Oversight Board

CONTENT MODERATION IN A HISTORIC ELECTION YEAR: KEY LESSONS FOR INDUSTRY

THE OVERSIGHT BOARD, MAY 2024

IMPROVING HOW META TREATS PEOPLE & COMMUNITIES AROUND THE WORLD

Executive Summary

In this historic year of elections, with the populations of at least 80 countries around the world set to participate, there has never been a more critical time for democracy, human rights, and open and fair societies. In the first three months of 2024 alone, people in Bangladesh, Pakistan, Indonesia and Taiwan went to the polls. Elections throughout the rest of the year are already underway in India and expected across several other countries and regions, including in South Africa, Mexico, the European Union, the United Kingdom and the United States.

Social media platforms play a central role in civic discourse with their impact on democratic processes the subject of extensive debate. While platforms can enable a more transparent electoral process by broadening access to information, they can also be used to incite election-related violence or spread falsehoods to try and manipulate public opinion and influence outcomes. Inaccurate enforcement by platforms can exacerbate these abuses. This is why it is essential the actions of private tech companies, which control the flow of so much political information, are scrutinized.

The Oversight Board, an independent body of 22 human rights and freedom of expression experts from around the world and across the political spectrum, has made the protection of elections and civic space one of our seven strategic priorities. We believe it is crucial that social media platforms defend an open civic space in which people, including members of political oppositions, human rights defenders and marginalized voices, can freely express their opinions, share information and participate in democratic processes. In this year of elections, it is especially important to identify ways in which social media companies can better safeguard the integrity of elections, while respecting freedom of expression. At the Oversight Board, our recommendations have already led to Meta committing to better practices, but more work needs to be done, including by Meta, and other platforms and regulators.

This paper draws on our analysis in relevant cases, which consider emblematic pieces of content on Meta's platforms, to highlight some of the ways social media companies can better protect political speech and counter online challenges to the safe and reliable conduct of elections, under the guidance of international human rights standards. Through the collective insights gained from these cases, we also share our key lessons for industry that are described in full in this paper's final chapter.

Nine Key Lessons for Industry

- Policies are one part of the story, but enforcement is equally as essential. This demands that social media companies dedicate sufficient resources to moderating content before, during and after elections.
- Companies must set basic global platform standards for elections everywhere. They must ensure they do not neglect the dozens of elections taking place in countries or markets considered less lucrative because this is where the human rights impact of not implementing such standards can be most severe. Platforms that fail to deliver should be held accountable.
- Political speech that incites violence cannot go unchecked. Quicker escalation of content to human review and tough sanctions on repeat abusers should be prioritized.
- Platforms must guard against the dangers of allowing governments to use disinformation, or vague or unspecified reasons, to suppress critical speech, particularly in election settings and around matters of public interest.
- Policies that suppress freedom of expression must specify the real-world harms they are trying to prevent, to ensure they are necessary and proportionate to the harm.
- Lies have always been part of election campaigns, but technological advances are making the spread of falsehoods easier, cheaper and more difficult to detect. Clear standards need to be set for AI-generated content or "deepfakes" and other types of manipulated content, such as "cheap fakes."
- Journalists, civil society groups and political opposition must be better protected from online abuse as well as over-enforcement by social media companies, including at the behest of governments and other parties.
- Transparency is more important than ever when it comes to preserving election integrity. Companies must be open about the steps they take to prevent harm and the errors they make.
- Coordinated campaigns aimed at spreading disinformation or inciting violence to undermine democratic processes must be addressed as a priority.

The Importance of Freedom of Expression to Elections

Guided by International Human Rights Standards

The international community expects companies, including social media platforms, to respect human rights. The main human rights standard we apply in our decisions is Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which broadly protects freedom of expression. On elections, the UN Human Rights Committee has affirmed that "free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential," (General Comment No. 34, para. 13).

Protecting Political Speech

Many of our cases emphasize the high protection that political speech receives under human rights law because of its importance to public discourse and debate. In the Altered Video of President Biden case, in which we reviewed a video that had been altered to make it appear as though the US president is inappropriately touching his granddaughter's chest, we emphasized that mere falsehood cannot be the sole basis for restricting freedom of expression under human rights law.

As part of our review, we found that Meta's Manipulated Media policy, which governs how AI-generated content is moderated, was riddled with gaps and inconsistencies, including treating content that portrays people saying something they did not say differently to content showing people doing something they did not do. It also treated types of audio and audiovisual media inconsistently.

While we left up the altered video in this case, despite it showing President Biden doing something he did not do, we urged Meta to revisit its policies on manipulated media to ensure content is removed only when necessary to prevent or mitigate specific harms. These harms needed to be better defined. We also recommended the company eliminate the distinctions and inconsistencies outlined above. Finally, we encouraged social media companies to rely on labeling of AI-generated content as an alternative to removal, except when the content violates other policies. Meta has announced that it is acting to implement our advice, which will provide people with the context they need to make informed decisions about content.

In other cases, we have also decided to protect political speech, even at times it could be considered forceful in nature, seeing this as a necessary part of public debate, provided there is no direct connection to potential offline harm.

For example, we instructed Meta not to take down news reporting of a politician's speech in Pakistan's Parliament, which contained a classical reference that while violent in nature was not literal or likely to lead to actual harm (Reporting on Pakistani Parliament Speech). On another occasion, the majority of the Board determined that a controversial expression of opinion on immigration was not hate speech because it did not contain a direct attack on a group based on a protected characteristic (Politician's Comments on Demographic Changes). In both cases, we decided that the content, while potentially offensive to some, constituted protected political speech, and should stay up.

Our guidance to Meta has been clear: if it is going to take down content relating to politics, especially in the context of elections, that removal must be necessary to prevent or mitigate genuine offline harms. The right of voters to hear the views of politicians is essential.

The Challenges to Safeguarding Elections Online

Challenges to electoral integrity and the safeguarding of democracies must be taken extremely seriously. These range from incitement to violence by political leaders to disinformation that can undermine faith in electoral processes. Design choices made by platforms have often made the issues of disinformation worse, or incentivized these actions, with the rise of generative AI threatening to exacerbate this further. This is why a firm commitment to transparency is key. Users and other stakeholders must see the actions that companies are taking to counter these problems and whether they are learning from past mistakes.

Violence and Intimidation by Political Leaders

Although freedom of expression is generally protected under international human rights law, it may be limited under certain circumstances. Elections and political transitions can often be marked by escalating tensions, with social media sometimes used in settings where there is a heightened threat of violence. In 2021, we addressed the issue of post-election violence by looking at whether Meta was right to suspend former U.S. President Donald Trump from its platforms in the wake of the January 6 U.S. Capitol riots (Former President Trump's Suspension).

We have also looked at leaders inciting violence in other election settings, for example, in the Brazilian General's Speech case. In both these decisions, we found that Meta should have acted more quickly against the encouragement or legitimization of violence. We recommended that during a period of a heightened risk of violence, such messages should not be protected under the guise of the right to protest. We additionally said in the Brazil decision that content removal of individual posts is relatively ineffective when they are part of a coordinated action to disrupt democratic processes. Platforms need to be better at preparing for and responding to such crises.

Election integrity efforts (ensuring fairness of the voting process) and crisis protocols, which set out best practices for platforms to follow during extreme events, are essential solutions. We recommended in both the decisions above that Meta establish a framework for responding to high-risk events. Meta responded by creating a Crisis Policy Protocol, a policy guiding its response to crises when its regular processes are not sufficient to prevent harms. This tool can be applied to electoral controversies such as procedural disputes and contested outcomes, which are often fast-moving crisis situations. In the Brazil decision, we also recommended Meta establish a framework for evaluating and publicly reporting on its election integrity efforts worldwide, including adopting metrics for success, providing relevant data for the company to improve its overall content moderation system, including for both organic and paid content. Meta has committed to do this in the latter part of 2024. Given the urgency of the issue we are closely monitoring this commitment. Information drawn from these metrics should help Meta decide how to deploy its resources during elections and draw on local knowledge to address coordinated campaigns aimed at disrupting democratic processes, as well as set up feedback channels and determine effective measures when political violence persists after an election's formal conclusion.

Of course, incitement to violence is not always confined to the immediate run-up or aftermath of elections. In the Cambodian Prime Minister case, we required Meta to remove a violating post from the then Prime Minister Hun Sen targeting the political opposition with violence months ahead of scheduled elections. Despite the public interest in Cambodians hearing from their prime minister through social media ahead of elections, the virulence of the crackdown against opposition forces led the Board to call on Meta to suspend Hun Sen's Facebook page and Instagram account for six months. Given the volatile situation in Cambodia, we concluded his threats to the political opposition could not be justified as "newsworthy" content and had a high likelihood of causing physical harm.

Unfortunately, Meta's ultimate decision not to suspend Hun Sen's account sets a potentially dangerous precedent for rulers elsewhere who frequently use Meta's platforms to threaten and intimidate critics. A number of international human rights groups spoke out urging Meta to follow the Board's recommendations.

Given the stakes, we are closely monitoring how Meta implements the other recommendations we made in these cases and will continue to hold the company to account for delivering on its pledges.

In the Former President Trump's Suspension case, we had urged Meta to clearly explain its strikes and penalties process for restricting accounts with severe violations of content policies, something the company has since become much more transparent about. In the Cambodian Prime Minister case, we additionally recommended Meta to revisit its policy on restricting accounts of political leaders not only during crises but also in situations where the state is pre-emptively suppressing political expression with violence or threats of violence. We also said that Meta should update its review prioritization systems so that potentially violating content from heads of state is consistently and quickly reviewed by experts – and removed when it poses a risk of likely imminent harm.

We have recently explored another boundary of the limits of political speech, this time in the Greek 2023 Elections Campaign cases. Our majority decision supported the removal of two posts for violating Meta's Dangerous Organizations and Individuals policy. To safeguard election integrity, we recognized it was appropriate for Meta to limit the freedom of political candidates and parties campaigning on its platforms when they specifically reference endorsements by and symbols of proscribed individuals and groups who are known to be connected to violence. However, we also noted that Meta's rules could be clearer, given that the company does not disclose the list of entities it designates as dangerous. Since this decision, the Supreme Court of Greece has prohibited this political party (the Spartans) from participation in the upcoming EU elections, as it had "offered their party as a cloak" to the former spokesman of the banned Golden Dawn.

Finally, we also recognize there are instances when government officials and politicians, including election officials, are harassed or attacked. To partly address this, our policy advisory opinion on Sharing Private Residential Information advised Meta not to allow the sharing of such information when protests are organized in the vicinity of a high-ranking government official's private residence, and security measures may not be in place to guard the safety of those inside.

Risks of Over-Enforcement

The period before and during elections, as well as subsequent inaugurations and transfers of power, often involve heightened communications and information-sharing among users. These are the times that count: when Meta and other companies have a heightened responsibility to get enforcement of their content policies right. Several of our cases have pushed Meta to improve its record on enforcement errors.

One particular problem, especially common during elections, is that governments pressurize platforms to remove lawful content on the (sometimes pretextual) basis that it violates a platform's policies. As a minimum requirement, we have insisted that Meta inform users when their content is removed due to a government request, which the company now does. In the UK Drill Music case, in which we found Meta had removed lawful music that did not in fact violate the platforms' policies on the request of a police force, we recommended that the company adopt a globally consistent approach to receiving content removal requests from the state, make data on these requests public and assess for systemic biases in content moderation decisions resulting from government requests. This recommendation to increase transparency around government takedown requests and make them public was also highlighted in the Öcalan Isolation's case and our policy advisory opinion on Removal of COVID-19 Misinformation.

Another recurring issue we've seen in users' appeals is the difficulties the company faces in distinguishing between figurative political criticism and credible threats prohibited by the Violence and Incitement policy. The Iran Protest Slogan case illustrated our serious concerns that this type of over-enforcement could severely hamper protest movements aiming to promote human rights. Here, we decided that a particular slogan ("marg bar Khamenei" translated as "death to Khamenei," Iran's Supreme Leader) that was being used in the country's ongoing protests should be allowed. Prior to our selection of this case, substantial amounts of content with this slogan had been removed for supposedly inciting violence. An overly literal application of the policy was preventing protesters from expressing their discontent with the regime on Meta's platforms. In our decision, we highlighted that rhetorical political statements, which are not a credible threat, do not violate the policy and do not require a newsworthiness policy exception to be applied. We also recommended changes to the Violence and Incitement Community Standard to protect obviously rhetorical political speech during protests, all with the aim of enabling people to freely voice criticism of their governments.

A policy that often leads to over-enforcement is Meta's Dangerous Organizations and Individuals policy, which prohibits glorification, support and representation of individuals, groups and events the company designates as dangerous. While the policy pursues a legitimate aim, it has in practice all too often led to the arbitrary removal of content posted by users reporting on situations involving those groups, defending human rights or drawing unobjectionable analogies.

In a recent policy advisory opinion, in which we dive deeper into hard policy questions that Meta is considering, we advised the company to end its presumption that the word "shaheed" (loosely translates as "martyr" in one meaning) always denotes praise when referring to designated individuals (Referring to Designated Dangerous Individuals as "Shaheed").

This should ensure more accurate enforcement of what Meta has described as its "most moderated word," ensuring political expression is better respected. We also pushed for additional clarifications and, importantly, asked Meta to clearly explain to users how Meta's automated system is used to generate predictions about potential content violations of this policy.

All the recommendations above have emphasized the human rights responsibilities of social media platforms to address adverse impacts on individuals and society, rather than be guided by political or business interests.

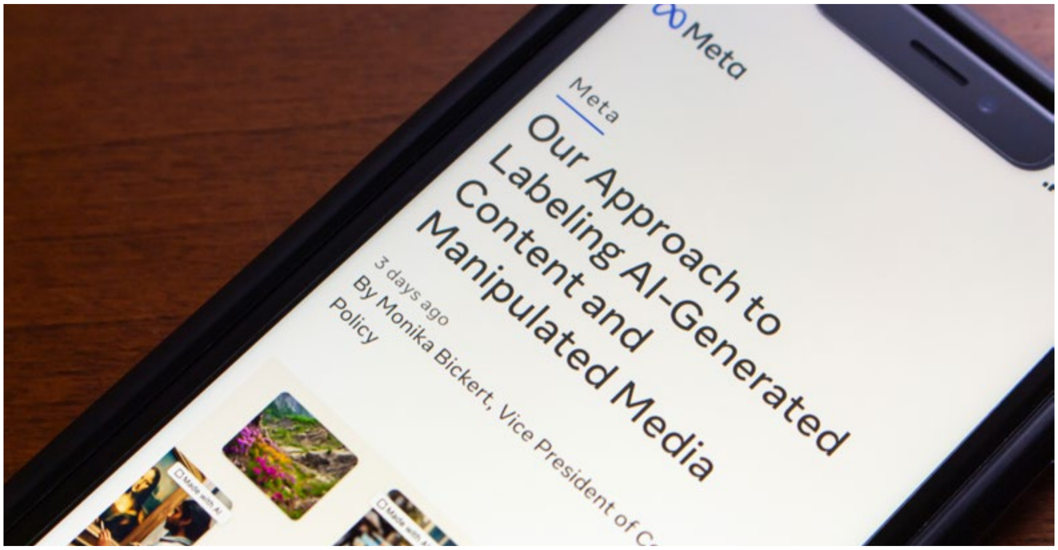
Disinformation

Disinformation can take various forms, presenting distinct harms in relation to elections and the safeguarding of democratic space, fueling polarization and undermining confidence in the integrity of a democratic process. Misleading content can also create distrust in government institutions, civil society and the media. On the other hand, the question of what information is true or false (or misleading) is often a legitimate part of democratic disagreement. Governments and powerful actors sometimes use the presence of misinformation as a pretext for suppressing uncomfortable truths. This is why the attempt to combat harmful misinformation is complex, with these issues particularly pertinent during elections.

Amplification and Coordination of Disinformation

Various actors – including governments – use social media to undermine democratic processes, with tactics for spreading disinformation evolving. Although Meta has identified and removed inauthentic accounts attempting to interfere with elections and the company partners with fact-checkers to label some forms of false or misleading information, coordinated disinformation campaigns continue to proliferate.

Despite efforts to address inauthentic behavior, Meta's design and policy choices, in particular its newsfeed and recommendation algorithms, have enabled disinformation narratives promoted by networks of influencers to gain traction and spread, sometimes leading to offline violence. This prompted us in part to recommend that Meta undertake a comprehensive review of how these choices contributed to the electoral fraud narrative and tensions that culminated in the U.S. Capitol riots of January 2021 (Former President Trump's Suspension). We have also advised Meta to explore measures to reduce organic and algorithmically driven amplified harmful content (Claimed COVID Cure case and policy advisory opinion on Removal of COVID-19 Misinformation). In the same policy advisory opinion, we urged Meta to conduct research analyzing accounts amplifying or coordinating health misinformation campaigns.


Meta
Meta
Our Approach to Labeling AI-Generated Content and Manipulated Media
3 days ago
By Monika Bickert, Vice President of C
Policy
Made with AI
Made with AI

Similar considerations should apply to election-related harmful disinformation, and we have raised the importance of users having the means to appeal Meta's decision when the company demotes their content based on a fact-checker's rating of "false," "misleading" or "altered" (Altered Video of President Biden).

Manipulated Media

Users can create malicious manipulated media that undermines democratic processes and worsens political conflict. While generative artificial intelligence (AI) threatens to make this worse, cruder methods like "cheap fakes" are more common and can be just as harmful. As outlined above, in the Altered Video of President Biden case, we urged Meta to amend its Manipulated Media policy to address various enforcement blind spots. Meta has since announced it will implement the Board's recommendations in full.

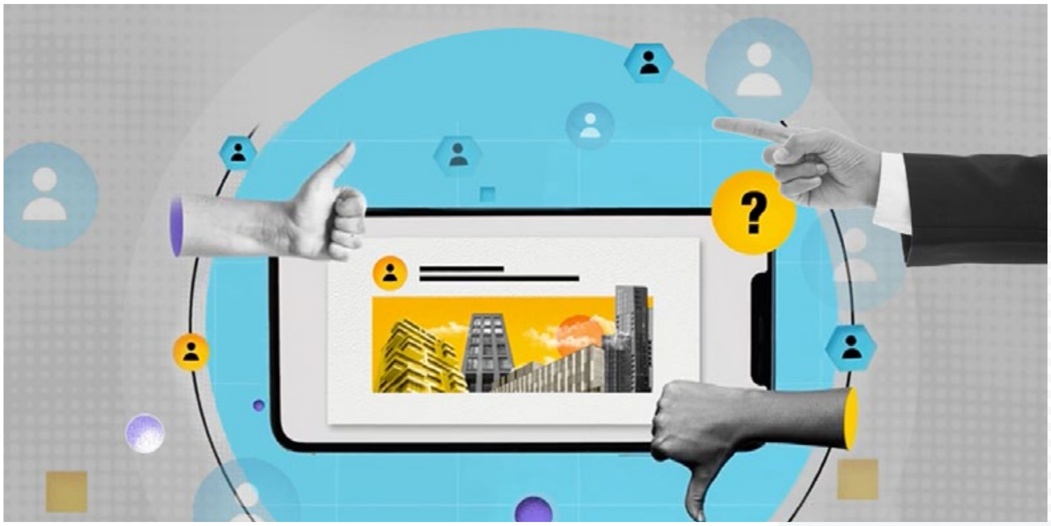
Political Advertising and Influencer Marketing

Paid advertising on social media platforms often obscures the true source of this content, which can be exploited as part of disinformation campaigns. One of the ways Meta currently addresses questions of attribution is to require adverts to show a "Paid for by" label. At a minimum, political adverts must comply with the same set of Community Standards that applies to all content on Meta's platforms as well as its ads policy. Nonetheless, stakeholders have raised concerns about the inaccurate enforcement of Meta's Community Standards when it comes to political ads. We have previously noted reports of political ads using manifestly untrue statements to attack the legitimacy of the Brazil elections on Meta's platforms (Brazilian General's Speech).

According to a civil society group, there were similar developments in the Myanmar and Kenya elections. Our recommendation, to create a framework with success metrics for evaluating the effectiveness of the company's election integrity efforts, was partly a response to this phenomenon. This would help Meta strike a balance between retaining legitimate criticism of elections as part of a healthy public debate and removing content that constitutes real attempts to undermine a voting process.

Political messaging, including that shared by micro-influencers and nano-influencers, can also seed narratives created to influence public opinion about political candidates or the polls. This development further challenges our concept of what is authentic content and what is not. Meta's definition of ads about social issues, election or politics provides some detail but the term "social issues" is itself broad. It encompasses "sensitive topics that are heavily debated, may influence the outcome of an election or result in/relate to existing or proposed legislation."

Partly in response to our recommendation (Altered Video of President Biden) for Meta to better inform users of the origins of manipulated media, the company also now requires advertisers to disclose their use of AI to create or alter a political or social issue advert that was "digitally created or altered to depict a real person saying or doing something they did not say or do."



Conclusion: Key Lessons for Industry

From the high protection that political speech warrants to the measures that platforms can take to better counter the spread of falsehoods online, we remain committed to the protection of elections through our decisions on emblematic cases and the issuing of recommendations that pursue best practices in content moderation. Drawing on the collective insights gained from this ongoing elections-related work, and other related cases, we have identified the following key lessons for those working to preserve electoral integrity on social media platforms. These are guidelines primarily for industry, but we hope they will help influence other stakeholders as they work to hold companies to account.

- **Policies are one part of the story, but enforcement is equally as essential during rapidly escalating situations.** This demands that social media companies dedicate sufficient resources to moderating content before, during and after elections, doing so on a global scale irrespective of whether they have political or economic interests in the affected country. Disputed elections can all too easily lead to crisis and conflict. It is imperative that social media companies have sufficient expertise in the local language and context to guide their global elections policies and practices there.

- **Companies must set basic global platform standards everywhere and platforms be held to account for failing to deliver.** It is important that companies do not neglect the dozens of elections taking place in countries or markets considered less lucrative because this is where the human rights impact of not implementing such standards can be most severe. While resources are finite, the harms of unchecked disinformation or incitement to violence are just as acute in often-overlooked regions, with instability in one location fueling instability in another and emboldening bad-faith actors elsewhere.

- **Political speech that incites violence cannot go unchecked. Quicker escalation of content to human review and tough sanctions on repeat abusers should be prioritized.** This is especially important when considering content from heads of state and senior members of government that could potentially incite violence. While people have a right to see what is "newsworthy," harmful content that outweighs the public interest, and which fundamentally undermines the election process, must be expedited for human review and removed when necessary. If politicians repeatedly break the rules, they may need to be suspended from online platforms. This is most critical at election time when the ability to amplify harms, threats and intimidation is greatest.

- **Platforms must guard against the dangers of allowing governments to use disinformation, or vague or unspecified reasons, to suppress critical speech.** This is particularly relevant in election settings and around matters of public interest.

- **Policies that suppress freedom of expression must specify the real-world harms they are trying to prevent, to ensure they are necessary and proportionate to the harm.** When dealing with misinformation and disinformation, there are clear tensions between allowing freedom of expression and access to information – essential to democratic processes – versus protecting people from real-world harms, especially violence. Policies should reflect risks to physical safety and security, as well as the risks of intimidation, exclusion and silencing. Speech that does not pose harm must not be suppressed under the guise of misinformation.

- **Lies have always been part of election campaigns, but technological advances are making the spread of falsehoods easier, cheaper and more difficult to detect. Clear standards need to be set for AI-generated content or "deepfakes" and other types of manipulated content such as "cheap fakes."** If something is harmful, it should be treated as such. The fast-moving pace of changes in technology means that policies can become outdated, which creates gaps and allows abuses to proliferate. It is vital that when designing policies and processes, companies are clear about the end goal or ultimate harm they are trying to prevent, with global stakeholders consulted as part of this process.

- **Journalists, civil society groups and political opposition must be better protected from online abuse as well as over-enforcement by social media companies.** Ensuring that political opponents and civic actors can express themselves is essential to a fair electoral process and an area that social media companies must prioritize, especially in countries where freedom of expression is routinely suppressed. Protestors and others who peacefully criticize their governments should be protected. Government requests to remove content must be scrutinized with human rights considerations in mind. Striking the right balance during a crisis is not easy but having consistent policies to deal with these situations, while providing additional support and training to moderators, are important first steps.

- **Transparency is more important than ever when it comes to preserving election integrity.** People need to clearly see where disinformation and other harmful content is coming from, what form it is taking and what impact it is having. Companies must be open about the steps they take to prevent harm, the errors they make and clearly set standards on how they can improve. This includes committing to greater clarity about state-backed takedown requests, which have the power to unduly silence opponents.
- **Coordinated campaigns aimed at spreading disinformation or inciting violence to undermine democratic processes must be addressed as a priority.** Such campaigns undermine trust in democratic processes by making it more difficult for people to find accurate information, enabling harassment of those who express political dissent and spreading falsehoods as though they are accepted facts. Social media companies should improve their design and policy choices to ensure that disinformation narratives are not amplified.

Acknowledgements

This paper was developed by Oversight Board members Michael McConnell, Pamela San Martin and Afia Asantewaa Asare-Kyei, and reviewed by all Board Members. This collaborative effort benefitted from contributions by: Jenny Domino, Senior Case and Policy Officer, and Carly Miller, Data and Implementation Officer, with editing and approvals from: Simona Sikimic, Communications and Engagement Manager, Neena Dhillon, Content and Editorial Lead, and Andrew Smith, Head of Case and Policy.

Oversight Board