ISD | Powering solutions to extremism, hate and disinformation

US Election Platform Preparedness

Isabelle Frances-Wright, Ellen Jacobs, Clara Martiny, Max Read, Ella Meyer

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

Overview

Ahead of the increasingly polarized US 2024 presidential election which has seen an increasingly obscured information environment, and a significant increase in hate speech and violent threats, ISD has assessed platforms' policies, public commitments and product interventions.

ISD looked across six major issue areas: platform integrity, violent extremism and hate speech, internal and external resourcing, transparency, political advertising and state-affiliated media. While divergences were common across the platforms, one of the most alarming throughlines across all issue areas was a lack of transparency, particularly when it came to the details of platforms' policies, safety efforts and resourcing.

In the US, there is a lack of regulation that allows for the external assessment of platforms' adherence to their own policies and commitments by researchers and regulators. This creates an environment in which the public (and lawmakers) have a limited and obscured understanding of the nature and scope of harms present on the platforms voters use daily. Barring transparency regulation that requires the companies to disclose certain information on how they design and run their platforms, policymakers, academics and civil society are unable to fully assess how effectively platforms are enforcing their own polices and safeguarding users. It is impossible for anyone other than the platforms to fully know what is happening on the inside, from what information is amplified by their algorithms to up-to-date information on harmful trends. Because of this, the US elections are at inherent risk of being influenced or manipulated because election administrators, policymakers, law enforcement, and, critically, voters might not have the accurate information they need in the time they need it.

In the recent European parliamentary elections, the European Commission was able to release election integrity guidelines under the Digital Services Act (DSA) and then conduct stress tests with the platforms to ensure their readiness. The US, however, does not have the regulatory framework needed to require the platforms to adhere to guidelines, let alone preparatory stress tests or risk assessments. Even if it did, they could not be effectively evaluated without the US also having meaningful transparency requirements to ensure comprehensive compliance.

The lack of regulated transparency around safety policies also allows the platforms to publish vague and subjective harm thresholds to determine if violative content or accounts will receive punitive measures. However, without defining these thresholds explicitly, it is difficult to ascertain which actions would meet specific thresholds or which context matters. This in turn makes it impossible to know if a policy is applied consistently or if it is capturing most violative content. This vagueness creates space for the platforms to intentionally under-invest in monitoring and mitigation efforts, and make decisions in high-risk situations that do not align with their own stated policies. When they later face criticism, they can provide clarification or implement further changes to policies instead of creating and implementing robust policies in the first instance.

Further complicating matters is the lack of any meaningful AI legislation ahead of the election, and the expectation that platforms should decide how and when AI-generated content is permissible in political advertising.

Where platforms have clear commitments and policies, ISD intends to understand how well these are being enforced and implemented through a series of subsequent analyses.

Key Risks Identified

Policy exceptions for politicians pose significant risk and place policy enforcement at the whim of platform executives. Policy exceptions for political figures, candidates and news outlets on social media platforms can profoundly impact the dissemination and perception of information, especially in critical moments like the 2024 election cycle. These exceptions, while designed to account for public interest and freedom of expression introduce ambiguity and inconsistencies in content moderation, leaving platforms with significant discretion that can essentially nullify their own product policies when politically expedient. These exceptions led to a chaotic approach to safety policy in high-risk moments throughout the 2020 elections, and given the ongoing ambiguity and lack of transparency, we seem poised to face the same harms again.

AI-generated content remains a risk that platforms may be ill-equipped to deal with. When we assess the platforms' policies, we find find vague or confusing definitions, and an overreliance on users self disclosing AI-generated content. This is likely to contribute to an information ecosystem where AI-generated content and the specter of AI-generated content make it even more difficult for users to ascertain which information is accurate.

Divergences and obfuscations regarding fact checking create uncertainty. A lack of transparency and efficacy in platforms' fact checking mechanisms (whether through third-party fact checking partners or community flagging) may lead to a fractured information ecosystem in a polarized election environment.

Political Ads remain a potential vector of misinformation when exempt from fact checking. Most platforms' ads moderation and advertiser verification processes leave open the possibility of using ads as a powerful tool to undermine trust in election processes and results. Platforms also give advertisers the ability to target ads based on location, interests and demographic traits, allowing advertisers to pay to more efficiently reach their target audiences. Ads face different – sometimes lower – standards for fact checking and moderation, specifically for political accounts, making it easier for those buying ads to publish false or misleading information on the platform and reach a wider audience. Ad transparency features and disclosures also remain insufficient to assess how well ad moderation policies are enforced.

The lack of regulated transparency allows platforms to obfuscate on internal and external resourcing. The platforms' reluctance to disclose specifics about staffing, linguistic diversity and collaborations with external experts hinders public accountability and leaves questions about their ability to effectively address the complex, evolving challenges of election misinformation and foreign influence operations. Notably, Meta and X have reportedly reduced the number of staff working on election integrity, raising concerns about their preparedness for upcoming elections.

Threats from state actors remain hidden, with no mechanism to force public disclosure or resourcing. Transparency in reporting on identified disinformation networks and covert influence operations, particularly those run by foreign state actors targeting social media users, is critical for understanding and mitigating the impacts of misinformation on public discourse. Across the board, platforms could improve their transparency reporting by providing more details and reporting on disinformation networks and clear information on the evidentiary thresholds for identifying and acting against these networks. Additionally, platforms need to provide timely reports and updates to their users, allowing voters to understand the landscape in real time.

Lack of transparency in harmful content trends. Without mandated transparency requirements or comprehensive voluntary disclosures from the social media companies, there is very little information available about trends in harmful content on platforms, such as the type of false information or extremist narratives appearing. This is particularly concerning in an election context, where harmful content can directly influence voter behavior and there is limited time available to correct or mitigate the effects of misleading or harmful information.

Researchers are flying blind in an information ecosystem filled with threats. The ability of academic and civil society researchers to analyze social media's impact on elections and democratic processes is severely hampered by inconsistent and often restrictive API and/or data access policies across major platforms. Meta, X, YouTube, Snap and TikTok each present unique barriers to comprehensive research, ranging from partial access and project-specific approvals to high costs and exclusionary access criteria. These limitations, coupled with instances of platforms blocking potentially critical

research, create significant blind spots and undermine efforts to identify threats to elections and hold platforms accountable when they are not addressed. These limitations are more alarming when in some instances they are proactive choices by platforms who have retired tools or limited access to previously available data from independent research.

Influencer advertising policies regarding campaigning are often vague and unenforceable. Influencers are powerful messengers due to their large followings, ability to reach wide audiences and capacity to influence people's actions, whether it is buying a certain product or voting for a specific candidate. Yet the platforms policies governing the use of influencers for political purposes are often vague, confusing and hard to enforce, and while paid political advertising is not necessarily problematic, transparency on who is funding the content and their motivations is crucial.

Inconsistent policies around election outcomes may mislead voters. Platforms have taken markedly different approaches to moderating both false claims about widespread election fraud and premature claims of victory. Based on the platforms they choose to access; users may be exposed to contradictory and confusing information as to the results and legitimacy of past and future elections.

Ephemeral content remains a black box. Ephemeral content such as livestreams and 24-hour stories pose a significant challenge for researchers studying online harms, particularly during elections. The transient nature of this content, combined with inadequate archiving and access policies from platforms, creates substantial blind spots in research efforts. Unlike more permanent posts, these fleeting formats leave little to no digital trail for post-hoc analysis. This makes it nearly impossible for researchers to comprehensively track, analyze or study the spread and impact of harmful content shared through these increasingly popular mediums. This gap in observable data significantly hampers efforts to understand and combat real-time threats to election integrity and public discourse.

The platforms miss proactive opportunities to deamplify harmful content and effectively promote healthy civic discourse to engage voters, including the prevention of violence and hate. There has been an uptick in political violence around the world. This is often compounded during elections. Despite this notable increase in political violence and hate speech, the social media companies have largely maintained reactive policies that fail to address the root causes or stem its spread. Furthermore, they have decreased their investment in the tools and personnel needed to manage the spread of this content, making it more likely that it falls between the cracks and reaches more people.

In Depth Findings

Policy exceptions made for politicians pose significant risk and place policy enforcement at the whim of platform executives.

Policy exceptions made for political figures, candidates and news outlets on social media platforms can profoundly impact the dissemination and perception of information, especially in critical moments during an election cycle. Exceptions, while designed to account for public interest and freedom of expression, introduce ambiguity and inconsistencies in content moderation. This leaves platforms with significant discretion that can essentially nullify their own product policies.

Exceptions are also where the judgement and philosophical leanings of the platforms' executive teams begin to have an outsized impact. In high visibility, high-risk incidents, trust and safety experts' determinations have at times been superseded by executives lacking in the same relevant expertise. This poses the question, if policies do not apply when clarity and a rules-based approach is most needed, what purpose do they serve? Across the platforms, we generally see a patchwork of vague approaches, with exceptions for political figures, journalists and "newsworthy" content often ill-defined.

In 2020, **Meta**'s "newsworthiness" policy defined speech coming from a political figure as inherently newsworthy, regardless of subject matter or policy violations. This exemption also affected political ads, where political figures continued to hold policy exemptions, but were also exempt from third-party fact checking. Following the election, Meta reversed this policy, deeming that a political figure's speech did not inherently qualify as "newsworthy." While this may have been viewed by some as a more restrictive and therefore safer approach, the overarching newsworthiness policy, which had allowed political figures' violative posts to persist, remains in place. Despite this policy change, 13% of newsworthiness exceptions provided between June 2022 and June 2023 were for posts by politicians.

In Meta's current "newsworthiness" policy, they notably provide no clear definition of what is considered newsworthy, instead saying "We've found that determining the newsworthiness of a piece of content can be highly subjective. People often disagree about what standards should be in place to ensure a community is both safe and open to expression." They do say that "We remove content, even if it has some degree of newsworthiness, when leaving it up presents a risk of harm, such as physical, emotional and financial harm, or a direct threat to public safety."

X's policy for public interest exceptions provides arguably the most detail on both definitions and process. It applies only to accounts that are both "high profile" and run by a political candidate, elected official, political party or political appointee. X's definition of public interest states: "We consider content to be in the public interest if it directly contributes to understanding or discussion of a matter of public concern," such as by adding to a debate, adds information to their public role, adds context to ongoing geopolitical events or issues, or there is value in preserving it as a matter of public record." Still, this is highly subjective, and it would be easy to make the case that every post from a prominent political figure could meet these criteria. X also outlines a process in which senior Trust & Safety leaders make the final call on a piece of content, yet we have seen this has not always been the case in the companies approach to some content during the 2020 election.

X also states "We recognize the desire for these decisions to be clearcut yes/no binaries. Unfortunately, the reality is that they can't be. This is new territory for everyone — a service being used by world leaders to communicate directly to their constituents or other leaders, and at times, announce policy — and every decision we make sets a new precedent." It is hard to make the argument that this is "new territory for everyone" given the vast number of elections during which Twitter and now X has had to grapple with the potentially harmful speech of political figures. One area in which X provides useful transparency to its users is via a label applied to content that allows users to understand when content has been left up due to a public interest exception.

YouTube and **TikTok** both rely on a broad exceptions policy categorized under Educational, Documentary, Scientific and Artistic (EDSA) criteria. The publicly available descriptions of these criteria are extremely vague, citing the broad categories with very little by way of definition, which presents obvious issues when such a significant amount of content related to the election

could be categorized as "educational" or "documentary" in nature. It should be noted though that YouTube provides more detailed examples of what would fall under an EDSA exception than TikTok, as well as what would not receive an exception ("Instructions on how to build a bomb that's meant to injure or kill others" being a noticeably high bar).

In addition to more detailed definitions, what YouTube and TikTok both lack is the level of public transparency that X and Meta both provide as to when an exception has been made, via either content labels or within dedicated transparency reports.

When it comes to enforcement of political figures content at the account level versus the content level, TikTok has arguably taken the most balanced and transparent approach, relying on a clearly defined rules-based system. When a government, politician or political party account (GPPA) "reaches the strike limit set for all accounts, they'll be temporarily ineligible to appear in the For You and Following feeds for 90 days. If a public interest account posts content during high-risk times that promotes violence, hate or misinformation that could undermine a civic process or contribute to real-world harm, we may restrict that account from posting content for a period of 7 to 30 days, depending on the severity of the violation and surrounding risk." This approach imposes restrictions on a sliding scale, removing violative content and accounts from the ecosystem (particularly during sensitive events) while also balancing freedom of speech concerns by only permanently removing the account for the most severe violations.

AI-generated content remains a risk that platforms may be ill-equipped to deal with.

Tackling synthetic and manipulated media is a crucial battleground for social media platforms as they navigate a new wave of harms emerging in an age of AI-generated content. The effectiveness of their policies and detection mechanisms will be paramount, especially in the context of elections, where the integrity of the democratic process is at stake.

When we assess the platforms' policies, we find a landscape so mixed and vague it is hard to comprehend how these policies will be enforced, particularly without scaled detection capabilities. This is likely to further contribute to a polluted information ecosystem where AI-generated content and the specter of such content will present voters with an even greater challenge when it comes to assessing the origins or veracity of online information.

Meta's previous Manipulated Media policy, which was from 2020, stated, "We remove misleading manipulated media that has been edited or synthesized in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say." However, the glaring caveat to this policy, which was criticized by Meta's own Oversight Board, was that the policy primarily targeted manipulated media involving speech, creating enforcement gaps for content that manipulates visual elements without altering speech. In response to the Oversight Board's feedback, Meta will "apply "AI info" labels to a wider range of video, audio and image content," shifting the focus from the removal of content to labeling.

X's Synthetic and Manipulated Media policy, while detailed, is undercut by the lack of any guiding principles as to what enforcement mechanisms will be applied and when they apply. In some instances, X may delete a post which includes content that poses "a serious risk of harm to individual or communities." In others, it may apply a warning label, present a warning if a user attempts to share the post, reduce the visibility of the post, turn off likes, turn off replies, or provide a link to X's policies.

TikTok's Edited Media AI-Generated Content policy, like other platforms, relies heavily on the use of labeling and somewhat vague definitions. TikTok does not consider misleading AI-generated content of a public figure to be inherently harmful, as long as it includes "the AIGC label" or "a clear caption, watermark or sticker of your own," which presents obvious challenges in that captions are often not read and portions may be hidden if they extend beyond a certain length. Additionally, captions are not retained if a TikTok user downloads the post and subsequently re-uploads the media. TikTok does say that, even when appropriately labeled, content showing public figures in certain contexts, such as making an endorsement or being endorsed, is not allowed. According to TikTok, an "endorsement signal" - which would lead to the content being removed rather than just labeled – could include actions such as "non-verbal response cues." TikTok's policies should be further clarified to ensure that there are consistent standards, particularly around what constitutes "significant harm" or an exhaustive list of contexts (such as examples of endorsements) that are disallowed for AI-generated content of public figures.

Snap's policy stands apart from the other platforms in its clarity and ease of comprehension where it states they prohibit "manipulating content for false or misleading purposes." Under its guidelines on harmful, false or deceptive information, the company clearly states that

its "teams take action against content that is misleading or inaccurate."

Under its misinformation policies, **YouTube** prohibits technically manipulated or doctored content that misleads users and "may pose a serious risk of egregious harm." Under its election misinformation policies, it reiterates that "certain types of manipulated content" would not be allowed. While the election misinformation policies do outline some categories that are explicitly prohibited, it is still unclear what instances of manipulated media of candidates, public figures or election officials would be prohibited.

Many of the platforms have now chosen to rely heavily on labeling AI-generated content, rather than downranking or removing the content. While labeling can play a role in a platform's policy on AI-generated content, it cannot be the sole or even most important mechanism. In instances where creators are required to self-label AI-generated content, this creates a risk that creators could forget or in the case of bad actors, willfully choose not to self-label. An obvious example here are extremist groups that are already using AI technologies to create propaganda. Even when platforms label this content once it has been identified, the label can be easily overlooked, buried in a caption or may not provide enough context. As we have seen recently, if a platform relies too heavily on labels, it can begin mistakenly labeling content that is not AI-generated, which could further users' confusion over what is real or fake.

For TikTok, X, YouTube and Meta, the reliance on broad undefined terms within their policies risks chaotic enforcement, accusations of malfeasance or censorship, and a dramatically weakened information ecosystem. Clearer policies and consistent enforcement will be critical to safeguard against the dissemination of doctored content and aid users to begin to reliably identify manipulated media. With the proliferation of AI-generated content and the wide accessibility of sophisticated tools growing, the platforms must respond with robust, transparent and enforceable policies.

Divergences and obfuscations regarding fact checking create uncertainty

Collaboration with third-party fact-checkers represents a critical component of the strategy employed by social media platforms to combat harmful, false information. In fact, it often serves as the bedrock for platform integrity policy enforcement. Platforms frequently tout the number of fact-checking organizations they have partnerships with, but how these partnerships work and if they are effective is completely opaque. With TikTok admitting to having fact checked only 15 pieces of content in 6 months in the EU member states in which they have coverage, and YouTube producing seemingly false reports to the European Commission, it has become clear that some of these partnerships may be nothing more than window dressing,

YouTube utilizes a combination of automated fact-checking, through the Schema.org ClaimReview markup and collaborations with third-party fact-checkers supported by a significant financial commitment to the International Fact-Checking Network (IFCN). Despite these efforts, the platform faced criticism for an "insufficient" response to misinformation, pointing to challenges in the scalability and impact of its fact-checking initiatives.

Meta has established a network of partnerships with 90+ organizations certified by the IFCN to assess content across more than 60 languages. This collaboration is used to allow Meta to label and demote content identified as false, providing users with links to fact-checking articles. However, the turnaround times for these verifications remain unspecified, raising questions about the timeliness of the platform's response to misinformation.

X has taken a different approach, opting to have no partnerships with global fact checking organizations and instead relying on its community notes feature. Criticisms of community notes have primarily been that they are either slow to be applied, inconsistently applied or contain incorrect context. However, the sheer volume increase in members of the community notes program may help alleviate some of these criticisms. Specifically, X informed ISD that its program membership increased to 600,000 users from 100,000 users since October 2023. This increase in volume is a positive sign and the US election will prove a significant test of the evolution of this feature.

Snap differentiates itself by relying on an in-house fact-checking team, boasting a median turnaround time of under 1 hour for all categories. This rapid response mechanism positions Snap uniquely in terms of the speed of its fact-checking process.

TikTok has expanded its partnership with global fact-checking organizations, working with 17 partners to assess content accuracy in over 50 languages. However, like its peers, TikTok has not disclosed specific turnaround

times for its fact-checking process, and does not provide users specific insights on the results of fact checks when content may be false but not reach the severity of a content violation, unlike Meta and X.

The 2020 US elections tested the capacity and effectiveness of all platforms' fact-checking partnerships. Despite the partnerships, the platforms still struggled to control or provide context to a significant volume of false claims which circulated, many of which may have actively disincentivized voters from participating in the election.

Political ads remain a potential vector of misinformation when exempt from fact checking

Social media platforms' varied approaches to political ads do little to mitigate the risks posed by paid amplification of election disinformation - and in some cases platforms' policies make ads an easier conduit to spread false election claims. All platforms claim to ban political ads that include false information about voting methods, eligibility and premature claims of victory. However, beyond these easy-to-detect forms of voter suppression content, the platforms' ads policies leave wide gaps for bad actors to exploit.

Research has shown that even the most restrictive policy approach taken by any of the platforms – **TikTok**'s comprehensive ban on political ads – is not airtight; political ads and sponsored political content have still managed to pass through the approval process. If TikTok has no systems in place to fact-check political ads because its policies disallow them, but automated detection systems fail to flag at least some political ads, this leaves the platform vulnerable to ads containing false claims that it will not have the resources or expertise to mitigate. TikTok also does not have a searchable ad library for the US, which makes it extremely difficult to assess the volume of political ads accidentally making it through the approval process.

Other platforms, like **Meta**, have taken more nuanced approaches to moderating political ads and allow politicians to run ads with false claims so long as they are in the politician's own words. Political ads on X are subject to fact-checking through community notes, which have the limitations noted in the previous section. Snap provides no exemptions from fact-checking for political advertisers and YouTube does not have a clear fact-checking policy but does prohibit false information that undermines trust in democratic processes, including in political ads.

Meta's approach to election denial claims in political ads is illustrative of the haphazard and insufficient approach platforms have taken on election disinformation in political ads. Meta's policy bans content that calls into question the validity of upcoming elections but allows false claims about legally certified results of previous elections. This opens the door for advertisers to obliquely call into question whether future elections will be fair by promoting disinformation about past elections.

This year will also be a major test of how well platform approaches to AI-generated content and synthetic media in ads mitigate harms in an election context. Meta and YouTube require political ads that use synthetic or selectively edited media to disclose that fact and include a label on the ads. X on the other hand does not require labeling for ads that include synthetically generated content but does claim that ads are subject to the platform's general Synthetic and Manipulated Media policy that prohibits "synthetic, manipulated or out-of-context media that may deceive or confuse people and lead to harm ('misleading media')."

Finally, the tools platforms have put in place in the name of transparency are welcome, but insufficient. As ISD and others have documented, ads have run on Meta's platforms without required 'paid for by' disclaimers or with deceptive disclaimers. The ad libraries platforms put in place under pressure from researchers and policymakers have deficiencies that limit their effectiveness, making it difficult to assess how consistently and quickly they are enforcing their ad standards. The data on US-targeted ads made available by X, for example, does not have an interactive ad library akin to the one it is required to have for EU-targeted ads. Meta keeps ads in its library that have violated policies, but does not specify what policies were violated.

The lack of regulated transparency allows platforms to limited information on internal and external resourcing

The intentionally limited information on the scope of and investment in election integrity efforts from the platforms raises concerns about the companies' preparedness – and willingness - to address the multifaceted challenges present ahead of an election.

Meta had previously indicated the presence of a dedicated team focusing on election integrity, which it highlighted during the 2022 midterms. Yet, following the 2022 election Meta reportedly made cuts to the team despite launching more surfaces that need

support, such as Threads. The company is running its Election Operation Centers, which has members of its threat intelligence, data science, engineering, research, operations and legal teams that conduct real-time monitoring. However, these types of cross functional working groups may distract from the absence of dedicated teams, and it is unclear how much time members of the groups dedicate to election integrity efforts.

Similarly, **Snap** also relies on a cross-functional working group which includes misinformation, political advertising and cybersecurity experts but limited information is available on the group's size or allocated dedicated work time.

In September 2023 it was reported that X's election integrity team was reduced by half, Musk tweeted that same month "Oh you mean the "Election Integrity" Team that was undermining election integrity? Yeah, they're gone." This move, particularly ahead of crucial election periods, signals potential vulnerabilities in X's ability to effectively manage election integrity and misinformation challenges. X CEO Linda Yaccarino acknowledged the insufficient number of moderators the company now has at an earlier Senate Judiciary hearing, saying they needed more. Like Meta and Snapchat, X utilizes a cross-functional elections working group.

YouTube, under Google's umbrella, benefits from a Global Election Integrity team, indicating a broad and potentially well-resourced approach to election-related issues. However, the specific focus on US elections and the team's size remains less defined publicly.

TikTok does have a dedicated Election Integrity Team, "which is staffed by multi-disciplinary experts in democracy, elections, civil society and technology."

YouTube, Snapchat and TikTok have not made cuts to their election integrity teams, suggesting a potentially more stable foundation for their ongoing and future efforts in this domain.

In addition to internal resourcing, external partnerships can also play a critical role. Partnerships with outside experts or vendors who can perform dedicated monitoring can help make a company's ability to detect and mitigate harm more comprehensive and effective. However, information about these partnerships is sparse.

Meta provides public information about its work with external experts that support identifying hate speech and violent extremist content, including a partnership with the Middlebury Institute of International Studies, an ongoing partnership with Search for Common Ground, and their Trusted Partners program. YouTube specifies that it works with Google's Threat Analysis Group to combat influence operations from foreign adversaries. Snap confirmed it works with third-party vendors services as well. TikTok has its US Content Advisory Council. Aside from these specifications, the platforms generally provided very little clarity about what their coordination with external parties looks like, such as whether their engagement is a one-time workshop or a more consistent consultation.

As social media platforms continue to play a pivotal role in shaping electoral discourse, the structure, scope and resourcing of election integrity teams will be crucial in navigating the challenges of election interference. The recent staffing cuts at Meta and X, coupled with the general lack of transparency across platforms, highlight areas of concern that need to be addressed proactively. Strengthening these teams, ensuring their linguistic diversity, and committing to transparent disclosure of election integrity efforts are essential steps for safeguarding the integrity of future elections.

Responses to threats from state actors remain unclear with no mechanism to force public disclosure or increased resourcing

In examining the commitments and practices of Meta, X , YouTube, Snap and TikTok regarding the ongoing publication of findings on disinformation networks, there is a lack of detail, transparency, and, in some cases, delayed reporting timelines that prevent platforms users from knowing if they have been exposed to content from a state-backed influence operation until months after an election has passed.

Meta currently commits to publishing Adversarial Threat Reports and data on Coordinated Inauthentic Behavior (CIB) networks quarterly through its Threat Disruptions page. While this represents a step towards transparency, the criteria for acting against these networks remain unclear, including the evidentiary threshold required for Meta to publicly disclose and address a disinformation network. Additionally the platform only reveals identified disinformation networks to its users months after it identifies and removes them. This could be a problem in the upcoming US elections as by the time Meta publishes its Fourth Quarter findings for 2024 (likely in February 2025), major key election dates will have already passed, and users will until then be unaware of

disinformation networks that targeted them during October or November.

X previously offered partial transparency through biannual Transparency Reports, which included sections on disclosures and election integrity. This ceased in 2022. X's future commitment to maintaining the depth of these reports is uncertain, raising questions about the platform's dedication to transparently sharing insights on disinformation efforts. In its response to ISD, X stated that their "goal is to return to a more regular cadence of global public disclosures," but did not indicate when the global public disclosures would resume.

YouTube, under Google, provides content removal data via the Google Transparency Report center and a quarterly Threat Analysis Group (TAG) bulletin and, unlike Meta, updates in between the bulletin with in-depth details of disinformation networks found on YouTube or other Google products.

Snap does not commit to publicly releasing findings on identified disinformation networks, representing a significant gap in transparency compared to its peers. In its response to ISD, Snap claimed that the platform's architecture "makes the use of our platforms by [disinformation networks] highly inefficient." It is unclear how often Snap monitors for disinformation networks on its platforms.

In May 2024, **TikTok** introduced a new dedicated, monthly Transparency Report on covert influence operations and disinformation networks. The report provides top-level information on how networks were detected (internally vs externally), network accounts, total followers and a brief description. Unlike the other platforms, TikTok publishes these statistics monthly – as of early July 2024, the platform has already published through May 2024. Additionally, TikTok provides a brief note on the removal of accounts associated with previously disrupted networks attempting to re-establish their presence on TikTok. While the report does not include details to the level of Google's TAG insights, it provides updates much more regularly than all the other platforms assessed.

Enhancing the detail, transparency and timeliness of reporting on disinformation is essential for social media platforms to rebuild public trust and effectively counter the effects of politically motivated influence operations.

Lack of transparency in harmful content trends

Due to a lack of transparency, harmful content trends often only emerge when whistleblowers come forward or enough widespread harm has occurred that a pattern becomes noticeable. This is especially problematic in election contexts, where time is of the essence and misleading or harmful information can have an outsized impact on democratic processes.

As reported in the previous segment, there are no official plans for **X** to share information publicly on identified disinformation networks or malicious actors, even during the election period, nor would they share information on harmful content trends relevant to the election e.g. an uptick in violent incitement aimed at disrupting electoral processes.

YouTube does provide more information on harmful content trends, primarily through its Threat Analysis Group (TAG). TAG's quarterly bulletin includes aggregate data on malicious actors and other harmful content trends; however, unlike its reporting on coordinated disinformation networks, this information does not include more in-depth updates in between bulletins, and does not seem designed to provide real time content theme updates during time bound events such as an election campaign period.

Meta publishes its quarterly Community Standards Enforcement Report, which includes information on dangerous organizations, hate speech, violence and incitement, violent and graphic content, and bullying and harassment, though like YouTube and X, disclosures on recent violative trends relevant to the election are unavailable.

In the 2020 election cycle, **TikTok** published daily updates on violative content trends being actioned within their Election Operations Center. It is unclear if this approach will be taken again in 2024.

Snap does not release public findings on malicious actors or harmful content trends. It cites a similar explanation for its lack of reporting on disinformation networks, saying it has not observed malicious actors or harmful content trends that relate to election integrity on its platform to date.

ISD is also not aware of any effort from YouTube, TikTok or Snap to inform users whether they liked, shared, commented on, or interacted with any content that turned out to have violated policies. For example, when

a user on X likes, replies or shares content that is later fact-checked by Community Notes, the user receives a notification. Meta also notifies users who try to share content that has been proven false or have shared content that is later proven to be false.

Researchers are flying blind in a polluted information ecosystem

The work of academic and civil society researchers is crucial in identifying state and domestic influence operations, threats of political violence and the overall impact of social media on democratic processes. However, access to platform data via Application Programming Interfaces (APIs) or other comparable means remains a contentious issue, with various platforms adopting differing policies that significantly affect the ability of researchers to conduct independent, timely and impactful studies.

Meta offers partial API access to academic researchers, emphasizing partnerships while implementing an authentication process to safeguard user data. In August Meta shut down CrowdTangle, the platform's transparency tool, despite calls from researchers particularly concerned about its deprecation ahead of the US election. Its replacement, the Meta Content Library rolled out with an application that is arduous and tailored to academic organizations, making it difficult for other researchers to gain access in time for the election. It is also not available for journalists, who used CrowdTangle for their reporting.

X significantly altered its API access policy in early 2023, moving from free to high-cost paid access, thereby limiting the data volume available to researchers. This change has raised concerns about the sustainability of independent research on the platform, especially for those examining election integrity and misinformation.

YouTube provides API access primarily to academic researchers affiliated with higher education institutions, excluding a broader range of civil society organizations that often engage in more rapid and potentially critical research. This exclusion represents a gap in the ecosystem of independent platform analysis.

Snap does not provide API access for external research, creating a blind spot in the understanding of its impact on elections and misinformation.

In the 2020 election cycle, **TikTok** published daily updates on violative content trends being actioned within their Election Operations Center. On September 4th, TikTok announced they would be taking similar measures this election cycle stating "To bring ongoing transparency to our work, today we've launched a new US Election Integrity Hub in our Transparency Center. We'll be providing continuous updates on steps we're taking to protect TikTok during the elections, including misinformation we're taking action on." Based on the updates provided in the initial rollout, it appears TikTok is leading in terms of providing real-time updates to its users on election related safety actions.

While platforms have legitimate concerns about user privacy and data security, the current approach to API access for researchers—particularly the exclusion of civil society organizations and journalists and the imposition of restrictive terms—limits the scope and impact of independent research on social media's role in society.

Influencer advertising policies regarding campaigning are vague and unenforceable.

Influencers play a critical role in shaping conversation in online communities across various social media platforms. With a video or a post, an influencer can encourage their millions of followers to buy a product, download an app or follow a new trend. Social media companies are more than aware of the attention and revenue an influencer can harness. In turn, they invest in influencers, drawing them in with funds and income opportunities for content creation in hopes of attracting more users (and therefore revenue). Similarly, political campaigns are increasingly recognizing how beneficial it is for a celebrity or influencer to endorse a candidate and promote voting or other electoral processes. And while paid political partnerships are not necessarily harmful, a lack of clear policies on paid partnerships and self-disclosure requirements certainly can be.

As mentioned in previous sections, each platform varies in its rigidity over political ads – **TikTok** has been the most stringent, banning political and issue-based ads and paid content on its platforms entirely. In 2022, ahead of the US midterm election, TikTok published a blog post reminding its users and influencers that paid political content was prohibited. The platform claimed it worked to "educate creators about the responsibilities they have to abide by our Community Guidelines and Advertising policies as well as FTC guidelines." Yet earlier this year reports emerged highlighting how a super PAC supporting President Biden's then ongoing reelection campaign paid micro-influencers in battleground states to post content encouraging people to vote in local and

state-level elections. This content was later removed by TikTok following Politico's inquiry.

YouTube's policies on monetizing content related to elections and issue-based topics are also vague and hard to navigate. YouTube claims to have various policies and guidelines for influencers to monetize videos, but ISD analysts found them difficult to identify within Google's website, and generally unclear.

Other platforms have less stringent and comprehensive policies when it comes to paid political partnerships. For **Meta**, paid partnerships are allowed, but the political campaign or group must be registered in the Meta ad library. Snap and X allow political advertising but do not allow influencers to post commercial content related to elections or issue-based topics. **Snap** restricts paid promotion of political messaging to "traditional ad formats." **X**, on the other hand, explicitly states that "geo-political, political, social issues or crises for commercial purposes" are not allowed under its Paid Partnerships Policy.

With a complete lack of regulation and disclosure requirements for social media influencers and content creators from federal bodies such as the Federal Election Commission, it is crucial social media platforms have clear language and policies. Otherwise, the content is harder to enforce and users are misled about the authenticity and motive of the content they come across on their feeds.

A divergence in approaches to election denialism could affect voters' trust in the 2024 election

The 2020 election cycle saw one of the largest declines in voter trust in the electoral process in recorded history. While misinformation may have come from specific actors, the platforms also played a role in disseminating false claims about the electoral system by acting too late to implement policy changes or take enforcement action to 'stop the spread.'

The platforms have three mechanisms by which they can prevent the further erosion of electoral trust: policies that prohibit false claims that an election will be rigged/fraudulent, effective enforcement of those policies, and the promotion of authoritative information. The third mechanism is the only one where all major platforms have made commitments; each says it has or will have features directing users to authoritative sources such as vote.gov.

Three platforms (**Meta**, **X** and **YouTube**) allow content that falsely asserts that the 2020 election was illegitimate, in YouTube's case with labels linking to the 2020 election Wikipedia page below relevant videos. **Snap** prohibits these claims and **TikTok** either downranks or removes them. While all platforms have some form of policy prohibiting false claims of mass election fraud in upcoming elections, the vagueness of the policy language – and the fact that three platforms allow unchecked false claims about previous elections – creates doubt about how they are enforced. For example, Meta states it will prohibit misinformation about "whether a vote will be counted," but it is not clear how that policy applies to claims of widespread fraud. Similarly, YouTube prohibits "false claims that materially discourage voting," and X prohibits "misleading information relating to votes not being counted." How these policies are interpreted and enforced will play a significant role in determining whether voters trust the official results of elections in 2024.

The other major information risk platforms will have to grapple with is the potential for premature claims of victory. Media outlets will "call" races based on projections before the official electoral verification process concludes. Platforms, therefore, must navigate the delicate task of deciding which signals to use as benchmarks for allowing claims of victory.

Platforms have taken various approaches to this issue. **TikTok** and **Snap** outright ban premature victory claims, while **Meta** bans them in ads but not in other content. **YouTube** does not have a clear policy, and it is not clear that **X**'s policy prohibiting "false content that directly interferes with participation in an electoral process" applies to premature victory claims. It is possible that X will rely solely on community notes to address this issue. The uneven approach to this issue across platforms leaves open the possibility that candidates will be able to claim victory before results are official on some platforms, but not others, leading to widespread confusion and distrust in official results.

Ephemeral content remains a black box

Livestreaming and other ephemeral content such as 24-hour Instagram, Facebook, Snap and TikTok 'stories' has presented itself as an increasingly popular, real time threat vector for hate speech, violent extremism and misinformation. Impermanent content's rise in popularity has led to it beginning to play a more pivotal role in shaping public discourse, especially during election cycles. The unclear mechanisms to enforce safety policies in ephemeral content, highlights vulnerabilities in real-time content moderation systems, risking the

unchecked spread of harmful content through stories and livestreams. Additionally, when platforms publish public-facing announcements regarding safeguarding elections across the world, livestreaming and stories are rarely mentioned, despite being one of the most at-risk product vectors given the challenges of moderating them in real-time.

Meta encompasses livestreaming within its broader Community Standards against misinformation and voter suppression. In 2020, Meta faced criticism for how it managed livestreams spreading misinformation about election fraud. Notably, live broadcasts related to the "Stop the Steal" movement often remained unaddressed until they had garnered significant viewership. More recently, Human Rights Watch documented how Meta "censored content" and "shadow banned" Instagram and Facebook accounts due to "spam," suggesting erroneous and automated application of the policy. This highlights how enforcement of policies for impermanent content cannot be entirely left up to automation.

Similar to Meta, the other platfroms assessed also require all content across livestream and other temporary content vectors to adhere to the respective platform's safety policies. However, the opaqueness surrounding enforcement of policies for livestreaming and other impermanent content moderation presents a clear risk. The lessons learned from 2020 and 2022 necessitate a forward-looking approach that emphasizes transparency, specificity, and adaptability in policies governing livestreaming and impermanent content, ensuring that platforms are well equipped to mitigate safety risks.

Preventing violence and hate does not seem to be a key focus area of the platforms in the context of the election

In addressing the pervasive issues of hate speech, extremism and the promotion of positive civic engagement on social media, platforms must enact comprehensive and effective strategies. The 2020 US elections highlighted the critical need for such measures. Yet, in platforms' 2024 election blog posts there was little to no mention of strategies or measures taken to combat hate and extremism. In recent years hate crimes have increased in the US and spikes in hate crimes have been detected during presidential campaign cycles. As hateful and extremist speech and actions rise, the platforms must increase their resources to fight them (particularly when they have had a part to play in this rise in the first place).

While all the platforms have seemingly robust and clear policies prohibiting the glorification or threats of violence, we have seen and continue to see significant enforcement failures, most recently in the context of the ongoing Israel-Gaza conflict and global elections in 2024.

As social media platforms continue to evolve, their commitment to combating hate speech, extremism and promoting positive civic engagement will be critical. The experiences from the 2020 election cycle underscore the importance of proactive, comprehensive strategies that address the complexities of moderating online content.

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org