INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA

The views expressed herein have not been approved by the House of Delegates or the Board of Governors of the American Bar Association and, accordingly, should not be construed as representing the policy of the American Bar Association.
Invisible Threats: Mitigating the Risk of Violence from Online Hate Speech Against Human Rights Defenders in Guatemala

MAY 2019

AMERICAN BAR ASSOCIATION
CENTER FOR HUMAN RIGHTS
All around the world, justice sector personnel and human rights advocates are facing unprecedented threats in the form of frivolous criminal accusations, funding restrictions and violent threats. The ABA Center for Human Rights provides pro bono legal assistance to human rights defenders facing reprisals. In the course of this work, it has noted a correlation between online hate speech and violence against leaders of marginalized communities, particularly in conflict-affected countries.

The longstanding use of threats and disinformation to discourage and discredit the work of human rights advocates has become more sophisticated in an online world. Social media companies provide a platform for human rights defenders to share information and express opinions. These platforms have been used to organize protests that have toppled dictatorial regimes and to call for the release of political prisoners. At the same time, they are increasingly being used to target and harass human rights defenders, including journalists, environmental activists and lawyers. Ease of dissemination and difficulties with attribution can make accountability for online hate speech and incitement challenging. Most concerning are systematic, state-aligned campaigns to denigrate and indirectly threaten human rights defenders and marginalized communities. The combination of the volume and visibility of such speech is particularly toxic in countries where there is political persecution of human rights defenders and impunity for violence against human rights defenders and marginalized populations more broadly.

Guatemala is one such country. After thirty years of armed conflict which claimed the lives of an estimated 200,000 civilians, Guatemala has struggled to control parallel criminal organizations that have captured state institutions. After making progress in recent years, a crackdown on anti-corruption efforts and the independence of the judiciary has jeopardized recent gains and compromised the checks and balances necessary to ensure that the upcoming presidential elections are legitimate. Against this backdrop, coordinated, state-aligned campaigns of online disinformation and coded threats pose a unique challenge for SMCs and contribute to the greater culture of impunity for violence against human rights defenders and marginalized communities while significantly increasing the risk of violence against them.

To address these challenges, the ABA Center for Human Rights asked the International Human Rights and International Law Clinic at the University of Connecticut School of Law to examine whether the content moderation policies of SMCs are sufficient to mitigate the risk of violence from online hate speech against human rights defenders in Guatemala. This study focused on the situation in Guatemala to illustrate the need for more targeted content moderation policies and practices. It is our hope that it will provide concrete and practical recommendations to SMCs operating in this and other high-risk areas, including post-conflict settings, and support them to engage in context-specific content moderation that narrowly focuses on hate speech and coded threats that are likely to contribute to violence in the real world. Greater attention to context and ongoing dialogue between knowledgeable civil society leaders and SMCs are clearly warranted to achieve the promise of these platforms for realizing and expanding on fundamental freedoms while ensuring accountability and transparency to protect those rights.
TABLE OF CONTENTS

EXECUTIVE SUMMARY ........................................................................................................... 1

INVISIBLE THREATS: MITIGATING THE THREAT OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA .......................................................... 5

I. INTRODUCTION .................................................................................................................. 5

II. THE HISTORY OF VIOLENCE IN GUATEMALA ................................................................. 8

   A. OVERTHROWN: MILITARY DOMINANCE IN GUATEMALA ........................................... 10
   B. JUSTICE IN THE AFTERMATH OF THE ARMED CONFLICT ........................................ 12
   C. TODAY’S POLITICAL LANDSCAPE ONLINE ................................................................. 13

III. CURRENT CONTENT MODERATION POLICIES AND PRACTICES ................................. 18

   A. CONTENT MODERATION PROCESSES ON TWITTER AND FACEBOOK .................. 19
   B. CONTENT MODERATION POLICIES ON FACEBOOK .............................................. 19

       1. CREDIBLE VIOLENCE ............................................................................................... 19
       2. HARASSMENT ........................................................................................................... 21
       3. HATE SPEECH .......................................................................................................... 22
       4. COORDINATING HARM ............................................................................................ 23

   C. CONTENT MODERATION POLICIES ON TWITTER .................................................. 23

       1. THREATS OF VIOLENCE .......................................................................................... 23
       2. ABUSE ....................................................................................................................... 24
       3. HATEFUL CONDUCT ................................................................................................. 26
       4. SPAM AND FAKE ACCOUNTS .................................................................................... 28

   D. PROCESS AND SANCTIONS .......................................................................................... 28

   E. WHY ARE THE CURRENT MODERATION PRACTICES STRUGGLING TO
      PROTECT DEFENDERS IN GUATEMALA? .................................................................... 29

       1. RELIANCE ON USER FLAGGING ............................................................................ 29
       2. CONTEXT .................................................................................................................. 30
       3. NET CENTERS .......................................................................................................... 31
       4. STATUS OF THE SPEAKER ...................................................................................... 31
IV. SOCIAL MEDIA COMPANY RESPONSIBILITIES .................................................................32
   A. CODED THREATS AND RACE-BASED INCITEMENT ...........................................33
   B. STATE PERSECUTION OF HUMAN RIGHTS DEFENDERS ......................................35

RECOMMENDATIONS & CONCLUSION ..................................................................................37

V. RECOMMENDATIONS ........................................................................................................37
   A. CONTEXT-DRIVEN CONTENT MODERATION .........................................................37
      1. HEIGHTENED SCRUTINY IN PROBLEMATIC AREAS ..............................................37
      2. LOCALIZED GUIDELINES AND PERSONNEL .........................................................38
      3. MONITORING AND DE-PLATFORMING KEY ACCOUNTS .......................................38
   B. ADDRESSING POLITICAL PERSECUTION IN SMCS CONTENT
      MODERATION POLICIES ............................................................................................38
   C. IMPROVE IDENTIFICATION OF HARMFUL POSTS ................................................39
      1. VERIFIED USERS AS ENDORSED CONTENT MODERATORS ..............................39
      2. IMPROVED APPEALS PROCESS .............................................................................40
      3. ONLINE TRAINING AND AWARENESS OF SOCIAL MEDIA LITERACY ..............40
   D. ORGANIZATIONAL AND STRUCTURAL CHANGES ..................................................40

VI. CONCLUSION ................................................................................................................41
At the request of the American Bar Association’s Center for Human Rights, the Human Rights and International Law Clinic at the University of Connecticut School of Law investigated online hate speech directed against human rights defenders in Guatemala. Violence against human rights defenders in Guatemala is pervasive and on the rise, and the Center for Human Rights requested a study of whether social media companies are effectively detecting coded language that would – but for the use of euphemistic language – violate content moderation policies that prohibit the posting of hate speech online. The issue of online coordinated defamation of human rights defenders has already been identified as a subject of considerable concern by both NGOs and the Inter-American Commission on Human Rights.

The report documents a coordinated campaign on social media to defame, harass, intimidate, and incite violence against human rights defenders in Guatemala. Social media users with substantial followings, and with formal and informal connections to the government, military and a certain section of the business community, routinely characterize human rights defenders as criminals, terrorists, and communists.

Using coded speech that invokes rhetoric from the thirty-six-year Guatemalan armed conflict and genocide, private actors operating with the implicit support of corrupt elements of the Guatemalan government, often evade scrutiny under the content moderation policies of social media companies (“SMCs”).

These campaigns contribute to conditions in which physical attacks on defenders are not prevented, investigated, or punished by the state. Although an empirical analysis of the causal relationship between online hate speech and real-world violence is beyond the scope of this study, there are grounds for concern that hate speech contributes to violence against human rights defenders by fostering a climate in which attacks against defenders—invoking physical violence as well as the use of false and punitive criminal charges—can be launched with impunity.

The report focuses on speech by “state aligned” actors because harassing speech that is aligned with the state, particularly in contexts characterized by state corruption and impunity, presents greater risks to state opponents than “mere” private speech. The report uses the term “state-aligned actors” to designate individuals who have contributed to or benefited from high-level government corruption and the

---

1 This publication has not been reviewed by the ABA House of Delegates and therefore should not be construed as representing ABA policy. The views expressed herein are those of the authors alone. The report was written by Muna Abbas, Elaf Al-Wohaibi, Jonathan Donovan, Emma Hale, Tatyana Marugg, Jonathan Sykes, and other members of the Human Rights and International Law Clinic at the University of Connecticut School of Law, under the supervision of Professors Molly K. Land and Richard Ashby Wilson. We are grateful to John Murphy and his team at the University of Connecticut’s, Digital Media & Design Social Media Analytics Command Center (SMACC) for conducting the social media analytics referred to in this report.


INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA

culture of impunity that permeates Guatemala. SMCs provide platforms where these actors can quickly and widely disseminate harmful speech that facilitates human rights violations. The impact of this harmful speech is further amplified by net centers that employ fake accounts to boost messages aligned with corrupt business and political interests.

Although the report highlights areas where content moderation of speech on social media is under-inclusive by failing to address speech that undermines the safety of human rights defenders, it recognizes that over-enforcement of content moderation standards can also cause harms to freedom of expression and opinion and can limit the extent to which individuals—including human rights defenders—are able to organize and advocate for change. The definition of “hate speech” in the rules developed by both Facebook and Twitter is extremely broad, and more robust enforcement of such a standard would have a significant impact on expressive freedom. We therefore endorse the calls for greater transparency, notice, and due process in content moderation represented by the Santa Clara Principles on Transparency and Accountability in Content Moderation.4

Recognizing the risks to freedom of expression and opinion of overbroad content moderation, the report does not advocate stronger enforcement, but smarter enforcement. Content moderation policies must be more narrowly targeted to address hate speech that is associated with heightened risk of violence, which necessarily requires consideration of context. Speech that may be fairly low-risk in an environment characterized by democratic institutions and rule of law could be extremely risky in a country plagued by corruption, political violence, and inequality. Thus, this report provides guidance to SMCs in developing more narrowly tailored procedures and policies. Toward this goal, the report makes two primary recommendations.

First, SMCs must take measures to prevent their platforms from being used to target, harass, and intimidate human rights defenders, or they risk being complicit in any resulting harms and injuries.5 Human rights defenders are not currently a protected group under Facebook or Twitter’s content moderation policies. Drawing on international law that prohibits states from targeting individuals for serious harm based on their political opinion, the report recommends that SMCs adopt policies that allow them to designate defenders as temporarily protected groups in contexts like Guatemala characterized by state-coordinated or state-condoned persecution.6 Violence against defenders is also more likely in the months preceding high-profile elections, such as the Guatemalan elections in June 2019. Guatemala’s history of armed conflict, coupled with the heightened tension surrounding elections, makes it especially susceptible to political violence fomented on social media.

Second, we recommend that social media platforms take steps to augment their ability to engage in context-specific content

---

6 Inter-American Commission, supra note 2.
moderation. United Nations experts and human rights activists and scholars have called for greater consideration of context in a variety of settings, ranging from efforts to understand the range of gender-based harms experienced by women online\(^7\) to avoiding censorship of protected speech with educational or historical value.\(^8\)

Recently, the genocide against the Rohingya in Myanmar has raised with particular urgency the importance of identifying coded speech threats. For example, the Independent International Fact-Finding Mission on Myanmar noted that it was difficult to identify online hate speech against the Rohingya because of “[s]ubtleties in the Myanmar language and the use of fables and allegories.”\(^9\) Efforts to integrate context into content moderation, however, are in tension with the challenge of moderating at scale.\(^10\)

Our investigation of hate speech in Guatemala illustrates the way in which coded speech is used against human rights defenders. The same word may have very different meanings and consequences in two different countries, even when those countries share a border and a language, such as Guatemala and Mexico. SMCs must dedicate more resources to distinguishing the coded meaning of speech and understanding the differential impact of the same words in distinct political settings. Companies can increase their ability to respond to context-specific meanings by creating interdisciplinary teams that draw upon the expertise of academics, linguists, and lawyers. Such efforts should focus first on hate speech in countries with weak rule of law and a history of violent aggression against protected groups.

The report also discusses several potential strategies for greater consideration of context. SMCs should provide heightened scrutiny of content in problematic or sensitive countries, engage localized personnel and guidance, and consider the greater impact of the speaker in monitoring and possibly de-platforming high profile accounts. SMCs can also improve flagging processes to facilitate the gathering of context-specific information by utilizing verified users as endorsed content moderators, creating and implementing online and social media literacy training programs, and creating appeals processes. These recommendations, when taken together, will increase protection for human rights defenders on the platform—both online and offline.

---


INTRODUCTION

Human rights work is often dangerous, particularly in countries like Guatemala where government corruption is rampant, impunity is entrenched, and a cohesive network of corrupt actors actively undermines civil society. Human rights defenders ("defenders") in Guatemala have endured harassment and state-sponsored violence for decades, but the rise of social media platforms has enabled the rapid and widespread dissemination of these threats and intimidation. In a highly polarized society, coordinated online attacks erode defenders’ credibility, undermine their efforts to challenge corruption and impunity, and make them more vulnerable to offline attacks.

While social media harassment of defenders is not unique to Guatemala, the country’s history of violence and genocide means that this harassment is more dangerous than in many other countries. Euphemisms and ‘coded’ words such as comunista (communist) and terrorista (terrorist) elicit powerful and negative associations with Guatemala’s genocidal armed conflict. By using coded terms, state-aligned online attackers have evaded the content moderation and hate speech policies of social media platforms to harass and silence defenders.

Drawing on this case study of Guatemala, the report concludes that social media companies should further align their policies with standards articulated in international human rights and international criminal law in two specific ways. First, they should establish additional review procedures that will allow them to take better account of coded speech that contributes to a climate of intimidation and violence but which may not necessarily constitute a specific personal threat. Second, SMCs should include human rights defenders as a protected category under their hate speech/harmful content policies in countries where defenders face persecution by the state or are not protected by the government from this harm.

Two of the most widely-used social media platforms, Facebook and Twitter, have come under immense pressure recently to reevaluate their content moderation policies. These companies have publicly acknowledged their duty to prevent human rights violations—both under the United Nations Guiding Principles and as the responsible hosts of some of the most popular websites in the world—and they have taken initial steps towards establishing a more protective regime of online speech regulation. SMCs are now subject to heightened public and governmental scrutiny regarding their obligations to their billions of users and wider society.

Facebook and Twitter have the resources necessary to protect vulnerable groups on their platforms, and the technical
challenges, while significant, are not insurmountable. Facebook’s revenue in 2018 was over $55 billion—higher than the annual GDP of more than 100 countries. Twitter reported revenue of over $2.4 billion in 2017. On the basis of income alone, the economic influence of SMCs rivals that of states.

Today, content moderation teams sift through millions of posts a day, making quick decisions based on complex rules and scenarios. Respecting free expression while simultaneously removing harmful or threatening content is a delicate balancing act and this is made more difficult where the harmful or threatening nature of the content is based upon language that is highly specific to the local context, which we refer to as “coded” hate speech. Identifying the harms in euphemistic language requires sophisticated cultural knowledge.

In Guatemala, language that may appear merely insulting and belligerent, but not obviously dangerous, contributes to an ongoing climate of violence, intimidation, and impunity with its origins in a bloody and genocidal armed conflict. Physical violence and punitive prosecutions of defenders occur in an environment saturated with online social media harassment and defamation campaigns decrying them as criminals, terrorists, and “destabilizers.”

Defenders are placed in danger when they are labelled “criminals” by authority figures: due to Guatemala’s poor rule of law and culture of impunity, “vigilantism in vulnerable neighborhoods [and] the lynching of suspected criminals” is common. Prior to his indictment by state prosecutors and arrest, one indigenous defenders was “subjected to acts of intimidation and a campaign of defamation on social media.” This is of particular concern given the “alarming intensification of violence” in Guatemala as noted by the UN Rapporteur on the Rights of Indigenous Peoples.

Euphemistic speech by state-backed actors present risks to vulnerable groups all over the world. Because content moderators often have insufficient cultural and contextual experience with the local variants of globally spoken languages, the risks associated with coded calls to violence often go unnoticed. This report argues that formal comprehension of a language is not enough; SMCs must evaluate language in its local and national contexts in order to adequately identify harmful speech.

In Guatemala, defenders include environmental activists, judges and prosecutors, anti-corruption activists, indigenous rights advocates, and the staff of international agencies such as the United Nations. All of these groups are systematically targeted on social media platforms, and the harassment does not only occur online. High profile defenders

---

18 Special Rapporteur on the Rights of Indigenous Peoples, supra note 16.
19 Id.
have had their accounts hacked, their movements stalked, and photographs of them altered to suit the narratives of perpetrators. In human rights and corruption trials, victims have been intimidated and threatened by unknown individuals taking pictures and uploading their personal information online. Even in rural areas with little internet or social media usage, there is evidence that online smear campaigns can still harm Guatemalan environmental campaigners and human rights activists.

Real world violence against human rights defenders has been associated with incendiary speech. In one case, online retaliation against a defender appears to have escalated violence against activists. A local activist had been shot after leading a protest against the environmental impact of a mining project in Guatemala. The incident was never properly investigated by the police. Rather, the community leader was the subject of frivolous criminal charges that were ultimately dismissed. After the shooting, the owner of the mine posted a letter drawing on invective of the era of the armed conflict, stating that the “people who oppose us are the same type of people who supported the civil war. They want to keep the people poor, uneducated, and totally submissive.”

The Foundation Against Terrorism (FCT), a far-right group mostly composed of a number of former military personnel, also posted statements attacking the organization that had represented the community. Thereafter, an employee of a local subsidiary of the mine was filmed threatening members of the community. Eventually, protestors were violently evicted from the entrance to the mine by Guatemalan police; at least 20 people were injured and 7 were hospitalized. Video shows police dressed in riot gear throwing rocks at protestors. Although the mine employee was ultimately convicted for making threats, the online harassment of the community escalated the likelihood of the resulting violence.

The coded language documented in this report deliberately draws on tropes used by elements of the military as a counter-insurgent tactic during decades of armed conflict in Guatemala to discredit any kind of political opposition. This language includes terms such as comunista (communist) and terrorista (terrorist), as well as accusations of treason and foreign collaboration. Indeed, the same photographs with foreign officials that activists might see as protective of their security have been reposted by state-aligned actors as evidence of defenders’ alleged collusion with “foreign

20 Interview 014.
21 Interview 015.
23 AMERICAN BAR ASSOCIATION CENTER FOR HUMAN RIGHTS, TILTED SCALES: SOCIAL CONFLICT AND CRIMINAL JUSTICE IN GUATEMALA 18 (2013).
24 Interview 001.
25 Id.
26 Id.; see also Fundación Contra el Terrorismo, Denuncia de la Fundación Contra el Terrorismo, No. 5(1), at 1 (2013), https://perma.cc/5V59-EJZS.
30 Bastos, supra note 27.
destabilizers.” Ex-military officers and other individuals closely associated with the actors responsible for human rights violations during the armed conflict, now use social media platforms to accuse those who support UN-affiliated anti-corruption investigators\textsuperscript{31} of attempting to overthrow the government.\textsuperscript{32} “Real Guatemalans,” argues one, are “those who do not sell the fatherland to foreigners.”\textsuperscript{33} Such language draws connections between today’s defenders and the opposition targeted in Guatemala’s armed conflict and genocide.

Language that could be dismissed in another context as simply polemical can have serious consequences in a country still transitioning from a period of state repression and genocide. Post-conflict Guatemala is characterized by widespread corruption and weak rule of law. According to the U.S. Department of State’s Bureau of International Narcotics and Law Enforcement Affairs, “[W]idespread corruption permeates public and private institutions and exacerbates [Guatemala’s] security, governmental, and economic challenges.”\textsuperscript{34} In such an environment, pervasive social media attacks on defenders and other opposition figures contributes to a climate in which violence is condoned and perpetrators are often not charged or prosecuted.\textsuperscript{35} Defenders say that the hostility of the government toward their activities is acute and creates a hostile environment.\textsuperscript{36}

In Guatemala, killings and harassment of defenders are inadequately investigated,\textsuperscript{37} and the rule of law is among the weakest in the world.\textsuperscript{38} According to one account, impunity for violent crimes stands at 98%.\textsuperscript{39} Corruption in the legislature is common, and many citizens resolve personal disputes violently.\textsuperscript{40} In rural areas, activists who oppose local corruption are subjected to explicit and implicit attacks on social media that correlate with persecution in their communities and in some cases, height of the circumstances, without giving concessions of any nature. Guatemala is our[s,] [l]et’s recover it!”); Giovanni Fratti (@frattigiovanni), TWITTER (Sept. 8, 2018, 10:36 PM), https://twitter.com/frattigiovanni/status/1038662448440594432 (“https://youtu.be/uz3zmATnu64 no more abuses of foreigners and no more occupation of foreigners insolent!!! Guatemala is our[s,] let’s recover it!!!”).\textsuperscript{34} U.S. DEP’T OF STATE BUREAU OF INT’L NARCOTICS AND LAW ENFORCEMENT AFFAIRS, INTERNATIONAL NARCOTICS CONTROL STRATEGY REPORT 172 (2018), Interviews 005, 006, 007, 013.\textsuperscript{35} Interviews 006, 013.\textsuperscript{36} U.S. DEP’T OF STATE, COUNTRY REPORTS ON HUMAN RIGHTS PRACTICES FOR 2017: GUATEMALA (Jul. 23, 2018), https://perma.cc/P464-3X3Y.\textsuperscript{37} The World Justice Project’s Rule of Law Index ranks the rule of law for countries across the globe; it ranked Guatemala as 96 out of 126 for 2019. WORLD JUSTICE PROJECT, RULE OF LAW INDEX 2019 INSIGHTS 8, https://perma.cc/4XJJ-UF7B.\textsuperscript{38} Gutiérrez, supra note 17.


\textsuperscript{32} Ricardo Méndez Ruiz (@RMéndezRuiz), TWITTER (Jan. 23, 2019, 6:12 PM), https://twitter.com/RMéndezRuiz/status/1088258198946091009) (“Guatemala expelled Iván Velásquez, a sad caricature of dictator, and his entire structure of criminal aliens. Venezuela’s turn is now.”).

\textsuperscript{33} Ricardo Méndez Ruiz (@RMéndezRuiz), TWITTER (Sept. 23, 2018, 2:42 PM), https://twitter.com/RMéndezRuiz/status/1043978830082428928 (“President @jimmymoralesgt, the real Guatemalans, those who do not sell the fatherland to foreigners, we expect from you on Tuesday before the @UN_is one speech to the
Although Guatemala is not the only country battling impunity, it is a case study in how government-aligned actors use social media platforms to pursue long standing campaigns of persecution and human rights violations. As such, it highlights the challenge facing SMCs to limit harmful uses of their platforms, and specifically how to respond to state-aligned campaigns against defenders that rely heavily on coded speech.

This report starts by reviewing the history of political violence in Guatemala in order to understand how a vocabulary of coded speech emerged that is invisible to content moderators on Facebook and Twitter. The report then discusses the responsibilities of SMCs regarding harmful speech on their platforms and examines the existing content moderation practices used by Twitter and Facebook as those policies apply to the speech at issue. The report concludes with a series of recommendations for consideration by SMCs.

Finally, a note on the methods used in this study: Over the course of our research, our team conducted 15 interviews with journalists, human rights defenders, government officials, and social media experts from Guatemala and the United States. For interviews conducted in Spanish, a competent Spanish speaker provided translations. The names of all interview subjects have been removed to protect their privacy. Many of the social media posts and accounts included in this study were initially identified by the American Bar Association’s Center for Human Rights. Our team also monitored over the course of several months the Twitter and Facebook accounts of twelve key figures in Guatemala identified by the ABA as relevant to this project. Four of these accounts were analyzed in early 2019 by the University of Connecticut's Digital Media & Design Social Media Analytics Command Center (SMACC) using Talkwalker software. Talkwalker is a social media listening platform that gathers publicly available social and digital media conversations through licensed agreements. The platform then uses artificial intelligence and engineering methods to analyze, organize and generate visual representations of the data.

THE HISTORY OF VIOLENCE IN GUATEMALA

The challenges inherent in moderating content on social media are compounded in Guatemala, where a violent national history has created a vocabulary of coded language that can be difficult for outsiders to identify. Facebook and Twitter’s policies, designed to provide a basis for removing direct threats as well as hate speech against individuals based on their group membership, are not triggered by these coded words and phrases, which often refer to communist sympathies or threats to family values. Guatemalan defenders note that discourse and legislation critical of women, communism, and the LGBTQ community often combines with these narratives to blur their boundaries. Content moderators may not recognize the strong negative emotions that these terms evoke for many Guatemalans.

Fluency in a language does not necessarily imply fluency in the cultural nuances of

---

41 Interview 004; Inter-American Commission, *supra* note 2, ¶¶ 146, 151-53.

42 Interview 005.

43 *Id.*
expression, and to effectively moderate content and inhibit the malicious use of their platforms, SMCs must take greater steps to understand the context that lies behind a user’s speech. Until moderation strategies can better cope with coded language, the work of defenders will continue to be undermined by social media in Guatemala and state-aligned campaigns of violence and impunity will thrive.

A. Overthrown: Military Dominance in Guatemala

For almost a century after gaining full independence from the Spanish empire in 1821, the government in Guatemala swung between outright dictatorships and authoritarian presidencies. Social-democratic reforms began in 1944 at the outset of the revolutionary movement that overthrew then-dictator Jorge Ubico. The social-democrat revolutionary government created a social security system and redistributed some land to indigenous Guatemalans, among many other structural reforms.

From the north, the United States government viewed these changes with alarm through the lens of the ongoing Cold War. When some assets of the American-owned United Fruit Company were nationalized, U.S. President Harry Truman concluded that Guatemala was moving towards communism. In response to aggressive lobbying from United Fruit, Truman authorized the CIA to stage a coup to topple the Guatemalan government. The coup deposed the democratically elected president in June 1954 and installed Lieutenant Colonel Castillo Armars, a member of the right-wing National Liberation Movement as President.

Armars’s regime marked the beginning of decades of political dominance by the military characterized by virulent anti-communist rhetoric—rhetoric that cultivated an extreme fear in the population of any association with communism. This fear is still exploited by state-aligned actors on social media platforms today. Under the banner of an anti-communist crusade, Armars reversed land reforms and reinstated property to powerful, wealthy landowners. Armars created the National Committee of Defense Against Communism, which was given the broad powers to surveil, arrest, and deport suspected communists without trial. Armars’s government suppressed dissidents and actively dismantled opposition parties, both by decree and through physical violence.

Armars was murdered in 1957 and replaced through sham elections by General Ydígoras Fuentes who continued the crusade against communism. Political opponents were imprisoned or disappeared. Outraged by the corruption of the Fuentes regime, dissidents and indigenous farmers joined forces and some

45 Id.
49 Id.
51 Id. at 195.
formed leftist rebel groups. The government’s military intelligence services and security forces embarked on a violent campaign against the political opposition who were termed the “internal enemy,” which was defined in the Counter-Subversive War Manual of the Guatemalan Army as:

all those individuals, groups, or organizations that, through illegal actions, seek to rupture the established order, represented by elements that, in pursuit of the goals of international communism, pursue so-called revolutionary war and the subversion of the country [...] those individuals, groups, or organizations that, while not communists, seek to rupture the established order.\textsuperscript{54}

The army and military intelligence applied this label to trade unions, associations, and religious organizations, thereby transforming non-combatant civilians into targets of violence.\textsuperscript{55} The Inter-American Commission on Human Rights found that the Guatemalan government progressively expanded the notion of the internal enemy to include intellectuals, artists, students, teachers, academics and union leaders—any who did not agree with the official line—and treated such groups as armed terrorists and insurgents.\textsuperscript{56} The same document says that the

“destruction of the

insurgency’s local political/administrative organization [...] is achieved through the capture and physical elimination of its active agents” and further notes that “the aim is not the pursuit of common criminals, but rather of individuals who are ideologically compromised but who are not participating in terrorist acts or combat operations.\textsuperscript{57}

This rhetoric drove the unrelenting slaughter of largely defenseless civilians and indigenous communities.

The most violent epoch in Guatemalan history was during the 1982-83 dictatorship of General Efraín Ríos Montt. Montt sanctioned a genocidal scorched earth campaign against the indigenous population in an effort to undercut support for the rebels. The military killed over 200,000 indigenous civilians and destroyed 440 indigenous villages.\textsuperscript{58} Additionally, Montt’s government increased efforts to force civilians to join paramilitary groups\textsuperscript{59} and commit violence against their own neighbors, whom were labeled Marxist rebels. Few of these crimes were ever prosecuted, creating a widespread climate of impunity that continues today.

Another military coup overthrew Montt in 1983.\textsuperscript{60} The 36-year armed conflict finally ended in 1996 with the signing of peace accords and the removal of some powerful

\textsuperscript{55} Id.
\textsuperscript{56} Id. ¶ 74.
\textsuperscript{57} Id. ¶ 73 (quoting the War Manual) (emphasis added).
\textsuperscript{58} Efrain Rios-Montt, TRIAL INTERNATIONAL (May 4, 2018), https://perma.cc/YZE8-4SKH.
\textsuperscript{59} Id.
\textsuperscript{60} Timeline: Guatemala, BBC NEWS, supra note 44.
military actors from public office.\textsuperscript{61}

\section*{B. Justice in the Aftermath of the Armed Conflict}

In 1999, the United Nations’ Commission for Historical Clarification published a report on the armed conflict, finding that 200,000 Guatemalans had been killed in the conflict and that the military and other government-backed security forces were responsible for 93\% of all documented human rights violations.\textsuperscript{62} According to the United Nations-sponsored truth commission, indigenous communities in the Ixil Triangle were subjected to genocide at the hands of agents of the state.\textsuperscript{63} In addition to those killed in the course of the armed conflict, over a million were driven from their homes or into exile.\textsuperscript{64} Villages were razed and starvation was widespread in the Mayan highlands.\textsuperscript{65} Fleeing civilians were relocated into military-controlled re-education camps and conscripted into the military under penalty of death.\textsuperscript{66}

In 2007, Montt became a member of Congress and retained immunity from criminal prosecution until his term ended in 2012. Twelve days after leaving office, Montt was indicted on charges of genocide and crimes against humanity\textsuperscript{67} as part of a national effort to combat impunity. He was convicted a year later on both counts and sentenced to 80 years in prison, but the verdict was overturned on procedural grounds by the country’s Constitutional Court. Retrial proceedings were initiated amidst national and international calls for justice, but Montt was declared mentally unfit to stand trial.\textsuperscript{68} Montt died in 2018 before the conclusion of his trial.\textsuperscript{69}

For the crimes committed during the internal conflict by Guatemalan government actors, international human rights monitors have commented that “virtually no one has been held accountable.”\textsuperscript{70} Beyond the lack of justice for officials like Montt, the culture of impunity for government actors continued long beyond the end of the armed conflict.

In 2006, the United Nations, in response to a formal request from the Guatemalan government, established an International Commission Against Impunity in Guatemala (CICIG).\textsuperscript{71} CICIG’s goal is to “investigate illegal security groups and clandestine security organizations in Guatemala,” and to “strengthen the rule of law in a post-conflict country.”\textsuperscript{72} In carrying out its mandate, CICIG has investigated paid assassins for criminal or political organizations, politicians, government officials, conspiracy schemes, and business representatives. CICIG’s work has led to the conviction of 310 government officials as well as the indictment of former President Otto Pérez Molina on corruption charges, which forced Molina to step down from

\begin{itemize}
  \item \textsuperscript{61} Id.
  \item \textsuperscript{62} COMM’N FOR HIST. CLARIFICATION, GUATEMALA
  MEMORY OF SILENCE: REPORT OF THE COMMISSION FOR
  HISTORICAL CLARIFICATION: CONCLUSIONS AND
  \item \textsuperscript{63} Id., ¶ 122.
  \item \textsuperscript{64} Larry Rohter, Guatemalans Formally End 36-Year Civil War, Central America’s Longest and Deadliest, N.Y. TIMES, Dec. 30, 1996, at A00008, https://perma.cc/9DMJ-HPFL.
  \item \textsuperscript{65} COMM’N FOR HIST. CLARIFICATION, supra note 62, ¶ 116.
  \item \textsuperscript{66} Id. ¶¶ 50, 69.
  \item \textsuperscript{67} TRIAL INTERNATIONAL, supra note 58.
  \item \textsuperscript{68} Id.
  \item \textsuperscript{69} Id.
  \item \textsuperscript{70} U.N. Special Rapporteur on the Rights of Indigenous Peoples, supra note 16, ¶ 98.
  \item \textsuperscript{71} CICIG, UNITED NATIONS: DEPT OF POL. AFFAIRS, https://perma.cc/6Z83-S4M9.
  \item \textsuperscript{72} Id.
\end{itemize}
The current President of Guatemala, Jimmy Morales, campaigned on an anti-corruption platform and initially supported CICIG’s efforts. Since assuming the presidency, however, Morales turned against the Commission in 2017 after it investigated him for campaign financing violations. Morales announced that the government would not renew CICIG’s mandate in August 2018, claiming that CICIG was “politically invading the country” and posing a threat to Guatemalan society and its notions of family values. In January 2019, President Morales, standing in front of a phalanx of military officials, withdrew from the treaty with the United Nations that established CICIG, a move that constituted a violation of international treaty law. Guatemala’s Constitutional Court overruled President Morales, who flouted the ruling by expelling CICIG personnel and impeding the reentry into Guatemala of the head commissioner of CICIG, Iván Velásquez, thus triggering a constitutional crisis.

As Morales has attempted to expel CICIG, one of his allies in the legislature, Fernando Linares Beltranena, has proposed amendments to the Law of National Reconciliation. The immediate effect of these proposals would be to extend amnesty to those accused and convicted of gross human rights violations including crimes against humanity, such as enforced disappearance, torture, summary executions, and sexual violence. The amendment, if adopted, would result in the release from prison of the few former military officials that have been convicted or are currently being processed “within 24 hours.” The UN High Commissioner for Human Rights responded to the proposal with deep concern for the “complete impunity for all those involved in some truly horrendous violations, including crimes against humanity.”

C. Today’s Political Landscape Online

Political and economic actors with ties to the military continue to exert influence in Guatemalan government and society. As one observer noted, “Ideologically conservative businessmen and high-ranking military officials from the middle classes . . . derived their influence by maintaining the army as an institution of social and political control.” The President of one of the organizations implicated in attacks on human rights defenders in Guatemala, the Fundación

---

77 Alfredo Ortega Franco, Guatemala on the Edge of the Abyss, ALJAZEERA (Jan. 10, 2019).
79 Sonia Pérez, Guatemala: advances plan to amnesty war crimes, ASSOCIATED PRESS (Mar. 6, 2019), https://perma.cc/ARY3-FC9V.
81 Id.
82 Id.
83 Interview 002.
84 Gutiérrez, supra note 17.
Contra el Terrorismo (Foundation Against Terrorism, or FCT), is Ricardo Méndez Ruiz Valdés. His father, Ricardo Méndez Ruiz Rohrmoser, was a military commander between 1980-82. The two military bases under Rohrmoser’s command, Cobán and Playa Grande, were deeply implicated in mass atrocities, crimes against humanity, and genocide. Méndez Ruiz Rohrmoser subsequently served as the Minister of Interior during the murderous dictatorship of General Ríos Montt.

In 2013, the Guatemalan Human Rights Ombudsman Jorge de León Duque condemned both FCT and Ruiz for their hate speech against human rights defenders, stating that Ruiz’s discourse on social media “feeds hatred and intolerance” and furthermore, “degrades, intimidates and promotes prejudice.”

FCT focuses its posts so often on the political left that “left” is the second most frequently used word on the FCT’s Twitter account. Tweets by Ruiz refer to Iván Velásquez, the director of CICIG, as a foreigner who comports with kidnappers and assassins, and indicate that he wants Velásquez’s head “as a trophy.” (Velásquez was the second-most used word on Ruiz’s Twitter account from 2017 through 2018.) Ruiz has also criticized foreigners in racially charged language, and he has used social media to deny the genocide against indigenous peoples that occurred in the 1980s, calling the genocide a “farce.” FCT and Ruiz have referred to activists in Guatemala as “communists.”

87 Id.
88 Id.
90 International Law and Human Rights Clinic, Content Analysis of Twitter User @RMéndezRuiz (2019) (unpublished).
91 For example, consider the following exchange: Ricardo Méndez Ruiz (@RMéndezRuiz), Twitter (Jun. 21, 2017), https://twitter.com/RMéndezRuiz/status/877688307450339330 (“If [US Ambassador Todd] Robinson called Idiots Sandra Morán, Amilcar Pop, Juracán or Taracena, maybe he would qualify as a foreigner son of a bitch.”). A user responded to Ruiz calling the ambassador a “black son of a bitch,” which Ruiz affirmed. Pedro Perez (@PedroPe18242815), Twitter (Jun. 21, 2017), https://twitter.com/PedroPe18242815/status/87773753426108417 (“Don Ricardo Well said, no matter who he is. The deputy is our representative and that black HDP has no right to insult.”); Ricardo Méndez Ruiz (@RMéndezRuiz), Twitter (Jun. 21, 2017, 5:42 PM), https://twitter.com/RMéndezRuiz/status/877742064720519168 (“That’s the point, exactly. May he call his officials in his country as he pleases; But here you have no right to insult them.”).
92 International Law and Human Rights Clinic, Content Analysis of Twitter User @RMéndezRuiz (2019) (unpublished).
93 Ricardo Méndez Ruiz (@RMéndezRuiz), Twitter, https://perma.cc/WGY4-48YE (“Declaration of Ricardo Méndez Ruiz Regarding the Abuses of Iván Velásquez and CICIG”); Ricardo Méndez Ruiz (@RMéndezRuiz), Twitter, https://perma.cc/AUQ9-R6CW (“The head of Iván Velásquez is going to be my trophy. If necessary, you have to kick it out.”).
and “terrorists,” terms that are widely used to delegitimize and intimidate human rights defenders in Guatemala. As of May 2019, Ruiz’s personal Twitter account has over 15,000 followers and 33,000 posts. The Foundation’s Twitter profile was suspended in late April 2019.

The attacks on indigenous and environmental groups are aligned with established economic interests in Guatemala, including the Committee of Agricultural, Commercial, Industrial, and Financial Associations (CACIF), which controls major exports in the country. Its economic dominance gives CACIF vast political power, and according to Guatemala’s former Minister of Foreign Affairs, “CACIF takes on a visible and belligerent role in times of conflict, most notably during administrations in which the institutionalized economic agents lose influence.”

Although our research did not identify specific mechanisms by which the speech of opinion-makers may contribute to offline violence, stigmatization, and harassment of defenders, such speech indirectly creates a climate of fear and intimidation that can silence opposition voices. Further, activists believe that denigrating speech emboldens those who use violence to reinforce their political and economic power in the country. This speech occurs in a context where the Guatemalan security forces have seized the equipment of rural radio stations and accused the staff of criminal offenses in an apparent attempt to silence indigenous voices and prevent the spread of information at the grassroots level.

Globally, the voices of opinion-makers are particularly concerning where they are further amplified by “net centers”—propaganda outlets known in other parts of the world as “troll farms” that create an illusion of broad consensus on an issue by flooding social media feeds with retweets from fake accounts. In Guatemala, observers have noted a pattern consistent with...
with the presence of net centers with apparently coordinated efforts across fake accounts to amplify state-aligned messages.\textsuperscript{106} These centers may be located in nondescript office buildings but inside, dozens of persons may be running hundreds of fake social media accounts and rapidly disseminating politically charged attacks against specific opposition individuals and institutions in Guatemala.\textsuperscript{107}

The activities of net centers are frequently aligned with the Guatemalan state. When President Morales declared Iván Velásquez \textit{persona non grata}, numerous accounts associated with such centers began to circulate content defaming Velásquez and CICIG.\textsuperscript{108} While some of the accounts were clearly fake, their message was rapidly disseminated by real Guatemalans on social media.\textsuperscript{109} Net centers distort public perception by creating fake accounts that follow real politicians; the popularity of those politicians is then artificially inflated, which in turn lends them authority in the eyes of other Guatemalans.

Coordinated slander has become a popular tool of what two investigative journalists call an “entrenched elite” in Guatemala.\textsuperscript{110} They find that the operation of net centers, dating back to as early as 2008, has become a highly profitable industry, with clients paying upwards of $280,000 for their services.\textsuperscript{111} Although their connections to the government and to business remains unclear, the journalists found that the activities of net centers often aligns with the interests of the Guatemalan political and economic elite.\textsuperscript{112}

Analysis of the posting habits of the accounts we reviewed reveals consistent efforts to discredit political opposition in Guatemala. One prominent voice on Guatemalan social media is Rodrigo Polo, and he makes extensive references to Iván Velásquez, CICIG itself, and “foreigners”.\textsuperscript{113} On his popular YouTube channel, Rodrigo Polo attacks journalists and human rights defenders as part of a Soros-funded conspiracy to bring a “leftist” government to power in Guatemala.\textsuperscript{114} He has attacked anticorruption activists and representatives of victims of the armed conflict, spreading baseless accusations of corruption and graft against them.\textsuperscript{115} Polo was temporarily banned from Twitter in 2018 for bot-like behavior, which implies use of automated software to retweet posts, like posts etc..\textsuperscript{116} He regained access to his account but has since been

\textsuperscript{106} See, e.g., Currier & Mackey, supra note 22; Harald Waxenecker, \textit{This is how the #JimmySeQueda was made: the netcenter of @rodrigopolo, @rMéndezruiz and @pirulismo, NOMADA} (Nov. 16, 2017), https://perma.cc/L4PS-Y9V5; Henry Bin, \textit{Asi operan los netcenters en Guatemala, en plena impunidad, CON CRITERIO} (Feb. 19, 2018), http://concriterio.gt/asi-operan-los-netcenters-en-guatemala-en-plena-impunidad/; A Corruption Spat in Russia Endangers Crime-Fighters in Central America, \textit{THE ECONOMIST}, Apr. 28, 2018, https://perma.cc/T5L4-R3DL.

\textsuperscript{107} Currier & Mackey, supra note 22.

\textsuperscript{108} Id.

\textsuperscript{109} Id.

\textsuperscript{110} Id.

\textsuperscript{111} Luis Assardo, \textit{Guatemalan troll factories}, A MEDIUM CORPORATION (Feb. 24, 2018), https://perma.cc/Y8PL-J4XE.

\textsuperscript{112} Id.

\textsuperscript{113} Rodrigo Polo, Debunking Luis von Ahn, (Sept. 30, 2018), https://medium.com/@RodrigoPolo/debunking-luis-von-ahn-404a97985fd4

\textsuperscript{114}https://www.youtube.com/watch?v=jnk4VZrYbak

\textsuperscript{115}https://medium.com/@RodrigoPolo/indignados-contratistas-73080d7d60a5

\textsuperscript{116} Rodrigo Polo, Suspended on Twitter, A MEDIUM CORPORATION (Feb. 25, 2018) https://perma.cc/WV4S-38JG. For a discussion of bot-like behavior, see University of Cambridge, \textit{Celebrity Twitter accounts display 'bot-like'..."
banned again from the platform. Posts by Ricardo Méndez Ruiz and his audience are also extremely negative. Social media analytics using Talkwalker indicate that the Ruiz’s tweets have an overall net sentiment of -66% for the month of February 2019. Judges, terrorism, anti-terrorist, and the left are associated with some of the most negative sentiment for Ruiz. Ruiz posts frequently, and we counted more than 400 posts in the month of February. Despite a modest engagement rate of 8%, the content itself was highly viral, developing nearly 43,000 additional posts. If the engagement rate is modest but the subsequent reposting rate is high, this suggests that bots may be targeting a post for tweeting and retweeting to amplify Ruiz’s messages.

The online harassment of Thelma Aldana, the former Attorney General of Guatemala, illustrates the pervasiveness of the speech campaigns against defenders. Aldana worked closely with CICIG head Iván Velásquez from 2014 through 2018. Together, they brought 60 high-profile corruption cases, resulting in the resignation of then-president Otto Pérez Molina. President Jimmy Morales replaced Aldana with a new attorney general in May 2018. Aldana announced a Guatemalan presidential run and has become one of the most targeted individuals in Guatemala.

The specter of Guatemala’s “internal enemy” has not disappeared. State-aligned actors continue to characterize dissent as an affront to the nation, calling defenders guerillas, communists, destabilizers, foreign invaders, terrorists and criminals. The harassment is so pervasive that at least one defender has given up on flagging hostile posts. State-aligned actors continue to deny the genocide and the violence that persists today, especially in indigenous communities.

The campaign of dehumanization of defenders appears to be continuing the government’s long campaign against the internal enemy.

---

117 See Account Suspended (@RodrigoPolo), TWITTER, https://twitter.com/@RodrigoPolo.
118 Net sentiment uses artificial intelligence to analyze social and digital media conversation. Most conversations are neutral in nature, but others are clearly negative or positive. Net sentiment eliminates the neutral conversation and calculates a sum on the basis of the formula (positive - negative).
120 Engagements indicate the number of times a user interacted with a tweet (including likes, replies, retweets etc.), and the engagement rate is the number of user engagements divided by the number of views.
123 Jami López, Thelma Aldana, ex-prosecutor of Guatemala, confirms that she would like to become president of her country, EL SALVADOR.COM (Jan 15, 2019), https://perma.cc/NKB7-X4K6.
124 Interview 014.
INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA

Facebook and Twitter have both come under immense pressure over the past few years with respect to their content moderation policies. Both companies have elaborate content moderation policies that address a variety of different forms of harmful speech—from copyright violations to threats of violence. Both employ moderation teams that sift through millions of posts a day, making instant judgements about how to balance free expression and the harms of speech. The purpose of this section is to consider the practices and policies of these companies and how they apply to the government-aligned hate speech directed at defenders in Guatemala today.

A. Content Moderation Processes on Twitter and Facebook

Facebook and Twitter both employ human moderators to work alongside artificial intelligence and assess content that has been reported by users. Although Facebook has publicly committed to hired over 7,000 moderators, it is unknown how many are employed by Twitter. Nor is it apparent how moderators function within their content moderation structure because Twitter has not publicly disclosed its content moderation practices.

Harmful content is first identified at both companies through user flagging. In certain instances, such as child pornography and, in recent years, extremist content and hate speech, the companies may use algorithms to identify harmful material on their platforms. Otherwise, both companies rely on users to identify harmful content that is then sent to a human reviewer to be assessed for its consistency with their content moderation policies. Each has created a flagging mechanism for users to tag content that they believe violates the platform’s terms of service.

The processes for identifying problematic content vary between Facebook and Twitter. Facebook allows users to report abusive content through a “Give feedback or report” link. Twitter outlines how someone should respond to offensive posts on an “About online abuse” page which suggests unfollowing, blocking and, ultimately, reporting to Twitter. Twitter also urges someone who feels threatened to contact local law enforcement. The process for reporting on Twitter is outlined on the platform.

At Facebook, a three-tier structure guides human content moderation. The bottom tier (Tier 3) consists of thousands of moderators situated in various geographical locations around the world clicking through hundreds of thousands of posts per day. They are the first reviewers of reported content, judging its admissibility using Facebook’s content moderation policies, and looking at everything from attacks on race or sexual orientation to violent or pornographic content.

---

128 Koebler & Cox, supra note 15.
129 Interview 009.
130 Id.; Zuckerberg, supra note 11.
135 Andrew Arsht & Daniel Etcovitch, Commentary—The Human Cost of Online Content Moderation, JOLT DIGEST (March 2 2018), https://perma.cc/3SSR-H4MT.
136 Klonick, supra note 132, at 1640.
If a Tier 3 moderator is unsure of whether a post is appropriate, they will send it to the mid-level moderator (Tier 2), who is usually the supervisor of a content moderation center who has more experience in the field.\(^\text{137}\) Finally, if the Tier 2 moderator still cannot determine whether the post is appropriate, the matter is escalated to the top of the chain to Tier 1 moderators, who are often lawyers or policy makers, at the main Facebook HQ offices in Palo Alto, California.\(^\text{138}\)

**B. Content Moderation Policies on Facebook**

Facebook’s community guidelines prohibit among other things credible threats of violence, hate speech, harassment, and the coordination of harm or injury. The majority of coded speech campaigns against defenders in Guatemala do not clearly violate these policies. The coded nature of the speech means that the threats embedded within the language and inherent in the coordinated attack are not apparent when any individual post is viewed in isolation. Facebook’s policy on hate speech does not cover some of the more offensive posts because it only prohibits attacks based on protected characteristics. Merely targeting someone for their human rights work on behalf of protected communities is insufficient.\(^\text{139}\) It is possible that the campaign itself might constitute harassment, but the coordinated activity in question seems to fall short of the threshold that Facebook describes for coordination.

**1. Credible Violence**

Facebook’s community standards prohibit credible threats of violence. Among other things, this policy prohibits 1) “[c]redible statements of intent to commit violence against any person, groups of people, or place”; 2) “[a]ny statement of intent to commit violence,” “[c]alls for violence or statements advocating violence,” and “[a]spirational and conditional statements of violence” against vulnerable persons or groups such as activists and journalists, public individuals, or groups of people; and 3) “[m]isinformation that contributes to imminent violence or physical harm.”\(^\text{140}\) Facebook assesses the credibility of statements based on the amount of detail included, such as whether the statement is accompanied by a target and a bounty, describes a weapon or building, or identifies location, timing, or method of the attack.\(^\text{141}\) Facebook also states that it works with local law enforcement if it appears that a person is in imminent danger.

Although the coordinated campaign of coded speech is perceived as threatening by activists and advocates in Guatemala, the posts—when viewed in isolation—do not trigger Facebook’s credible violence policy.

---

\(^{137}\) *Id.*  
\(^{138}\) *Id.*  
\(^{139}\) FACEBOOK COMMUNITY STANDARDS, OBJECTIONABLE CONTENT, https://perma.cc/WZ99-V2TM.  
\(^{140}\) FACEBOOK COMMUNITY STANDARDS, CREDIBLE VIOLENCE, https://perma.cc/FSNA-ADWD.  
\(^{141}\) *Id.*
In this post,142 the FTC is referring to the PNC, or the National Civil Police (Policía Nacional Civil), which was created as part of the 1996 Peace Accords with the goal of reforming the police and creating a clear separation between the police and military.143 Because of the role of the military in the repression and the control it has historically maintained over the police and the overall security situation in Guatemala, the creation of a national police force subject to civilian oversight has been controversial. A complete separation of police and military was rejected in a 1999 national referendum.144 The post is also addressing the National Reparations Program (Programa Nacional de Resarcimiento), which was created in 2003 to provide compensation to the largely indigenous victims of the armed conflict.145 

The post targets Helen Mack and Rosalina Tuyuc, two of the most prominent defenders in Guatemala. The post accuses the PNC and the PNR of corruption, alleging with respect to the latter that it simply funnels money to the “guerrilla terrorists” while injured veterans are denied medical care.

Viewed in isolation, this post is clearly not a statement of intent to commit violence, a call for violence, advocating violence, or even aspirational support for violence. Nonetheless, it can be seen as contributing to a climate in which violence against indigenous Guatemalans can occur. The primary recipients of the reparations program are indigenous communities that were the victims of the genocide. This post calls these recipients “terrorists”—derived from the government’s use of the word “terror” to disenfranchise indigenous and other vulnerable groups during the conflict, which has since been adopted by FCT and applied (particularly in the post 9/11 context) to associate current activists with the terrorists from before.”146 The post also evokes sentiments of nationalism, alleging that (non-indigenous) army veterans are denied care while the indigenous “guerrilla terrorists” enjoy a “piñata of resources.”

promotes racism and ethnic conflict, was denounced by the same supposed victims who she provides restitution to, to waste her resources.”

142 Translation: “Helen Mack: Commissioner of Political Reform. If any institution of the State has been permanently characterized by corruption, it’s the PNC, from overpricing/overcharging to sequestering/stealing and extortion by some agents. Has she ever denounced this? All citizens must denounce offences and she is in this position of almost absolute power. There must be criminal responsibility and she must respond. Rosalina Tuyuc: Another ex-director of the National Reparations Program which has been a piñata of resources for guerrilla terrorists and their families— in contrast, the Army veterans crippled by mines placed by Subversives and in ambushes, they’ve been given no medical attention. This activist who


144 Id. at 440.


146 Fundación Contra el Terrorismo, Denuncia de la Fundación Contra el Terrorismo, supra note 26.
The posts do not reference violence, but they do reinforce negative perceptions of indigenous communities by evoking language that was commonly used during the armed conflict to justify human rights abuses against these individuals.

The post could be construed as “[m]isinformation that contributes to imminent violence or physical harm.” Researchers are increasingly pointing to the way in which misinformation and disinformation (“fake news”) online can contribute to violence, and this post provides such an example. Nonetheless, it is unclear how Facebook interprets this provision, and it is also unlikely that the post contributes to “imminent violence or physical harm.” It is not this individual post, but rather a campaign of similar posts, that is undermining the rule of law in Guatemala.

2. Harassment

Facebook prohibits harassment, which includes “[t]hreatening any participant in public discourse with violence in an attempt to intimidate or silence them” as well as “[p]ost[ing] content about anyone, including a public figure, that contains a name, photo, or video of the individual” and any statements of intent to commit, calls for action, or statements advocating lethal violence. The posts we reviewed for this report are attempts to silence defenders, and they often contain names and photos, but they do not explicitly reference violence.

For example:

This post is discussing the Molina Theissen case, which was a prosecution brought against former military generals for the illegal detention and rape of Emma Molina Theissen and the disappearance of her 14-year-old brother, Marco Antonio. The PGT is one of the now-defunct communist parties in Guatemala, the Guatemalan Workers Party (Partido Guatemalteco del Trabajo) that existed between 1952 until it merged with other leftist organizations after the conclusion of the Peace Accords. The picture displayed is of Iduvina Hernández, with a smaller picture of Marco Antonio. In this post, Ricardo

148 FACEBOOK COMMUNITY STANDARDS, CREDIBLE VIOLENCE, supra note 140.
149 FACEBOOK COMMUNITY STANDARDS, HARASSMENT, https://perma.cc/RBW5-C6DR.
150 Ricardo Méndez Ruiz, FACEBOOK, https://www.facebook.com/pg/MendezRuizV (“When I affirm that Edgar Gutierrez and Iduvina Hernandez are Marxist terrorists responsible for my kidnapping, I don't lie. Read here the documentary proof of what I claim. These are the kind of people who support Colombian Velasquez and the corrupt part of the private sector. They seek a conviction against our veterans for the Molina Theissen Case, in which the plaintiff is a terrorist of the [Guatemalan Party of Labour], the same faction they belonged to.”).
152 MODERN GENOCIDE: THE DEFINITIVE RESOURCE AND DOCUMENT COLLECTION 929.
Méndez Ruiz argues that former-government minister Edgar Gutiérrez and defender Iduvina Hernández are “Marxist terrorists” and criminals, and that criminals are supporting Iván Velásquez (and thus the work of CICIG). It is alleging that the complainants in the Molina Theissen case, as well as Gutiérrez and Hernández, are also “terrorists” associated with a now-defunct communist party. In short, this is a defamatory post that is part of a larger smear campaign discrediting human rights defenders and their work. It does not, however, explicitly reference or call for violence.

3. **Hate Speech**

Facebook prohibits hate speech which it defines as “a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.” Facebook defines an “attack” as “violent or dehumanizing speech, statements of inferiority, or calls for exclusion or segregation,” and it separates such attacks into three tiers of severity:

- **Tier 1 attacks** are those that “target a person or group of people who share one of the above-listed characteristics or immigration status,” where the attack is violent or dehumanizing speech. Dehumanizing speech involves referring to or comparing the target to, among other things, insects, animals, filth, disease, or criminals.

- **Tier 2 attacks** target persons or groups sharing the above-listed characteristics, when the attack includes “[s]tatements of inferiority or an image implying a person's or a group's physical, mental, or moral deficiency” and expressions of contempt or disgust.

- **Tier 3 attacks** are “calls to exclude or segregate a person or group of people based on the above-listed characteristics” and “[c]ontent that describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics.”

Many of the posts attacking human rights defenders do not unequivocally constitute hate speech because it is not clear they are directed against an individual based on a protected characteristic. For example, the prior posts targeting Rosalina Tuyuc and the work of the PNR is implicitly targeting indigenous communities and activists. The posts do not seem to be targeting them based on their indigeneity, however, but rather their political activities—such as seeking reparations for victims of human rights violations. Nonetheless, the clear association of the PNR with the indigenous victims of the genocide, the targeting of Tuyuc herself, and the language of criminality, evokes the deeply embedded racism against indigenous peoples that fueled the genocide itself. In another other case, a post attacking a human rights defender was accompanied by an indirect allusion to, but not direct condemnation of, her sexual orientation.

---

154 Id.
155 Fundación Contra el Terrorismo Guatemala, [FACEBOOK](https://www.facebook.com/FundacionContraElTerrorismoGuatemala/) (Oct. 30 2015), “Sandra Morán will be the first lesbian n the Congress of Guatemala.”
4. Coordinating Harm

Finally, Facebook’s community guidelines also “prohibit people from facilitating or coordinating future activity, criminal or otherwise, that is intended or likely to cause harm to people, businesses, or animals.” This would seem to provide a basis to address coordinated slur campaigns against human rights defenders. However, the examples provided by Facebook indicate that the policy is directed to posts that themselves coordinate harm, rather than posts that are part of a coordinated campaign. For example, the policy prohibits “[s]tatements of intent, calls to action, or advocating for” among other things “[a]cts of physical harm committed against people.” This indicates that a statement must be part of the act of coordination—for example, by calling for violence or harm—to trigger this policy; merely being one post of many that are part of a coordinated attack would not seem sufficient.

C. Content Moderation Policies on Twitter

Twitter’s rules regarding permissible and impermissible content broadly restrict against “specific threats of violence” as well as “abuse and hateful conduct.” The coded speech campaigns in Guatemala against defenders do not clearly violate either of these policies. The coded nature of the speech means that it does not constitute a specific threat of violence.

Although the speech in question likely would constitute hateful conduct, this kind of speech is generally only prohibited under Twitter’s rules if it is directed at a vulnerable group, which does not include human rights defenders. Some of the tweets may however constitute abuse under Twitter’s policies.

1. Threats of Violence

Twitter’s rules provide that users “may not make specific threats of violence or wish for the serious physical harm, death, or disease of an individual or group of people.” Twitter prohibits both violent threats as well as posts that “glorify[y] acts of violence in a manner that may inspire others to replicate those violent acts and cause real offline danger, or where people were targeted because of their potential membership in a protected category.”

Twitter defines violent threats to be “explicit statements of one’s intent to kill or inflict serious physical harm against another person. This includes, but is not limited to, threatening to murder someone, sexually assault someone, break someone’s bones, and/or commit any other violent act that may result in someone’s death or serious injury.” Twitter specifies that “wishing or hoping that someone experiences serious physical harm, making vague threats, or threatening less serious forms of physical harm would not fall under this specific policy.” Twitter works with local partners to ensure higher-level review of threats against journalists and activists.

None of the tweets identified as online harassment directed at defenders in Guatemala would trigger Twitter’s policy prohibiting specific threats of violence. The closest example is a post from Ricardo Méndez Ruiz referencing an attack against a hydroelectric plant in Ixquisis, the site of violent and non-violent protests against the building of a dam in an isolated Mayan

157 Id.
158 Interview 017.
INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA

An indigenous region of the country:

This tweet could constitute a threat in the Guatemalan context, particularly due to its call to treat the attackers “[w]ithout any mercy.” Although the incident Ruiz was responding to was unlawful, it occurred in the context of ongoing indigenous protests of hydroelectric projects in Ixquisis, in which indigenous protesters are routinely labeled “terrorists.” The tweet says these “terrorists” must be “dealt with” by the security forces, evoking directly Guatemala’s history of the military violently targeting peasant and indigenous communities during the armed conflict. Nonetheless, the tweet does not likely constitute a direct threat of violence under Twitter’s rules, since it is not an “explicit statement[] of one’s intent to kill or inflict serious physical harm” on another. At most, this tweet is an indirect or “vague” threat of the kind that Twitter does not explicitly prohibit under its current policy.

Further, although this post condones violence, it would also not constitute glorification of violence under Twitter’s rules. Twitter only prohibits content as glorification of violence if it either “condones or celebrates violence (and/or its perpetrators) in a manner that may promote imitation of the act” or where “protected categories have been the primary target or victim.” Examples of prohibited content include glorification of “mass murders, terrorist attacks, or rapes and sexual assault.” This tweet, in contrast, is not targeting an individual based on their membership in a protected group, and it is not celebrating past violent acts.

2. Abuse

Twitter’s rules provide that users “may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behavior an attempt to harass, intimidate, or silence someone else’s voice.” Although the rule is described initially in Twitter’s documentation as applying broadly to “prohibit behavior that harasses or intimidates, or is otherwise intended to shame or degrade others,” it appears from Twitter’s explanation of this rule that it is limited to specific categories of abusive behavior, three of which are relevant to our analysis:

- Wishing or hoping serious harm on a person or group of people
- Using aggressive insults with the purpose of harassing or intimidating others
- Encouraging or calling for others to harass an individual or group

---

159 Ricardo Méndez Ruiz (@RMéndezRuiz), TWITTER (Oct. 31, 2018), https://twitter.com/RMéndezRuiz/status/1057677872700289024 (“They don’t just handle guns; They also wield explosive skills. Does anyone still doubt that this is about terrorism? As such, the security forces should treat them. Without any mercy.”).

160 Incendian tuberías de presión de hidroeléctrica en Ixquisis; Los atacantes provocaron una detonación en el área de balsa de regulación de la empresa, EL PERIODICO, Oct. 20, 2018, https://perma.cc/S7XC-GPYW.


162 TWITTER, RULES AND POLICIES: ABUSIVE BEHAVIOR, https://perma.cc/5SN6-G99H.
of people.

Based on the examples provided in Twitter’s content moderation rules, none of the tweets we reviewed would likely constitute “[w]ishing or hoping serious harm on a person or group of people.” Twitter provides as examples of this content statements such as: “I hope you get cancer and die”; “I wish that you would get run over by a car next time you run your mouth”; and “If this group of protesters don’t shut up, they deserve to be shot.” Although the use of terms like “communist” or “terrorist” or the post discussed above calling for action by the security forces would be understood by those familiar with Guatemala’s history as a wish or hope for serious harm to those resisting the government, they are not the kind of direct “wish or harm” illustrated by the examples provided by Twitter.

None of the tweets we reviewed would constitute “[e]ncouraging or calling for others to harass an individual or group of people.” This is defined in Twitter’s rules as content that “encourages others to harass or target specific individuals or groups with abusive behavior” including but not limited to “calls to target people with abuse or harassment online and behavior that urges offline action such as physical harassment.” Although many of the tweets are implicit calls to action, none are likely an explicit enough call to action to trigger Twitter’s policies.

For example, the following tweet contains both coded language and an implicit call to action:

In this tweet, Ruiz is referring to the human rights defenders and political activists that work for the organization CODECA, Comité de Desarrollo Campesino (Committee for the Development of the Peasants), a grassroots organization for protecting the rights of the rural (and mostly indigenous) poor in Guatemala. Ruiz calls democratic political organizers “terrorists” and says that they will “have to face a united and radical right” in the coming election. The reference to terrorists in this context is particularly concerning, since the activists of CODECA work in the same communities in Guatemala that suffered genocide during the armed conflict. The call for a “united” right provides an impulse for action, and it characterizes the necessary action as “radical.” Although to those versed in Guatemala’s history, this tweet would immediately raise concerns, it would not appear to trigger the harassment policy. Any call to action is veiled, at best, and the reference to the election makes this appear to be pure political speech, and thus the kind of speech that is generally entitled to the greatest level of deference.

At least some of the tweets we studied might constitute use of “aggressive insults with the purpose of harassing or intimidating others.” Twitter explains: “We

163 Ricardo Méndez Ruiz (@RMéndezRuiz), TWITTER (Oct. 31 2018) https://twitter.com/RMendezRuiz/status/1057609304608514048 (“It is a fact: the terrorists of

CODECA have a political party, the Movement for the Liberation of the Peoples (MLP). It will be up to them to face a united and radical right in the upcoming elections.”).
take action against excessively aggressive insults that target an individual, including content that contains slurs or similar language. Please also note that while some individuals may find certain terms to be offensive, we will not action against every instance where insulting terms are used.” One tweet we analyzed likely meets the standard of aggressive insults.

The expression in this tweet is certainly insulting and the volume of insults might be sufficient to constitute abusive use of such insults. Further, these insults explicitly reference prominent human rights defenders, lawyers, foreign diplomats, and activists in Guatemala—Todd Robinson, the former U.S. Ambassador to Guatemala; Iván Velásquez, the head of CICIG; Thelma Aldana, former Attorney General and former President of the Supreme Court of Guatemala; and Helen Mack, one of the most well-known human rights advocates in the country.

Either a target or a bystander can report abuse. However, Twitter emphasizes that context matters in understanding the impact of speech, and thus it is more likely to remove speech that is reported by the target of the speech. Thus, Twitter notes it “takes action against excessively aggressive insults that target an individual,” but that it may need to hear directly from the target in order to determine whether the speech is abusive.

3. Hateful Conduct

Twitter’s rules prohibit “hateful conduct,” noting that users may not “promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.” Twitter explains further that this policy is based on its commitment “to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category.”

165 Bávara Hernández (@HdezBarbara), Twitter (May 8, 2017) https://twitter.com/HdezBarbara/status/86169744330553441 (“We are parasites / We are (Vagos) / We are servile / We are lazy / We are useless / We are trans / We are corrupt / We are pot-heads / We are (Escorias) / We are gays, lesbians, bisexuals / We are anti-social / We are like Todd / We are like Velasquez / We are like Aldana / We are like Mack / We are the same shit”).
166 Twitter, Rules and Policies: Hateful Conduct Policy, https://perma.cc/88QF-HCRQ. In its hateful conduct explanation, Twitter provides a somewhat different definition: “You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.”
167 Id.
There is a range of speech prohibited under Twitter’s hateful conduct policy, including “Violent threats against an identifiable target”; “Wishing, hoping or calling for serious harm on a person or group of people”; “References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims”; “Inciting fear about a protected category”; “Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone”; and “Hateful imagery.”

Although some of the tweets reviewed for this report constitute attempts to incite fear, they do not trigger Twitter’s moderation policies because they are not obviously directed at individuals belonging to a protected category. Incitement of fear only encompasses content that is designed to “incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, ‘all [religious group] are terrorists’.” Tweets we reviewed that allege that defenders are criminals or otherwise engaged in illegal activity are not directed at individuals on the basis of their membership in a protected category.

For example:

In this tweet, a picture of Helen Mack, a highly recognizable human rights activist in Guatemala, is positioned under a bright red word “Corrupt.” Give current battles against corruption in Guatemala as well as the work of CICIG and the high-profile prosecutions of corrupt politicians and businesspeople in the country, the word “corrupt” is highly charged and clearly implies a claim that Helen Mack is a criminal. This tweet would not, however, violate Twitter’s moderation policies because Helen Mack is not being targeted on the basis of any membership in a protected group, defined by “race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.” Twitter looks for the use of racial slurs, tropes, or references to particular events in order to determine whether the speech is targeting someone based on their membership in a particular group, and this tweet does not contain any explicit racial slur or trope. The reference to “left” might be evidence that she is being targeted for her work as a human rights defender, but Twitter does not prohibit hateful conduct that targets journalists or activists.

Similarly, many of the tweets contain “[r]epeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.” The purpose of this rule is to prohibit “content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes.

---

168 #GuateSoberana (@PiruLismo), TWITTER (Apr. 19, 2017) https://twitter.com/PiruLismo/status/85477637577079296 (posting a picture of Helen Mack with superimposed text showing “left” “corrupted.”).

169 Interview 016.

170 Id.

171 Hateful Conduct Policy, TWITTER (last accessed May 17, 2019) at
about a protected category.”\textsuperscript{172} Many of the tweets in question are focused on dehumanizing, degrading, and reinforcing harmful and negative stereotypes about defenders, such as the tweet containing the list of slurs analyzed in the preceding sub-section. Nonetheless, this tweet would not trigger Twitter’s hateful conduct policy because they are not slurs, epithets, tropes or degrading content directed at individuals belonging to a protected category.

4. \textit{Spam and Fake Accounts}

Twitter’s rules prohibit, among other things, “scripting the creation of content in such a manner as to interfere with or create an undue burden on Twitter.”\textsuperscript{173} Users cannot use fake accounts to “manipulate the conversations on Twitter.”\textsuperscript{174} They are also prohibited from engaging in “bulk or aggressive activity that attempts to manipulate or disrupt Twitter or the experience of users on Twitter to drive traffic or attention to unrelated accounts, products, services, or initiatives.”\textsuperscript{175} Coordinated activity including through net centers in Guatemala, such as automated coordinated boosting of tweets, may trigger these policies.\textsuperscript{176}

D. \textit{Process and Sanctions}

Both Facebook and Twitter lay out on their websites the consequences of breaching their content moderation policies. For Facebook, the repercussions of violating their community standards depends on the seriousness of the violation as well as the person’s history on the site. Twitter can remove individual posts as well as a person’s profile and may notify law enforcement personnel if it believes someone poses a “genuine risk of physical harm or a direct threat to public safety.”\textsuperscript{177} Similarly, a violation of Twitter’s User Agreement can result in a ban on new content or even a permanent ban.\textsuperscript{178} Twitter will also monitor and remove new accounts created by banned users.

Facebook provides users with categories for identifying why they view the post as problematic. Categories include violence, harassment, false news, hate speech or “something else” which generates a further list of categories. A statement also appears which tells users to contact local law enforcement or emergency services if they feel that someone is in immediate danger.

Twitter’s process of reporting potential harmful posts includes several additional steps. It initially allows users to identify abusive or harmful tweets and then gives them the opportunity to elaborate on why they believe they fall into this category but giving them options such as it “includes targeted harassment” or its “threatening violence of physical harm” (see below). In the third step, users are then asked who

\begin{itemize}
  \item \textsuperscript{172} Id.
  \item \textsuperscript{173} Id.
  \item \textsuperscript{174} Id.
  \item \textsuperscript{175} Id.
  \item \textsuperscript{176} Interview 016.
  \item \textsuperscript{177} Id.
  \item \textsuperscript{178} \textsc{twitter, rules and policies: hateful conduct policy}, supra note 166.
\end{itemize}


the threats, harassment or hate is directed towards – the user themselves, someone else, or a protected group.

E. Why are the Current Moderation Practices Struggling to Protect Defenders in Guatemala?

Although content moderation practices are constantly evolving to adapt to new technology and changes in society, there are identifiable gaps we have found with respect to online attacks of defenders in Guatemala. First, because of lack of transparency around user flagging for terms of service violations combined with the pervasive nature of this speech, it is unclear how much of the speech in question is being flagged for human review. Second, because of the context-specific nature of the coded language and the fact that it is not typically directed at a vulnerable group, posts targeting defenders would in most instances not trigger Facebook or Twitter content rules even if they were more systematically flagged. Third, the existence of “net centers” that coordinate attacks using hundreds of accounts result in broad dissemination of harmful posts which are harder to remove. Speech that is removed is also often re-posted under new accounts. Fourth and finally, content moderation policies of Facebook and Twitter do not provide a means for considering the status of the speaker in evaluating the harmful impact the speech may have, and it is widely recognized in the social science literature that speakers with authority have greater influence on behavior.179

1. Reliance on User Flagging

Twitter and Facebook both rely on user flagging to identify violations of their terms of service. Flagging alone, however, is not a reliable mechanism for identifying harmful speech because it relies on users, who may not be equally able or motivated to flag harmful content. For Twitter, certain objectionable content is only considered for removal if the target of the post flags the concern. Twitter says that context is important when evaluating for harassment and needs to hear from the person being targeted to review this type of content.180

Experts we interviewed about the content moderation policies of Facebook and Twitter expressed the view that companies are over-reliant on user flagging to trigger a review.181 This can result in many dangerous posts remaining online and going unnoticed by site moderators. Further, lower technological literacy may hinder defenders’ capacity to flag.182

Although both Facebook and Twitter have procedures in place for reporting, doing so requires having sufficient awareness and technological knowledge to navigate the

180 TWITTER, RULES AND POLICIES: ABUSIVE BEHAVIOR, https://perma.cc/VC3T-WXPK.
181 Interview 009.
182 Interview 007.
process. In many cases, especially for high profile defenders, they may be unable or unwilling to report every incidence of threatening or abusive language on social media. In our interviews, at least one defender reported that they only contact social media platforms when their account has been hacked. Indeed, given the volume of tweets identified, individual user flagging of all instances would likely be impossible. The volume of abusive posts has led some targets to only react to the most egregious threats or just tune out and ignore the threats. Further, targets may not be even aware they are being targeted online. “Flagging fatigue” by defender’s who are frequently attacked coupled with a process that depends on users reporting every instance of harmful behavior represents a troubling aspect of content moderation practices.

Increasingly, SMCs are experimenting with flagging through artificial intelligence. Facebook, for example, has complemented user flagging with AI. Yet deploying artificial intelligence to flag content comes with the risk of over-censorship. Algorithmic identification is imprecise and unlikely to allow for contextual cues that might be necessary to distinguish extremist speech and parody, documentary footage, and protest.

### 2. Context

Although the content moderation policies of Facebook and Twitter both recognize the importance of context in evaluating harmful speech on their platforms, neither company’s content moderation policies or practices provide a basis for considering the implicit meaning of coded threats against defenders in specific country contexts. Understanding the context of a post is one of the most vital components of knowing whether it is dangerous. SMCs have acknowledged how difficult it is to contextualize posts and are constantly reviewing their practices and guidelines to ensure that their systems can moderate posts accurately, but they have not yet found ways to more systematically integrate geographically localized meanings into their policies and practices. Content moderation practices operate more effectively in American and European contexts in part because those who design content policy are better able to understand languages and the cultural nuances of speech in those areas. Even though SMCs have content moderators for every single language in the world, this does not mean that content reviewers will be able to pick up on geographically-specific meanings. Even in areas of the world where a common language is shared – like Latin America, for example – many countries’ dialects employ words and

---

183 Interview 007.
184 Interview 015.
185 Id.
186 Id.
187 Interview 008.
188 Koebler and Cox, supra note 15.
190 Koebler & Cox, supra note 15.
192 Interview 009.
slang which have altered meanings, that carry a different weight, or contain unique or historically-specific terms which will be misunderstood or go unnoticed to someone who is unfamiliar with the area. Thus, it may not be the case that Spanish speakers will fully understand the context of posts in all Spanish-speaking countries unless they are provided with or have acquired a working knowledge of the current context, terminologies, slang, and cultural nuances. For example, posts attacking Guatemalan defenders with words such as “communist,” “terrorist” or “destabilizer” are more dangerous than when these words are used in Colombia or Ecuador. As such, even a Spanish-speaking moderator with no knowledge or understanding of Guatemalan current affairs would likely be unable to distinguish between the varying resonance of such words.

3. Net Centers

Even if an individual post is taken down, duplicate posts from multiple other user accounts (including fake profiles) carry the same harmful online attacks. “Net centers” in Guatemala are able to manage hundreds of online accounts to disseminate messages about individuals and groups. Employees involved in such “centers” have described how they were told to post messages of support for and attacks on various people involved in public life in a coordinated manner, and especially in election periods. Even when the original post is removed, sharing and retweeting content by other accounts leads to harmful information spreading.

4. Status of the Speaker

The main perpetrators of online hate speech towards defenders in Guatemala often have large followings. At least some of these actors are affiliated with the institutional political and military power structure and have thousands of followers on Twitter. Given their authority, their messages carry more weight with their audiences.

Twitter and Facebook do not appear to have procedures or policies in place to take account of the influence that a speaker might have due to his or her position. Under Twitter’s content moderation rules, for example, they consider a number of contextual factors, including whether

- the behavior is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest.

---

193 Koebler & Cox, supra note 15.
194 Interview 009.
195 Currier & Mackey, supra note 22.
196 Id.
197 Ferdy Montepueque, Mis días en un ‘net center’ de Baldizón y Jimmy Morales, DIARIO DIGITAL, May 5, 2016, https://perma.cc/Y83C-NWKQ.
198 Currier & Mackey, supra note 22.
199 See, e.g., Twitter Profile of Ricardo Méndez Ruiz (https://twitter.com/RMendezRuiz) (over 16,000 followers); Twitter Profile of Fratti Giovanni (https://twitter.com/frattigiovanni) (over 15,000 followers).
201 TWITTER, OUR APPROACH TO POLICY DEVELOPMENT AND ENFORCEMENT POLICY, https://perma.cc/AEP3-3X7P.
It does not appear from their written policies that Twitter or Facebook consider the status of the speaker in evaluating the harmful impact that a post may have. However, Twitter does take the newsworthiness of tweets into consideration, noting that “there may be the rare occasion when we allow controversial content or behavior which may otherwise violate our Rules to remain on our service because we believe there is a legitimate public interest in its availability.”

Thus, to the extent that a speaker’s status is relevant, it often weighs against removal. Facebook, for example, notes that conduct that would otherwise constitute bullying would be permitted if the target is a public figure.

Companies such as Twitter have occasionally “de-platformed” owners of social media accounts when these accounts are suspected of violating the platform’s rules. However, again, this occurs more often in America and Europe where SMCs have come under more pressure to act, and are more aware of the dangers that individual accounts can pose. Similar scrutiny and monitoring must be adopted to respond to powerful social media accounts in Guatemala and other countries in Africa, Asia, and Latin America.

**SOCIAL MEDIA COMPANY RESPONSIBILITIES**

Although social media platforms play an important role in creating space for expression and sharing, they can also host speech that contributes to human rights violations. Given the institutional violence and the lingering and systemic impunity in Guatemala, online speech in that country needs to be evaluated with particular care.

Content moderation policies and practices fall short of addressing the harms of the posts analyzed in this Report in two specific ways. First, they are not designed to be sensitive to the geographically specific and coded meanings of speech. Second, they do not provide protection to human rights defenders that are the targets of state persecution. This section draws on principles of international law to argue that SMCs should reform their content moderation policies and practices to address these two gaps.

SMCs should seek to align their content moderation policies and practices with international human rights and international criminal law as much as possible. Under the UN Guiding Principles on Business and Human Rights (the UN Guiding Principles), companies have a responsibility to avoid, mitigate, and remediate human rights harms that are “directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts.”

More broadly, the UN Special Rapporteur on Freedom of Expression and Opinion, David Kaye, has recommended that SMCs base their content moderation policies on

https://twitter.com/RMéndezRuiz/status/819734024893693953 (“@hurtadopaola, to say that he @frattigiovanni was a magistrate's advisor, when in fact it was OJ in general, is to miss the truth.”).

Interview 010.

UN GUIDING PRINCIPLES, supra note 5, Principles 11 and 13. The UN Guiding Principles are non-binding principles that were unanimously endorsed by the United Nations General Assembly in 2011.

---

202 Id.
203 FACEBOOK COMMUNITY STANDARDS: BULLYING, https://perma.cc/B3QM-VDML.
204 For example, during the production of this draft Report, Rodrigo Polo’s Twitter account (@RodrigoPolo) has been suspended. Polo is a political commentator who formerly worked for the Guatemalan Judiciary. Ricardo Méndez Ruiz (@rMéndezruiz), TWITTER (Jan. 12, 2017, 6:33 PM), https://twitter.com/RMéndezRuiz/status/819734024893693953 (“@hurtadopaola, to say that he @frattigiovanni was a magistrate's advisor, when in fact it was OJ in general, is to miss the truth.”).
205 Interview 010.
206 UN GUIDING PRINCIPLES, supra note 5, Principles 11 and 13. The UN Guiding Principles are non-binding principles that were unanimously endorsed by the United Nations General Assembly in 2011.
international human rights law. As the UN Rapporteur explains, aligning their policies with international human rights law would “enable companies to create an inclusive environment that accommodates the varied needs and interests of their users while establishing predictable and consistent baseline standards of behaviour.” This section discusses the international human rights and criminal law that counsels in favor of reforms that better address coded speech and which protect human rights defenders.

A. Coded Threats and Race-Based Incitement

Facebook and Twitter’s current policies do not provide a sufficient framework for responding to coded threats and race-based incitement to discrimination and violence. International human rights law requires states to prohibit race-based incitement to violence and discrimination. Article 20 of the International Covenant on Civil and Political Rights (ICCPR) requires that “[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.” Article 4 of the Convention on the Elimination of Racial Discrimination (CERD) is framed even more broadly, requiring state parties to prohibit “all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin.”

By and large, Facebook and Twitter’s policies are compliant with international norms prohibiting race-based inciting speech. Nonetheless, these policies do not address situations in which the racial nature of the speech is coded. In Guatemala, key figures utilize social media platforms like Facebook and Twitter to disseminate their hateful messages against defenders. Hate speech against defenders may not at first appear to meet the definition of Article 20 ICCPR and Article 4 CERD, since they appear to be targeted for their political activity, not their race or nationality. Nonetheless, some of the hate speech seems to be targeting indigenous activists on the basis of their protected group membership in a coded or implicit way. In Guatemala, many defenders are indigenous activists who are fighting for their land and economic rights. Prominent users that attack indigenous defenders as “communists” and “terrorists” are evoking the language of the Guatemalan genocide. Social media users also use words such as “Colombiano” (Colombian) to label defenders as outsiders aiming to agitate Guatemala’s political sphere, reflecting further discriminatory animus.

International criminal law also considers the context of speech in evaluating its impact. For example, international criminal law recognizes that the crime of incitement can be coded, and they base their evaluation of the meaning of speech on how a listener would understand it. The Statute of the ICTR, for example, provided

---

207 Kaye Report, supra note 8 ¶ 41.
208 Id. ¶ 43.
jurisdiction over the crime of “[d]irect and public incitement to genocide.” The ICTR Chamber held that speech could be “direct” when it specifically provokes another to commit a crime, and is “more than a mere vague or indirect suggestion.” In Rwanda, Akayesu gave a speech to a gathering of over 100 Hutus in which he rallied the crowd to eliminate what he referred to as the sole enemy. The ICTR was satisfied beyond a reasonable doubt that the audience had understood his call to fight against the “accomplices of the warriors” as a call to kill Tutsis. The Chamber in Akayesu explained:

the “direct” element of incitement should be viewed in the light of its cultural and linguistic content. Indeed, a particular speech may be perceived as “direct” in one country, and not so in another, depending on the audience. The Chamber further recalls that incitement may be direct, and nonetheless implicit .... [A]cts of incitement can be viewed as direct or not, by focusing mainly on the issue of whether the persons for whom the message was intended immediately grasped the implication thereof.

In emphasizing that consideration of context was essential to understanding whether the defendant’s speech constituted direct incitement to genocide, the tribunal acknowledged that threats are not always evident. A deep understanding of the way in which the intended audience would understand the speech is essential to evaluating the seriousness of the threat.

International criminal law also considers the corollaries of speech in the particular political circumstances in which it occurs. In Šešelj, for example, a Serb nationalist politician was charged with committing persecution as a crime against humanity by “[d]irect and public denigration through ‘hate speech’ of the Croat, Muslim and other non-Serb populations ... on the basis of their ethnicities.” The Appeals Chamber found that the defendant’s speech in the village of Hrtkovci was a violation of the fundamental right to the security of the Croatian people in the village, and that this was sufficiently grave to constitute a crime against humanity.

The Chamber asserted that Šešelj’s words had the effect of “infecting the village with hatred and violence, which led to the departure of Croatian civilians in the ensuing months, thereby expanding the wider attack against the non-Serbian population in Croatia and Bosnia and Herzegovina.” In reaching this conclusion, the Appeals Chamber looked at the impact of his words in the context in which they occurred. Šešelj’s speech created a climate of fear and insecurity in a system that was already plagued by violent attacks against a targeted group.

The authority of the speaker is a key

---

212 S.C. Res. 955, art. 2(3)(c) (Nov. 8, 1994).
214 Id.
217 Id.
element in assessing the risk that the audience will respond to a call to violence. International criminal tribunals have weighed the authority of the speaker in determining whether a speech act constituted an incitement and whether a reasonable trier of fact could connect a speaker’s calls to violence to subsequent crimes. The approach of international courts has been affirmed by numerous social science studies of authority and persuasion. The Dangerous Speech Project has integrated the strands of law and social science to produce a practical guide to evaluating harmful speech, noting, “When a speaker is unusually influential, this can make their speech more dangerous. Influence or authority can come from a variety of sources, including personal charisma, high social status, or official status such as political office...”

**B. State Persecution of Human Rights Defenders**

Facebook and Twitter’s policies protect human rights defenders from direct threats of violence but not hateful speech that significantly increases their risk of being targeted for a violent attack. International human rights law prohibits harassment of individuals based on their political opinion if the harassment is a) perpetrated or condoned by the government and b) associated with past harm or creates a fear of future harm. Article 19(1) of the ICCPR provides that individuals “have the right to hold opinions without interference.” In recognition of this right, refugee law provides an additional layer of protection to its exercise by affording individuals the status of “refugee” if they flee persecution based on their political opinion.

Harassment alone would not rise to the level of persecution based on political opinion; persecution requires that the individual have been exposed to in the past, or fear in the future, the infliction of “a certain level of serious harm” that is “more than mere discomfort, unpleasantness, harassment, or unfair treatment.”

Persecution of human rights defenders that is perpetrated or tolerated by the state violates international human rights law. Although state actors may legitimately engage in democratic political discourse and disagree with their opponents, they may not bring the resources of the state to bear on individuals to silence them because of their political opinion. In Guatemala, online harassment campaigns are part of a larger pattern of harmful behavior—including physical violence, threats causing emotional distress, and frivolous criminal charges—that are directed at human rights defenders in

---

218 See Blaškić TC §474, §601, Brdanin TC §273 and Gacumbitsi AC §181-2.
221 The Convention Relating to the Status of Refugees defines a refugee as one who, “owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country.” Convention Relating to the Status of Refugees, Jul. 28, 1951, 189 U.N.T.S. 137.
222 Martínez-Pérez v. Sessions, 897 F.3d 33, 39-40 (1st Cir. 2018).
retaliation for their political opinions. As part of their responsibility to avoid or mitigate human rights abuses with which they are linked, social media companies must ensure that state actors do not use their platforms to persecute defenders.

To constitute a violation of international law, the persecution must constitute state action. The online harassment of defenders in Guatemala may trigger the state’s responsibility as it could be argued that the government systematically fails to prevent, investigate, and punish private persecution.\footnote{Online harassment by private actors could also be an act of the state if the private actors are acting on the instruction of the government. International Law Commission, \textit{Draft Articles on Responsibility of States for Internationally Wrongful Acts}, Nov. 2001, Supplement No. 10 (A/56/10), chp.IV.E1, \url{https://perma.cc/EV9Y-}}\footnote{TSPB. Our research was not able to document the extent of the connection between the individuals engaged in online harassment of defenders in Guatemala and the Guatemalan government. \textit{Veładexe Rodríguez Case}, Inter-Am. Ct. H.R. (ser. L), No. 13, ¶ 172 (July 29, 1988).} States are obligated to protect individuals from violations of their rights by private actors, and systemic failure to do so constitutes a breach of their obligation.\footnote{TSPB. Our research was not able to document the extent of the connection between the individuals engaged in online harassment of defenders in Guatemala and the Guatemalan government. \textit{Veładexe Rodríguez Case}, Inter-Am. Ct. H.R. (ser. L), No. 13, ¶ 172 (July 29, 1988).} In Guatemala, the government does not protect human rights defenders from online or offline harassment. Activists and journalists in Guatemala are regularly targeted for harassment, violence, and even assassination because of their status as defenders. The government’s systematic failure to protect defenders violates the state’s international responsibility.
RECOMMENDATIONS

Greater attention can and should be paid to three dimensions of hate speech in Guatemala: the contextual meaning of speech, the use of social media by state-aligned actors to engage in political persecution of activists, and the measures used to identify potentially harmful speech. Organizational and structural changes may also aid in responding to the particular concerns that are raised in sensitive contexts such as Guatemala.

In addition to the specific recommendations below, we encourage SMCs to adopt the Santa Clara Principles in order to improve accountability and transparency in their content moderation policies and practices. The procedures that SMCs develop to deal with the problem of hate speech—including any changes they make to address coded hate speech against human rights defenders—will inevitably be abused by those seeking to stifle dissent. Reports have documented, for example, the way that states can exploit takedown mechanism to target and silence activists and opposition voices. Although it may not be possible to eliminate this kind of abuse, platforms can provide users with more information about what content is removed and what content remains, and they can offer additional opportunities to appeal and contest the decisions they make about content. Greater transparency about and accountability for their moderation decisions will not only help guide users but will also provide additional safeguards against the use of their rules against the populations the rules are intended to protect.

A. Context-Driven Content Moderation

Although Facebook and Twitter both recognize the importance of context in understanding the nuance and impact of harmful speech, their moderation policies and procedures do not provide a basis for assessing the geographically and culturally specific meanings of speech. Understanding speech in context may reveal embedded threats or racialized attacks that may not be otherwise apparent.

1. Heightened Scrutiny in Problematic Areas

We recommend developing policies and processes to identify and deploy greater scrutiny of speech in specially designated “critical countries.” The history of violence and discrimination in Guatemala, coupled with the polarization of society and ongoing political violence, create a climate in which normally permissible speech elevates the risk of violence sufficiently to warrant regulation. Recent research has identified ten distinct factors that influence the impact of harmful speech, including “a history of intergroup conflict between the in-group and out-group” and an overall increase in the “number of instances of inter-group violence” in the past twelve

---

months; “a major national political election in the next 12 months”; “significant polarization of political parties along religious, ethnic, or racial lines”; and emotional state of the audience.\textsuperscript{226} If these contextual factors are present in a particular country, SMCs would be justified in using broader moderation of harmful speech there.

2. \textit{Localized Guidelines and Personnel}

The meaning and intent behind the use of coded terms in Guatemala may not be evident to outsiders who do not know that country’s history of violence. We recommend exploring the feasibility of content moderation guidelines and procedures specific to countries identified as needing heightened scrutiny at particular times such as the ones noted above. This would allow posts to be vetted more contextually and to eventually develop strategies, rules, and procedures tailored to a geographic area. For example, in Guatemala, consulting with civil society groups who better understand the language, context, and subtle meanings of words used in that context would help platforms more accurately identify and evaluate harmful speech.\textsuperscript{227} Localized personnel would not only allow for more accurate moderation, but they could also alert companies to upcoming problems that require attention. The addition of moderators familiar with local languages and contexts may also promote greater expressive freedoms and avoiding unnecessary censorship of legitimate speech by providing moderators with a better basis for judging the extent to which content violates platform norms.

3. \textit{Monitoring and De-Platforming Key Accounts}

SMCs can also take greater account of the characteristics of the speaker in their response to harmful online speech. The impact of speech is predicted by several characteristics of the speaker, including the position of the speaker, the way in which the speaker is perceived by supporters, and the access of the speaker to the means to create or control communication.\textsuperscript{228} Messages from persons in positions of influence or authority are more likely to be acted on. Moderation policies might monitor the posts of influential speakers closely and take steps if a speaker in a critical country uses the platform to disseminate inciting messages. Identifying problematic accounts, especially of influential actors could be realized through the work of the localized guidelines and personnel.

B. \textit{Addressing Political Persecution in SMCs Content Moderation Policies}

Defenders in Guatemala have come under attack for their opposition to powerful groups and interests. Like other vulnerable groups, online attacks can leave defenders

\begin{footnotesize}
\footnotesubtext{226} Wilson and Kiper, \textit{supra} note 200; see also RICHARD ASHBY WILSON \& MATTHEW GILLETT, THE HARTFORD GUIDELINES ON SPEECH CRIMES IN INTERNATIONAL CRIMINAL LAW (2018), https://perma.cc/4MCG-2VQH.
\footnotesubtext{227} Twitter does something similar already, but it does not appear to go to the extent of analyzing context through local partners to better identify problematic Tweets. See Twitter Safety, \textit{Get the help you need from our partners around the world}, https://perma.cc/4KF7-RUA6.
\footnotesubtext{228} Wilson \& Kiper, \textit{supra} note 196. The impact of speech is also affected by the message (explicit or implicit calls for violent acts, dehumanizing messages or messages expressing disgust or revenge, and identifies direct threat and clear course of action). Twitter and Facebook’s content moderation guidelines do identify speech that, based on its content, might be associated with violence; they do not, however, address the contextual factors related to the specific country, audience, and speaker.
\end{footnotesize}
feeling unsafe offline. SMCs currently prohibit hateful conduct and speech against protected categories and characteristics such as race and gender, but not political views or affiliation. We recommend that SMCs develop specific guidelines to address government persecution of individuals based on their political opinion. Although it may seem initially difficult to distinguish online political persecution from ordinary political polemic, persecution requires the threat of serious harm. Even government criticism of human rights defenders—activists, journalists, advocates—would not rise to the level of persecution unless accompanied by a threat of serious harm.

Nonetheless, in contexts in which there are high levels of political persecution such as Guatemala, online harassment of human rights defenders should be prohibited. SMCs can accomplish this goal by allowing localized personnel to temporarily designate human rights defenders as a protected group. This designation would have to be accompanied by findings that defenders are the targets of political persecution in that country, and that the persecution is being carried out by or tolerated by the government. This can be determined by looking at whether defenders are targeted on account of their political views, or where the government orchestrates the attack or fails to prevent and punish the attacker. Relevant country-specific factors include the strength of rule of law norms and domestic institutions, histories of violence or attacks against defenders, current levels of violence and persecution, and the volume and coordination of online attacks, including through entities such as a net center.

C. Improve Identification of Harmful Posts

The current reliance on user flagging and user-reactive content moderation means that many problematic posts are left online. We believe that defenders in Guatemala generally do not report the hate speech they are subjected to, for a variety of reasons. Indigenous activists may not have an online presence, and defenders may not know how to report hate speech on social media. Moreover, the use of net centers and coordinated campaigns can result in large volume of posts and tweets which overwhelm any one individual’s ability to respond appropriately. Suggestions for improving flagging might include:

1. Verified Users as Endorsed Content Moderators

Trusted individuals or organizations can serve to identify problematic posts and alert social media companies through their content moderation processes. This would also allow for the identification of harmful posts without having to rely on targets alerting SMCs to posts which may put them in danger. Trusted flaggers can in some cases be more attuned to problematic posts if they have a greater working knowledge of the area, understand the context of the language used, and know how to navigate the content moderation process in general. Waze is an example of a company that uses data to track respected users as being trustworthy.

---

229 See G.A. Res. 429 (V), Convention and Protocol Relating to the Status of Refugees (July 28, 1951), article 1, defining “refugee” to include “being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion.”

230 Id. This notion is further covered by article 1 of the Refugee Convention, “is unable or, owing to such fear, is unwilling to avail himself of the protection of that country.”

231 Interview 014.
content moderators after having positive feedback on their accounts that the information they provide is accurate and reliable. A similar approach might allow Facebook and Twitter to effectively flag harmful speech.

2. Improved Appeals Process

Both Facebook and Twitter are in the process of developing appeals processes. For example, Facebook CEO Mark Zuckerberg recently announced the establishment of an independent body to review cases in order to ensure that its moderation decisions are made objectively and independently. Nonetheless, there remains room for improvement. At present, for example, Facebook allows users to appeal when one of their posts has been taken down, but not when reported content has not been removed, although it has announced its intention change this. Twitter also only allows users to appeal tweets that have been found to violate their content moderation policy, and it locks the account until the appeal has been determined. Allowing appeals for reported posts that remain online and giving users an opportunity to explain why a post should be taken down would make the process of flagging and moderation more transparent and comprehensible. As the Special Rapporteur on Freedom of Expression has recommended, SMCs should provide appeals processes both when content is removed as well as when it is not.

3. Online Training and Awareness of Social Media Literacy

SMCs already engage in outreach, but they could focus additional efforts in critical countries to raise awareness of how to recognize and report hateful speech on their platforms. SMCs could do this through messages on their platforms, or they might even require new users to watch a short video highlighting their policies regarding hate speech and processes for reporting before gaining access. This onboarding may also deter future harmful speech.

D. Organizational and Structural Changes

Facebook and Twitter are technology and advertising companies first and foremost, but their influence crosses geographical boundaries and affects the daily lives of billions of users. Developing and maintaining corporate social responsibility practices require the creation of a department with a team specifically dedicated to the promotion and protection of human rights. This team should be cross-functional and interdisciplinary and be equipped to analyze linguistic, cultural, political, and historical contexts to identify problematic areas where heightened scrutiny is required. These should be staffed by individuals knowledgeable and aware of human rights violations and sensitivities and well trained in human rights law and policy. Personnel from the human rights departments should be engaged at the highest levels of SMCs structures and be given a key role in how content moderation processes are

---

233 Zuckerberg, supra note 11.
235 Id.
236 TWITTER SAFETY, https://perma.cc/W9G5-3GP2; see also Sarah Perez, Twitter now lets users appeal violations within its app, TECH CRUNCH, https://perma.cc/4LYQ-ZTSW.
237 Kaye Report, supra note 8 ¶ 58.
CONCLUSION

Social media companies provide a platform for human rights defenders to share information and express opinions. At the same time, they are increasingly being used to target and harass human rights defenders, including journalists, environmental activists and lawyers. Ease of dissemination and difficulties with attribution can make accountability for online hate speech and threats challenging. Systematic, state-aligned campaigns to denigrate and indirectly threaten human rights defenders and marginalized communities contribute to a climate of violence and impunity that increases the risk of real-world violence. The combination of the volume and visibility of such speech is particularly toxic in countries, such as Guatemala, where there is political persecution of human rights defenders and systemic and historical impunity for violence against human rights defenders and marginalized populations. In order to mitigate the risk of violence from content on their platforms, social media companies should take the following steps, in line with international human rights and criminal law.

First, social media companies should, as a temporary measure, include human rights defenders as a protected category under their hate speech/harmful content policies in countries where defenders face persecution by the state or are not protected by the government from violent retaliation for their advocacy.

Second, social media companies should establish additional review procedures that will allow them to take better account of coded threats that contribute to a climate of violence, but which may not necessarily constitute a direct personal threat. In particular, these companies should (1) provide heightened scrutiny of content in problematic or sensitive countries; (2) engage localized personnel and guidance; (3) consider the impact of the speaker in evaluating the effect of speech on their platforms; and (4) improve flagging processes to facilitate the gathering of context-specific information.

Third, social media companies should utilize verified users as endorsed content moderators; create and implement online and social media literacy training programs; and create transparent appeals processes for challenging decisions to remove, or refusals to remove, flagged content. A transparent process is necessary to address the significant risk of over-inclusion that is present whenever entities engage in content moderation. A transparent review process is the only way to effectively ensure that such policies are applied in a manner that is narrow and proportionate to actual risk of harm.

Finally, the companies should provide heightened scrutiny based upon a review of relevant contextual risk factors drawn from the social science of incitement to violence, including but not limited to:

- “a history of intergroup conflict between the in-group and out-group” or an overall increase in the “number of instances of inter-group violence” in the past twelve months;
- “a major national political election in the next 12 months”; and
- “significant polarization of political parties along religious, ethnic, or racial lines”.

Implemented carefully, these recommendations could help ensure these companies deliver on their promise of
providing a platform for open debate without inadvertently enabling the denigration of vulnerable populations and the silencing of their greatest defenders.