2g.