A WORKING GROUP REPORT

Political Violence in the Digital Age

What Online Platforms Can Do to Mitigate Escalating Threats

Yaël Eisenstat
Director of Policy and Impact, Cybersafety Research Center

Justin Hendrix
CEO and Editor, Tech Policy Press

CRC CYBERSAFETY RESEARCH CENTER

Tech Policy .PRESS

JUNE 2026

Table of contents

Executive summary

As the United States approaches its 2026 midterm election cycle, the nation faces a volatile set of circumstances. Polarization and political violence have risen sharply: unprecedented levels of threats against public officials both on- and offline have coincided with a bout of assassination attempts and acts of targeted violence.

Polls consistently show that the vast majority of Americans do not support political violence, even if that margin is decreasing. And research has demonstrated that what leaders say has an outsized influence on individuals' willingness to act. Social media platforms are the megaphones through which political leaders and influencers reach the public, and the infrastructure through which divisive and potentially violence-inciting narratives are distributed, amplified, and targeted. The people who own and lead these online platforms are responsible for the choices they make in designing and governing them—and those choices have real consequences.

The people who own and lead these online platforms are responsible for the choices they make in designing and governing them—and those choices have real consequences.

A range of actors have demonstrated their increasingly dynamic use of platform products, including artificial intelligence (AI), to promote division and incite violence. Domestic violent extremists increasingly trade tactical and aesthetic playbooks with other online subcultures, and to a lesser degree with foreign terrorist organizations and hostile states. The technological architecture of the online information ecosystem has supercharged this cross-fertilization. There is evidence, including from Meta's own internal research, that the algorithmic targeting and amplification systems of large online platforms have at times led users—often young men and boys, who now perpetrate a large share of extremist and politically motivated violent acts—down radicalization pathways, helping violent narratives become more mainstream.

While perpetrators of violence are responsible for their own actions, and factors well beyond social media contribute to the increasing polarization we are experiencing, online platforms have enabled once fringe views to become more mainstream and are tools for recruiting individuals; distributing messages and narratives intended to incite political violence; and organizing groups that may carry

out violent acts. At the same time, online platforms' trust and safety policies and teams have been significantly degraded, as has the federal government's infrastructure for addressing foreign interference and election security.

It is against this backdrop that a group of experts—former tech company employees, policy makers, researchers and academics—convened to consider what role online platforms are playing in potentially fueling some of this violence and what they could and should do to mitigate these multidimensional threats. Even if comprehensive action feels far-fetched at this time, we felt it worthwhile to lay out what is needed to address the platform-level factors contributing to political violence, while avoiding undue moderation of speech. These goals are not mutually exclusive. On the contrary, the transparency measures we set out are designed as much to evaluate and incentivize speech protection as they are to assess platform enforcement against the worst and most dangerous forms of online mobilization—harassment, incitement to violence, terrorism, foreign hybrid and cognitive warfare, and more.

We recognize, however, that systems-based approaches are unlikely to take effect quickly, so we also lay out a set of immediate steps that can be taken to help mitigate an escalation of political violence as we head into another election.

Recommendations in brief, broken into immediate and longer-term categories:

WHAT PLATFORMS CAN, AND SHOULD, DO RIGHT NOW:

1

Prepare for the threat by defining, analyzing and addressing political violence:

Platforms should strengthen preparedness for political violence by investing in multidisciplinary expertise, robust threat assessment, and crisis response protocols for dangerous viral incidents. They should also increase transparency around election-related planning, clarify enforcement approaches, and engage external experts to assess readiness and accountability ahead of major elections.

2 **Clarifying and enforcing election and dangerous content policies:**

Platforms should publicly define and consistently enforce clear policies on election integrity, incitement to violence, harassment, AI-generated deception, foreign interference, and coordinated manipulation. This includes acting quickly against violent threats, temporarily slowing the spread of inflammatory viral content during crises, and improving cross-platform threat detection to identify emerging risks.

3 **Enforce rules equally, including against high-value users:**

Platforms should apply rules consistently regardless of a user's political affiliation, status, or audience size, including influential figures whose content may disproportionately shape public discourse. Companies should also prohibit monetization or paid amplification of content that incites violence or deliberately undermines democratic processes.

4 **Protect Targets of Online Threats:**

Platforms should prioritize protections for individuals facing heightened risks of harassment and violence, including election workers, public officials, and candidates. This includes faster responses to user reports, dedicated reporting channels for those in imminent danger, and clear commitments to enforcing policies against threats.

5 **Implement synthetic media disclosure and provenance standards:**

Platforms should expand efforts to detect and label deceptive AI-generated content, including violent or hateful synthetic media, by adopting common provenance and disclosure standards. Companies should also publish data on the prevalence of synthetic political media and related enforcement actions to improve public understanding and accountability.

6 **Publicly provide transparency reports and metrics on par with what is provided for the European Union:**

Platforms should disclose meaningful metrics about exposure to harmful content, including election misinformation, incitement to violence, and hate speech affecting U.S. users. Providing transparency comparable to EU standards would improve accountability and help researchers, policymakers, and civil society better assess and respond to online harms.

BROADER DESIGN AND BUSINESS CHANGES PLATFORMS CAN MAKE:

7

Stop incentivizing, profiting from, and rewarding divisive, inflammatory and dangerous behavior:

Platforms should reduce algorithmic and financial incentives that amplify inflammatory or dangerous content, particularly during politically sensitive periods. This includes redesigning recommendation systems to promote healthier engagement and removing opportunities to monetize content linked to election denialism, hate speech, or glorification of violence.

8

Build friction into the system:

Platforms should introduce design features that slow the rapid spread of harmful content, such as sharing limits, warning labels, or prompts encouraging users to pause before amplifying material. Evidence suggests these interventions can reduce misinformation and escalation during high-risk moments without requiring widespread content removal.

9

Improve data access and transparency:

Platforms should expand access to public-interest research data and provide greater transparency into content moderation decisions to improve scrutiny, accountability, and understanding of online threats. Restoring researcher access to platform data and offering clearer explanations for moderation outcomes would strengthen both public safety efforts and protections for free expression.

Contributors

This report was composed following a working group meeting of experts, and its recommendations represent a general consensus following input from the participants.

Contributors include:

Jeff Allen

Co-founder and Chief Research Officer, Integrity Institute

Luke Barnes

Senior Research Scientist, Technology & Democracy Program, NYU Stern Center for Business & Human Rights

J.M. Berger

Senior Research Fellow, Center on Terrorism, Extremism, and Counterterrorism at the Middlebury Institute of International Studies

Dean Jackson

Contributing Editor, Tech Policy Press

Daniel Kreiss

Professor, Hussman School of Journalism and Media and a principal researcher of the Center for Information, Technology, and Public Life at the University of North Carolina at Chapel Hill

Anika Collier Navaroli

Assistant Professor, Columbia University Graduate School of Journalism

Spencer Overton

Patricia Roberts Harris Research Professor of Law, George Washington University Law School; Founder and Faculty Director, Multiracial Democracy Project

Katie A. Paul

Director, Tech Transparency Project

Stephen Richer

Former elected Maricopa County Recorder

Dhanaraj Thakur

Director, Emerging Technologies Initiative, Multiracial Democracy Project, George Washington University Law School

Introduction

Political violence is on the rise in the United States. According to a summary of key trends from the Princeton University Bridging Divides Initiative, this rise is reflected in a number of statistics, from an increase in targeted violence and assassination attempts to an increase in the overall volume of threats and harassment against political figures from the local to the federal level. And with midterm elections looming, election officials are subject to what Brennan Center for Justice researchers call "increasingly violent rhetoric and attempts to criminalize their work that are fueled by disinformation and conspiracy theories."

An October 2025 poll by Politico and Public First indicates most Americans expect political violence to increase. Another poll from PBS/Marist suggests a growing number of Americans—now 30%—think violence may be necessary to get the country "back on track." At the same time, research shows that affective polarization—the gap between the positive feelings people feel toward their own group and distrust and animosity they feel towards others—has risen more rapidly in the US than in other Western democracies. This phenomenon appears to be uniformly distributed across the country.

The design, content moderation, and business decisions made inside technology firms that affect citizens' access and pathways to information and the shape and quality of online discourse deserve scrutiny, particularly in advance of contentious election cycles.

While phenomena such as political violence and polarization are not solely caused by social media, research suggests there is a troubling relationship between the dynamics of online platforms, including their algorithmic promotion of certain types of content for engagement, and the severity of the division in this country. Moreover, the prevalence of "content expressing antidemocratic attitudes and partisan animosity" has been shown to raise levels of affective polarization. Consequently, the design, content moderation, and business decisions made inside technology firms that affect citizens' access and pathways to information and the shape and quality of online discourse deserve scrutiny, particularly in advance of contentious election cycles.

The Prospect of More State Violence in the US

As *Extremism* author J.M. Berger has noted, "Definitions of political violence universally agree that the category includes state violence." A major contributing factor to the potential for political violence in the US in the period leading up the midterm elections is the demonstrated willingness of Trump administration officials, including and most significantly the President himself, to use inflammatory rhetoric and to justify state violence.

Deployments of military domestic security forces—from the Marines in Los Angeles and the National Guard in Washington, DC to Immigration and Customs Enforcement and Customs and Border Patrol agents in Minneapolis—are evidence of the administration's willingness to use force. As Carnegie Endowment for International Peace scholar Rachel Kleinfeld pointed out, "political violence from the government has grown and may be replacing what was previously right-leaning vigilantism. Government violence—especially from ICE—rose in 2025. In 2026, as of March 19th, ICE has killed two protesting Americans and at least 13 people died in ICE custody. In 2025, law enforcement used less-lethal munitions in 22% of protests with police engagement, almost four times the rate in 2024."

Tech firms have already shown a willingness to side with the state in such scenarios, complying with government demands to remove apps and social media groups used by citizens to report on such deployments and share information with neighbors. Last year, amid violent immigration raids, Google and Apple removed apps used to track the movements of ICE agents, while Meta removed a group that tracked ICE in Chicago. Platforms may be faced with difficult decisions in the event of state violence associated with the midterms, as Berger notes, putting them in the position of "determining whose political violence to support by endorsing and enabling the second faction's political violence against the first."

This concern is further compounded by the prospect that administration officials themselves could be significant sources of election misinformation, as was the case ahead of the January 6th insurrection at the United States Capitol. Platforms have already signaled that they do not want to aggressively fact check or enforce content moderation rules against election denialism.

Our working group of academics, civil society researchers, policy experts, and former tech employees gathered in January 2026 to assess the evolving threat landscape, what we have learned about the relationship between social media and political violence in recent years, and to compose recommendations for technology platforms to implement. We convened with a shared sense of urgency and recognition that the window for intervention is narrowing as polarization deepens, trust in institutions erodes, and acts of political violence become an increasingly normalized feature of American public life. We convened with an equal sense of urgency to chart a path forward that protects free expression and that fosters online spaces in which a genuinely free and fair competition of ideas can take place. This report captures the discussion and includes the recommendations where we found consensus from the working group.

We remain steadfast in our conclusion that these companies can, if they so choose, do more to help mitigate tech-fueled political violence.

This is the third in a series of occasional reports over the past six years. The first report was issued in 2020, when many of the social media companies still sought the expertise and input of research organizations and civil society. By the time of the second report in 2024, we were clear-eyed about the extent to which these companies would take civil society perspectives into consideration, but believed that, as we wrote, "there remains a critical role for independent experts to play in both shaping the public conversation and shining a light on where we believe these companies can act more responsibly." For this report, we remain steadfast in our conclusion that these companies can, if they so choose, do more to help mitigate tech-fueled political violence. Recent reporting appears to reinforce this: in the six months after Meta relaxed its content rules in early 2025, violent threats and hate speech against lawmakers on Facebook roughly quadrupled, according to research from the Center for Countering Digital Hate. As such, we are continuing to create a public record to push back against the idea that—as executives often argue in the aftermath of real-world violence that had an online nexus—placing blame on platforms is unfair.

Scope

01 **POLITICAL VIOLENCE:**

For the purpose of this report, we use the term political violence in its broadest sense, encompassing all forms of politically-motivated violence, including violence stemming from extremist and other violent online subcultures.

02 **ONLINE PLATFORMS:**

For this paper, we use the term "online platforms" to refer collectively to social media companies and a variety of messaging apps and forums. We did not focus on AI developers or generative AI companies, which also have serious challenges that merit equal consideration with regards to the potential fueling of political violence. We do, however, consider the distribution of AI-generated content and those potential effects.

03 **WHICH ONLINE PLATFORMS?**

Our recommendations are skewed towards the larger social media companies, which play a critical role through their algorithmic and recommender systems in the mainstreaming and normalization of increasingly violent and threatening discourse. Moreover, they possess greater resources to address systemic risks and harms to users than many medium or smaller-sized platforms. That said, a range of smaller fringe platforms play a critical role in the radicalisation of users and in the planning and operationalizing of attacks. As such, they warrant further scrutiny, in particular when it comes to the safety of minors. We acknowledge that small and medium-sized platforms can meaningfully contribute to political violence and election-related risks, and that those that are willing to address such concerns should be actively supported by the broader tech and online safety community through shared tools, expertise, and cross-platform coordination. But for the purposes of this report, we consider small platforms out of scope.

04 **TIME AND PLACE:**

While political violence is on the rise in numerous countries, we are focused on the United States for this report, with the 2026 U.S. midterm election as a potential flashpoint.

05 **STATE ACTORS:**

Our group also debated whether this report should address state actors that engage in or help fuel political violence, and what online platforms can and should do in the face of potential state abuse of power. We acknowledge that one of the significant changes in the spread of political violence in the US in recent years is the role that government actors themselves are playing. We do touch on some of these issues, but this paper does not wade into the larger legal and policy debates on government interference in platform actions or how these companies should address governments' roles in fueling political violence.

06 **OTHER STAKEHOLDERS:**

Our recommendations are focused on what the major platforms can do to ensure they are mitigating and not fueling potential political violence. There are equally important roles for government agencies, law enforcement, civil society, media, philanthropists and others to play to protect public officials, election workers, researchers, and other targets, but those are outside the scope of our recommendations.

Evolution of the online threat landscape

Over the past half decade, we have witnessed a profound transformation in the online threat landscape. Extremist groups as well as individuals sowing division and encouraging violence utilize a widening range of mainstream and fringe digital platforms and technological capabilities like AI to organize, communicate, and plan in a decentralized manner. Ideologically diverse extremist communities intermingle and cross-fertilize with a widening array of online subcultures. Domestic extremists and foreign terrorist organizations increasingly share online spaces and tactics with hostile states and cyber criminals, who have increasingly sought to leverage those networks, including to perpetrate violent attacks.

We are seeing an increasing trend of tech-fueled violence associated not with organized groups with a clear ideological program, but rather self-initiated individuals influenced by harmful online communities and de-centralized networks.

In parallel we are seeing an increasing trend of tech-fueled violence associated not with organized groups with a clear ideological program, but rather self-initiated individuals influenced by harmful online communities and de-centralized networks. This trend is plain to see in analysis from the Institute for Strategic Dialogue (ISD) of violence with a nexus to online radicalization across the US in 2025. This data—which showed a 27 percent increase in plots and attacks with a nexus to online radicalization compared to the previous year—revealed that only around half of such violent attacks or plots were associated with established ideologies, such as Salafi-jihadism or Neo-Nazi accelerationism. The remaining half were carried out by individuals with unclear or loosely defined beliefs, or inspired by broader subcultures of nihilistic violence—such as the "True Crime Community" (TCC) and the 764 network, which glorify mass murderers for their violent acts alone rather than any ideological affinity. These groups were the chief drivers of plots and attacks in 2025, linked to more than double the number of incidents as Salafi-jihadi ideology (the next largest category). This comes in a context of an increasingly young profile of violent radicalization, with Global Terrorism Index data showing children and adolescents accounted for 42 percent of terror-related investigations in Europe and North America in 2025, a threefold increase since 2021. It is crucial that frameworks for prevention and response reflect this evolving threat.

More decentralized harmful networks frequently operate across multiple services to reinforce group identity, share propaganda, and attract new followers. In the nihilistic 'Com' network, for example, large social media platforms such as X or Reddit are used to identify potential victims before perpetrators move conversations to private or encrypted messaging platforms, such as Telegram.

These trends come as public officials face growing threats and violent rhetoric online. ISD research found that violent online rhetoric targeting leading U.S. public officials increased more than threefold between 2021 and 2025—a 5 percent median increase each month. Threats against officials are growing across the aisle, with President Donald Trump being the target of nearly half of the violent rhetoric analyzed. These threats most often emanated from individuals with a range of grievances rather than organized extremist groups, a parallel to broader shifts towards more decentralized political violence. The data shows how threatening language is being normalized in online spaces, alongside a rise in incidents of political violence.

As for security threats to the election, we expect to see a number of similar online tactics as we saw in the 2024 cycle, including swatting; doxxing; bomb threats; threats to election officials, law enforcement and candidates; and a continuous slew of often AI-generated election conspiracies and disinformation. Despite platform policies to address these problems, including commitments to label deceptive election-related AI content, research identified consistent failures by platforms to do so. With the rolling back of federal government support for election security and countering foreign interference, such nefarious activities are likely to proliferate.

Trends in recent political violence acts

A number of political violence acts in the U.S. in recent years demonstrate some of the trends highlighted in the evolution of the online threat landscape:

Attempted Assassination of Donald Trump, September 15, 2024

The attempted assassination of President Donald Trump in Florida in September 2024—two months after the first assassination attempt in Pennsylvania—demonstrates how motivations for political violence can develop in the absence of organizational ties but still be shaped by repeated exposure to online content. The suspect had maintained an extensive online footprint, characterized by repetitive posting of anti-Trump sentiment, support for Ukraine, and engagement with conspiratorial narratives. In this case, online spaces did not provide a direct route to coordination or tactical planning. Instead, they appear to have reinforced grievances and sustained attention on a specific political figure over time. The pathway to violence remained individual, but it was not formed in isolation; research on online radicalization suggests repeated digital exposure can reinforce extremist hostility over time.

The Fatal Shooting of Charlie Kirk, September 10, 2025

The fatal shooting of Charlie Kirk highlights how political violence is increasingly culturally hybrid, with meaning constructed in part through online subcultures rather than coherent ideological frameworks. The suspect, Tyler Robinson, was reportedly motivated by opposition to Kirk's views on LGBTQ+ issues, but investigators found no evidence of involvement in organized extremist movements. Ammunition casings were inscribed with references drawn from gaming and internet culture, including phrases linked to Helldivers 2 and "Bella Ciao," alongside deliberately ironic or humorous content that the suspect described as "mostly a big meme," highlighting how ironic internet humor is used to meme-ify attacks.

Arson Attack of Governor Josh Shapiro's Residence, April 13, 2025

The arson attack targeting Pennsylvania Governor Josh Shapiro's residence demonstrates how political violence can emerge from digitally-reinforced grievances that translate into offline action. The suspect appeared to have been motivated by a combination of personal grievance, anti-government hostility, and conspiratorial beliefs. These views were expressed through online activity, including hostile commentary toward political figures, but not in a way that pointed to one platform in particular, which suggests engagement with a diffuse digital ecosystem. Here, the online dimension sits at the level of reinforcement rather than coordination: it provides a setting in which grievances are repeated, validated, and directed towards political actors.

Evolution of the platform policy landscape

Ahead of the 2024 U.S. election cycle, ISD produced a comprehensive comparative analysis of platform preparedness, policies, and enforcement and reviewed a series of safety measures from large platforms relevant to the prospect of election-related violence. Since that election, the major tech platforms have implemented what a February 2026 report from the Center for Democracy & Technology called "consequential changes" to their content moderation policies "that will influence the information environment ahead of the 2026 elections."

This evolution of the platform policy landscape has been accompanied by substantial reductions in the trust and safety functions at the major platforms, with rolling layoffs of thousands of trust and safety workers in recent years. For most platforms, this has meant relaxing their posture. As the CDT report put it:

> *"These changes suggest that the 2026 midterm elections will take place in a materially different information environment than recent election cycles; one marked by reduced reliance on traditional fact-checking, evolving creator and monetization frameworks, expanded latitude for hateful speech, and diminished transparency."*

At the time of publication, some of the major platforms had yet to show any public indication that they are taking specific actions to prepare for this volatile election cycle in the US. In February, Meta issued a blog post on its preparations, framing them generally as a continuation of its past election procedures and vaunting its move to community notes. Google/YouTube, TikTok, and X have yet to publish analogous statements on specific preparations for the U.S. midterms, even as each of these companies have made substantial policy changes that deserve more scrutiny.

There is also little evidence as yet that the major technology platforms intend to collaborate, as they did in 2024, to address the threats of AI-generated mis- and disinformation. The "Tech Accord to Combat Deceptive Use of AI in the 2024 Elections" was signed by 27 companies—including Google, Meta, and TikTok—at the February 2024 Munich Security Conference. Evidence of its impact was limited, and the arrangement has not been renewed since the 2024 cycle despite calls from lawmakers to do so. As CDT concluded, the "absence of a similar mechanism going into 2026 will stymie the ability to share best practices and information about threat indicators and responses while weakening the ability of AI developers, researchers, and civil society to detect and counter abuse of their tools."

In the cases of Meta and X, changes in policy appear to stem directly from the owners of the platforms themselves. Shortly after purchasing Twitter (now X), Elon Musk began reinstating previously banned accounts under what he called an "amnesty," including known far-right figures, QAnon followers, and others who were previously banned for violating platform rules. Later he reinstated white supremacist

and self-declared "proud incel" Nick Fuentes, amongst others. And Musk himself has engaged in spreading hate-based conspiracy theories to his more than 239 million followers. While Mark Zuckerberg has not gone as far or engaged with dangerous rhetoric online himself, he did announce substantial changes in Meta's content moderation policies, including relaxing restrictions on hate speech, in January 2025. According to a report from GLAAD, users reported a sharp increase in hateful content in the months following the changes.

Recommendations

Against this backdrop, platforms are unlikely to match the scale of their 2024 preparations, but there are still steps they could take that remain within reach ahead of November. While it may seem futile to provide recommendations in this context, we believe it is important to continue providing potential interventions, policies, and design changes that the companies could employ. We do not anticipate that platforms will simply "do the right thing" without the proper regulatory and business incentives in place, but we aim to publicly document what companies could actually do, if they choose, to act more responsibly. The days where companies could say "we could not have seen it coming" are behind us.

Beyond the lessons from political violence in the US—including the January 6th insurrection—social media companies should also draw on experiences from around the globe where their platforms were used to help spark mass violence. One of the starkest examples is one that should leave all of these companies asking what they can do to ensure they never contribute to political violence: Facebook failing to intervene to prevent its platform from being utilized as part of a genocide in Myanmar in 2017.

Below are nine recommendations that our working group agreed could all be implemented. The list is not exhaustive and includes recommendations we and others have made before, but bear repeating. Some will say that the companies are not financially or legally required to take any of these steps. However, as companies that are headquartered in and rely on all that open democracies provide, they do bear a responsibility to act as good stewards of the communication spaces they have come to dominate.

Where to draw the line between legitimate expression—political protest, anger, even some forms of hateful speech—and genuine incitement to violence is hard. So is keeping platforms from being exploited by extremists, terrorists, and hostile states that seek to spread division and provoke violence. These are difficult questions, and they belong to a broader national conversation about the role of social media in society—a conversation that should be grounded in better data and greater transparency going forward.

We prioritize concrete measures that companies could take to prepare for, monitor, and reduce the risk of political violence in and around the U.S. election, and how they might work with civil society and the broader community to mitigate that risk.

These recommendations take a narrower focus. We prioritize concrete measures that companies could take to prepare for, monitor, and reduce the risk of political violence in and around the U.S. election, and how they might work with civil society and the broader community to mitigate that risk.

These platform-level recommendations fall into two broad categories. The first includes things that all companies can do in real-time to address political violence and mitigate the threats that are manifesting online. The second includes the larger design and product decisions companies could make to change how their platforms are incentivizing certain behaviors and start prioritizing public safety and democracy ahead of pure growth and profit. The second bucket may sound aspirational, but they are all achievable if the platforms commit to doing all they can.

Category 1: What platforms can, and should, do right now

The policies and enforcement decisions of every online platform are critical during moments of heightened tension and potential violence. The platforms should:

1 Prepare for the threat: defining, analyzing and addressing political violence

Platforms should engage cross-sectorally to ensure robust standards are established for threat assessment and crisis planning, and ensure they have the multidisciplinary expertise required to understand the contemporary threat landscape. **Platforms need to put in place measures 1) to better understand the evolved threat landscape, 2) develop incident preparedness and response protocols in relation to any potentially dangerous viral online incidents, and 3) determine and make clear to the public what their policies are and how they will enforce them going into this election.**

This will require significantly more transparency around current platform approaches to scenario planning, crisis protocols and training, and proactive engagement with the wider expert community. Lessons can be drawn from cross-industry and multistakeholder protocols established to address viral online incidents in the context of terrorist and violent extremist attacks, such as the industry-led Global Internet Forum to Counter Terrorism (GIFCT) Incident Response Framework or The Christchurch Call Crisis Response Protocol. Research has shown that online threats spike significantly following real-world acts of political violence, risking the escalation of further violence. **Platform preparedness in relation to the viral spread of discourse that threatens to incite violence in such contexts can help de-escalate in moments of crisis and heightened polarization.**

Ahead of the 2026 midterms, it will be essential for a broader community of practice to be able to systematically assess adherence to stated commitments, evaluate whether platforms are maintaining a similar level of readiness compared to previous elections, and assess the extent to which they are communicating about these measures effectively to the public.

As political tensions continue to rise in the US, and as part of any efforts to prepare for a potential surge in violence-promotion on their services, platforms will also have to answer a difficult set of questions when it comes to who is the speaker and who is the target, including how systematically they choose to enforce their policies. For example:

- When government officials or agency spokespeople are engaging in inflammatory language, for example baselessly targeting a group of people in a way that could lead to violence against them, will platforms enforce their rules against them? Will they protect users against threats that might stem from government officials? If they decide not to, will they explain their rationale?
- How will doxxing and harassment policies apply to law enforcement officers, or to protesters? Will all be afforded the same protections, assuming that these protections still exist?
- What will happen if the federal government designates an activist movement or protesters as "domestic terrorists?" Will they then fall under platform rules for terrorist organizations? Will platforms be clear on how they are defining terrorist organizations and who exactly the definitions apply to? Will such organizations have a right to appeal platform actions?

How these companies answer these questions and how they choose to enforce their rules could have profound implications for public safety and the spread of potential political violence.

2 Clarifying and enforcing election and dangerous content policies

All of the large online platforms have certain policies in place related to election integrity; AI-generated content; inauthentic coordinated campaigns; violence and threats; even hate speech and foreign interference. However, these policies have changed and evolved and are at times confused and unclear. At a minimum, platforms should publicly disclose what their policies will be in all core related domains in the lead up to the election, especially in relation to the sharpest tip of dangerous activity—incitement to violence, harassment, doxxing, AI-generated election deception, foreign interference and manipulation—and enforce those policies consistently. We recommend:

- **Agree to aggressively and objectively enforce existing content moderation policies for harassment and incitement to violence:** As platforms move to a more hands-off approach, increasingly relying on AI over human intervention, even for serious threats, and with some favoring "community notes," platforms must be clear on where their red lines are. Community notes are not a fast or robust enough solution for content that goes viral in a crisis situation and could inspire violence. In particular, platforms should be clear that there is a zero-tolerance policy for threats against election workers, judges, politicians, candidates, and other public officials.
- **Downrank virality of inflammatory content:** During acute crises, platforms should limit the virality of content that triggers certain pre-determined alarms—as described, for example, in the above bullet—around political topics or other issues that explicitly encourage violence. Platforms can temporarily

limit the reshare velocity and slow the spread of viral content long enough for either automated or human review. This does not mean that the companies need to necessarily remove the content, but they can build in clear signals of when to cap sharing capability long enough for protective measures. (Here, more data about past interventions such as Facebook's "Break the Glass" measures would be instructive.)

- **Conduct cross-platform threat detection:** Political violence is often organized across platforms; sometimes the planning or incubating takes place on a smaller or medium-sized platform, then it is amplified on a bigger one. The more mainstream platforms can use weighted signals for when content from certain platforms starts to proliferate more broadly. At a minimum, using cross-platform signals could potentially trigger a higher bar for review or moderation.

3 Enforce rules equally, including against high-value users

Policies are only as good as their enforcement. And when that enforcement skews towards protecting high-value users and certain government-aligned actors—or even just the perception that this is the case—the entire system comes into question. People with the largest social media audiences have an outsized potential to influence the conversation, including around elections and other emotion-laden topics. There is ample research showing how what elites say creates a ripe environment for actions. As such, those with the highest potential to inspire or incite violence should, at the very least, be held to the same standard, not a lower one.

The top recommendation here is simple and bears repeating from our last report:

- **Online platforms should uniformly apply and fairly enforce their existing rules for all users, regardless of status, political orientation, or follower count.**

On X, this problem is compounded by the current blue check mark system. Anyone who pays a monthly subscription fee can benefit from higher visibility on the platform, regardless of the quality of their content. According to two recent reports from the Tech Transparency Project, even individuals on the U.S. terrorist list and other U.S.-sanctioned individuals had these paid, premium accounts. Additionally, posts from X owner Elon Musk—who has engaged in targeting people and spreading hate-based conspiracy theories—not only appear to be exempt from platform rules, but studies suggest have also been artificially amplified.

Mark Zuckerberg announced a relaxation of Meta's hate speech policies and an end to fact-checking in the US shortly after the 2024 election, but it is unclear how the company's controversial "newsworthiness" exception—which allows those deemed newsworthy to violate platform policies—is currently being applied.

While TikTok does have a "public interest" exception, the company has historically stated that it enforces its rules across the board, regardless of political status or influence. Recent statements from a TikTok whistleblower, however, allege that enforcement decisions prioritized politicians despite a range of dangerous content in order to maintain certain political relationships. It is unclear how the rules, and enforcement, might change under the company's new US ownership.

In addition to clarifying the above and explaining to the public how they plan to address these thorny enforcement issues, the companies should be clear that profiting from incitement to violence is not an option, regardless of stature.

- **Nobody, regardless of status or business importance to the companies, should be able to monetize content that incites violence.**
- **Nobody, especially those with large audiences or followings, should be allowed to use targeted advertising or paid promotion of content that deliberately undermines democratic processes or engages in inflammatory rhetoric that could inspire followers to violence.**

4 Protect Targets of Online Threats

Both online and offline violent threats towards election workers and other public officials—across the political spectrum—have dramatically increased in recent years, with female candidates, particularly women of color, receiving the brunt of online abuse. Research has repeatedly shown that social media users feel unsupported when they try to report threats to the companies. User-generated reports often go unaddressed for too long, if at all. Moderation decisions can feel arbitrary, especially as there is little transparency into how and why the enforcement (or lack thereof) decisions were made.

In addition to mitigating these threats:

- **Online platforms must be more responsive to user-generated reports, especially for those in imminent danger.**
- **Social media companies should, at the very least, make clear that they will not tolerate threats against public officials and election workers and put their money where their mouth is: resource the response,** including by creating a special process for these officials to flag threats against them, and/or having a dedicated line for officials to use.

From conversations with public officials who have been directly targeted, it appears that most (or all) of the major social media platforms have become even less responsive to targets in recent years. Where once there was at least a government contact in major platforms, it is not always clear if there is anyone to directly call anymore. As these companies move to automate more of their trust and safety work, this raises a potential concern that it will be even more challenging to reach someone in the moment when it might be most critical, as a threat is unfolding.

We understand and acknowledge the challenge for these companies to clearly distinguish between certain online behaviors and to weed out the noise from what constitutes a real threat. And the targets are becoming a larger range of communities, including public officials, law enforcement, schools, protestors, everyday citizens. Even trust and safety workers inside the companies themselves have faced threats for doing their jobs.

But when it comes to violent, politically-motivated threats against those who are working to secure elections and serve the public, these companies have an obligation to do more to provide actual support and respond in a timely and effective manner.

5 Implement synthetic media disclosure and provenance standards

At the Munich Security Conference in February 2024, tech firms including Google, Meta, TikTok, OpenAI, and others signed a pact to detect and label deepfakes designed to deceive voters, and later that year major platforms pledged to join the Coalition for Content Provenance and Authenticity (C2PA). While disclosures and provenance standards will never solve the problems that AI-generated content pose, these efforts should be encouraged and extended beyond elections to include media that depicts acts of violence.

In addition, platforms should be compelled to share more data on this phenomenon, including the volume of synthetic political media detected, the percentage of political content that bears valid C2PA credentials, data on enforcement actions taken, and other information that could help industry, policymakers, and civil society better understand these phenomena and craft solutions. Initiatives such as YouTube's recent effort to extend protections to politicians and journalists should be the standard rather than the exception.

6 Publicly provide transparency reports and metrics on par with what is provided for the European Union

Platforms should publicly provide data, through reports and metrics, about how many exposures to policy violating content occur on their platform, and in particular how many exposures occur for U.S. based users to content that violates their policies against election misinformation, incitement to violence, and hate speech. To properly respond to these risks, civil society and policy makers need to understand the scope of the problem that exists on platforms.

Platforms are already providing these metrics in Europe, as a result of various policies and regulations there. For example, in the Transparency Center for the Code of Practice on Disinformation, large platforms have previously disclosed the number of views that content violating election integrity policies has received before being removed from the platform. These metrics both help inform civil society's response to mitigate the harm and create accountability for the platforms to improve the safety of their services.

Category 2: Broader design and business changes platforms can make:

7 Stop incentivizing, profiting from, and rewarding divisive, inflammatory, and dangerous behavior

Online platforms should not incentivize users to engage in inflammatory and divisive behavior. There is a body of research, including leaked documents from Meta's own internal research teams, showing that engagement-based algorithms designed to maximize the time users spend on the platform, and by extension audience growth and advertising revenues, often reward the most inflammatory rhetoric. This is a design choice: platforms can take steps to reduce or counterbalance these incentives. Examples include:

- Moving away from engagement-based ranking and consider interventions such as bridging algorithms and other research-backed approaches.
- Penalizing the distribution of accounts connected to users or pages which repeatedly share violative or borderline content, especially that which is related to elections or civic issues;
- Slowing the distribution of posts which spread rapidly beyond the network of the account where they originated until those posts can be more closely reviewed;
- Adjusting the balance of false positives and false negatives in AI classification systems related to civic processes, violence, or hate speech;
- More closely reviewing or penalizing the distribution of web domains which routinely produce content which violates a platform's terms of service or community guidelines.
- Platforms should be transparent about when they take these actions, while ensuring they are not undermining security.

Artificial intelligence may introduce new tools or opportunities to implement or improve these approaches. However, these tools should not replace seasoned Trust & Safety staff, especially in the run up to a politically dangerous election.

Designing feed algorithms to reward different behavior is also within each of these companies' technical capabilities. Third-party nonprofit and academic recommendations for design choices which encourage pro-social behavior are plentiful, and provide examples for platforms to consider during sensitive periods, if not more generally.

Nobody should be able to profit off of spreading hate and inciting violence.

For those chasing follower counts and engagement in order to monetize their social media feeds, it is not difficult to understand that the more inflammatory the post, the more likely to build up engagement. Nobody should be able to profit off of spreading hate and inciting violence. There should be clear red lines limiting what content is eligible for monetization or profit-sharing.

Removing the ability to profit from election denialism, hate speech, glorification of violence, and potential incitement would diminish the financial and engagement incentives from those who use this type of content to grow their audiences and extract monetary value.

8 Build friction into the system

A proven way to slow down the spread of false information, conspiracy theories, hate speech and potential calls to violence is to build friction into online platforms. WhatsApp, for example, began implementing message forwarding limits in 2019 after false information spread via the platform led to multiple lynchings in India. Ahead of the 2020 U.S. election, Twitter built in friction in several ways, including adding warning labels, prompts asking users if they read the article before being able to retweet, and a prompt to "encourage people to add their own commentary prior to amplifying content." As Twitter explained: "Though this adds some extra friction for those who simply want to Retweet, we hope it will encourage everyone to... consider why they are amplifying a Tweet." Following the election, Twitter acknowledged that this friction actually did contribute to less election misinformation on the platform. But it reversed course and later removed these design features. Unfortunately, companies are not currently incentivized to build friction into their systems because their business models continue to rely on keeping people engaged as often and long as possible, not in slowing anything down.

9 Improve data access and transparency

Meaningful transparency and platform data access for users and independent public interest research is an essential foundation for detecting and mitigating threatening online activities, understanding the evolution of the threat landscape, and assessing how platform systems may be amplifying threatening and violent narratives. It is also a prerequisite for the protection of speech online, as greater user-facing transparency regarding content moderation outcomes, combined with strong researcher data access would enable stronger independent scrutiny of platforms' content moderation systems, process-

es, and ultimately decisions, including the evaluation of over-moderation or unjustified removals of content or accounts. Similarly, meaningful transparency and data access would allow for an assessment of the effectiveness and proportionality of any other election-related platform policies or interventions.

However, despite calls for better protection of online speech, and a fast-evolving threat landscape, both of which require better scrutiny and investigation, data access has been progressively shuttered by platforms, including Meta and X. In parallel, we face an increasing challenge from tech-fueled violence linked to more closed or private platforms—for example online spaces like Discord servers—where it is more challenging or less ethically justified for open source researchers to gain access to conduct robust real-time threat monitoring, while ensuring analysis does not drift into active engagement or even entrapment.

In the absence of mandated platform transparency—enshrined in social media regulations in Europe—there is considerably more that platforms can be doing voluntarily. The starting point should be providing access at-scale to data that is *already publicly available* to users of a platform (e.g. high-engagement content, comments, and key metrics such as likes or shares), but is difficult for researchers to collect systematically and at-scale in a way that enables more rigorous analysis. In many instances, this would simply require platforms to revert to previously available forms of access, as some of the platforms did use to provide more data access to public interest researchers. For example, X could waive the high fees introduced for API access for public interest researchers, or Meta could ensure its Content Library offers the same or better capabilities than its predecessor CrowdTangle.

Additionally, any platform could allow public interest researchers to independently access publicly available data via scraping by making minor changes to their terms of service. Civil society groups and independent researchers play a crucial role in providing evidence around election related harms but are in urgent need of increased legal safeguards to ensure public interest researchers are not threatened with litigation.

Finally, to enable more robust research into the risks posed to freedom of expression from platforms' content moderation systems and processes, platforms could provide users with greater transparency and more detailed explanations into why content is restricted or removed. This would give users more actionable feedback to understand why their content was restricted and better opportunities to dispute those restrictions when they feel they are incorrect. Researchers could then crowdsource examples of these explanations from groups of users who felt they had been unfairly moderated and analyse the associated content, or platforms could provide restricted access to selected researchers to analyse content moderation outcomes at-scale.

Conclusion

During a May 20 House Administration Subcommittee on Elections hearing titled "Examining Best Practices for Strengthening Election Security," Election Assistance Commission chairman Thomas Hicks told lawmakers that election officials across the country have "faced a growing number of threats to personal safety, including swatting incidents, suspicious packages, and bomb hoax threats targeting election officials and polling places." Such threats are taking a toll: a Brennan Center survey released in April found that nearly one in four local election officials is concerned about being assaulted at home or at work, and more than half worry that ongoing threats will make it harder to recruit and retain election workers.

With less than five months remaining before November's elections, technology platforms have a crucial role to play in helping to turn down the temperature, and they must be prepared in the event that political violence occurs. The recommendations we provide here should represent the floor, not the ceiling: minimum steps platforms can take in a broader commitment to an information environment that nurtures democracy rather than division and violence.