Free Press

DEMOCRACY DEFERRED

Social-Media Companies' Meager Commitments to Election Integrity in 2024

A Report by Free Press
Written by Nora Benavidez & Jessica J. González

April 25, 2024

Table of Contents

CONTENTS

BACKGROUND

On April 9, 2024, more than 200 civil-society organizations, researchers and journalists from across the globe sent a letter to 12 of the top technology and social-media companies, calling on them to improve election-integrity efforts in 2024. The letter was delivered to the CEOs of the following companies:

- Discord
- Google
- Instagram
- Meta
- Pinterest
- Reddit
- Rumble
- Snapchat
- TikTok
- Twitch
- X (formerly Twitter)
- YouTube

The letter included an urgent request that social-media and tech platforms "increase platform-integrity efforts to protect democracy in 2024, as at least 60 countries have national elections this year and there is evidence of continued democratic backsliding and rising authoritarianism across the globe." Specifically, the letter requested that platforms adopt six core policies and practices and respond by April 22, 2024 with their intentions to adopt or reject these initiatives.

THE SIX DEMANDS ARE AS FOLLOWS:

01 **Reinstate election-integrity policies. Continue to moderate election misinformation about the 2020 U.S. election results (specifically Big Lie content), and staff up platform-integrity teams to enforce policies across languages.**

02 **Do not allow hate and lies in political ads, require human review of all ad buys and label ads that contain generative AI.**

03 **Require disclosure when political content is AI-generated, prohibit deepfakes in political ads and promote factual election content.**

04 **Reduce distribution of election content that has been flagged and is awaiting review.**

05 **Hold VIP accounts to the same standards as all other accounts.**

06 **Improve transparency.**

OVERVIEW OF COMPANY RESPONSES

Eight of the 12 companies responded on or around the April 22, 2024 deadline. These included Google and YouTube (through a Google representative), Meta and Instagram (through a Meta representative), Pinterest (not linked here because Pinterest marked the letter privileged and confidential), Reddit, Snap and TikTok. Six of the eight responses, excepting the ones from Google and Snap, hovered at barely two pages of text, hardly enough to substantively answer the questions presented.

X confirmed receipt of the letter but has not responded substantively. Discord, Rumble and Twitch failed to respond altogether. This shows utter disrespect to the more than 200 civil-society organizations, researchers and journalists that signed the letter, as well as a shocking disregard for the precarious state of global democracy this year.

GLOBAL CALL FOR PLATFORM INTEGRITY 2024	
RESPONDED	G Meta P 👽 👻 TikTok ▶
CONFIRMED RECEIPT	X
DID NOT RESPOND	rumble

Letter responses as of April 22, 2024

MAIN TAKEAWAYS

Several troubling themes emerged in the letters from the eight responsive companies.

01

First and foremost, no company responded with a direct yes-or-no answer on whether it would adopt or reject the six demands. To the extent that companies committed to elements related to any of the six demands, they did not provide timelines for compliance as requested.

02

None of the companies explicitly committed to taking down Big Lie content. This is particularly troubling given mounting evidence that Big Lie content erodes trust in democratic institutions and likely discourages participation in elections.

03

None of the platforms committed to staffing up trust and safety teams to allow for more human content moderation, despite reports that tens of thousands of employees — including significant portions of trust and safety teams at Meta, Twitch, X and YouTube — have lost their jobs over the past 18 months.

04

Four of the eight responsive platforms have signed a voluntary AI elections accord, touting their participation as evidence of a commitment to safeguard against AI harms. This is a positive step, but it is merely a promise — not a fulfillment of such a commitment. Free Press analyzed the accord and noted:

> *Whether these tech giants actually do the work of better detecting, labeling and debunking AI-generated disinformation remains to be seen. Lofty principles and promises like those expressed in this accord are nice, but the rubber meets the road when it comes to implementation. There are deeply complex questions about how to moderate AI-generated content … The companies must do more than what is outlined in the accord.*

There were a few pleasant surprises as well:

01 **All of the responsive platforms, except for TikTok, promised to hold VIP accounts to the same content-moderation standards as regular user accounts, though experts at a recent press briefing expressed doubts about platforms' follow-through on these commitments.**

02 **Most of the companies stated that they would moderate content across various non-English languages, though these commitments varied in depth and breadth and companies' recent track record on this point is cause for skepticism.**

03 **Most of the platforms that responded have plans to flag generative AI that aims to confuse or misinform people, though again commitment levels vary.**

The following provides further detail about the extent to which each of the 12 platforms were responsive to the letter's demands. This is not an evaluation of whether platforms are actually enforcing policies to the extent they promise to do so.

Such research is important and ongoing, and also difficult to do given widespread reluctance from social-media and tech firms to allow open and affordable access to their systems. Over the years we have witnessed many empty promises from tech platforms, so tech-company words should be taken with a grain of salt. Civil society can, must and will continue to study whether the platforms are living up to their commitments.

PLATFORM ANALYSIS

GLOBAL CALL FOR PLATFORM INTEGRITY 2024	
ADEQUATE	
PARTIAL	TikTok
INSUFFICIENT	G Meta
FAIL	X rumble

Discord — Fail
Already infamous for providing a home for violent extremism, Discord failed to respond to the letter. In early April, Discord announced that ads would be coming to the platform — advertisers should proceed with caution when deciding whether to advertise on Discord. Parents beware: Discord may not be suitable for children.

Facebook and Instagram — Insufficient
Meta summarized the activities for Facebook and Instagram, claiming its letter is an official response regarding all Meta platform activities for 2024. On its face, the response points to extensive work Meta is allegedly doing to safeguard elections and protect users. But it only scratches the surface.

As with every election cycle, Meta touts connecting users with credible information about elections and says it is staffing the “largest global fact-checking network of any platform,” with “nearly 100 independent fact-checking organizations around the world who review and rate content in more than sixty languages.”

For years, Free Press and allies have pushed platforms to invest significant resources in content moderation across non-English languages. After Free Press launched the #YaBastaFacebook campaign and whistleblower Frances Haugen revealed failures in Meta's enforcement of non-English content, the company finally committed to fighting misinformation across all languages. To see Meta reiterate its commitment to this work is a positive step, and we will be tracking the platform to ensure its actions meet its words.

Meta promises to prohibit ads that dissuade people from voting, but the company will not take action against Big Lie mis- and disinformation, committing to solely acting against lies about upcoming elections, not previous ones. Meta did not commit to increasing its staffing for critical teams, such as those focused on trust and safety or other content-moderation roles. This is especially notable given Meta's mass layoffs over the past 18 months — hitting trust and safety and other integrity teams — and its shuttering of a fact-checking program that had taken the company a half-year to build.

Meta advertisers who run ads about social issues, elections or politics are required to complete an authorization process and include a "paid for by" disclaimer disclosing the money behind the message. Meta maintains an exhaustive, publicly available Ad Library, launched in 2018, with ads saved for seven years. Experts and researchers value the tool since it was the first of its kind. But it's largely an opaque and clunky database. Users experience difficulties trying to understand the breadth of trends in political ads or to parse through other metrics to determine why they or others encounter specific content.

As in previous years, Meta will continue to block new political, electoral and social-issue ads during the final week of the U.S. election campaigns. This intervention has mixed results and in previous years has left small and local nonprofits unable to amplify get-out-the-vote messaging on Facebook and Instagram. Meanwhile, critical election-related disinformation is often seeded and spread to millions of users long before the final week leading up to a vote.

Meta will not prohibit AI-generated ads that use deepfake technology to falsely portray politicians or others. However, in certain cases it will begin requiring advertisers to disclose when they use AI or other digital methods to create or alter ads about social issues, elections or politics. Meta will begin labeling a wider range of video, audio and imagery as “Made with AI” when the company detects industry-standard AI image indicators or when people self-disclose that they’re uploading AI-generated content. Details remain sparse on the methodology and the process of review by automated tools and/or humans.

Meta has committed to treating all users equally, without giving special treatment to VIP users or candidates, marking a departure from its prior practice. In previous election cycles, Free Press and other civil-society organizations met with Meta executives, including Mark Zuckerberg, and revealed examples of violative content by VIP users — up to and including threats of violence. In the past, Zuckerberg and other Facebook officials declined to remove such posts, citing “public-interest” concerns in allowing users to see such content.

There is minimal transparency about the company’s practices. Meta makes it nearly impossible to understand how it deals with an array of problematic content, whether through labeling and other friction, downranking or removal, and beyond. As Free Press has documented in previous research on Meta’s pledges, attempting to gain clarity is like trying to find one’s way through a forest of ever-changing policy updates, contradictory community standards, newsroom announcements, blog posts, Terms of Service, business centers, advertising centers, help or customer-support centers, and more. There’s an excess of redundant internal linking in Meta’s most recent response, creating a loop back to sources that sound substantive but ultimately provide little insight into concrete actions and numbers when it comes to content moderation and enforcement processes.

Although Meta promises API access for researchers and others, this access has significant limits. All of the major platforms, including Facebook and Instagram, require advanced notice from researchers, who must be affiliated with universities to get access to companies' API. To access platform data, researchers must also first note what they intend to publish from this information. This presents significant barriers to API access.

Google and YouTube — Insufficient
Google issued its response shortly after midnight Eastern Time on April 23, 2024. It responded on behalf of Google Search and YouTube, both of whom share parent company Alphabet.

Google has policies against manipulated media, hate, harassment, incitement to violence and demonstrably false claims that undermine democratic processes. The company claims to enforce these policies in "an array of linguistic capabilities," including English, Spanish, E.U. member-state languages and all the major Indian languages. Google does not indicate whether it moderates content in African and Asian languages outside of those that overlap with the ones listed above. Like the other platforms, Google does not have a policy against Big Lie content and it does not commit to staffing up for better enforcement in 2024. It does commit to lifting up factual election information in search.

Particularly disappointing is Google's lackluster approach to moderating political advertisements. Essentially Google's policy is to comply with local law. Given that many local laws are failing to keep apace with technological advancements, this bar is far too low. Some — but far from all — local jurisdictions require verification processes for political advertising. In those contexts, advertisers must "prominently disclose when their ads contain synthetic content that inauthentically depicts real or realistic-looking people or events. This disclosure must be clear and conspicuous, and must be placed in a location where it is likely to be noticed by users. This policy applies to image, video, and audio content."

One highlight: Google offers very concrete guidance — including a couple of examples — about what clear and conspicuous disclosure should look like. That said, this policy does not cover a large swath of countries, including many countries where democracy is under threat or nonexistent.

Notably, it is unclear whether Google's policies against hate, harassment and demonstrably false claims that undermine democratic processes apply to political-ad content. The political-ad explainer page that Google's letter directed us to does not suggest these policies apply to political-ad content. Google fails to commit to human review of political ads or enhanced labeling and scrutiny of ads that contain generative AI in localities that don't require verification processes. This falls below what appears to be the industry standard as described in other respondents' letters, and is particularly disappointing given Google's immense reach as the go-to search engine and its deep pockets.

Google's letter sheds very little light on how it aims to improve transparency and strengthen researcher access to data.

YouTube's election-integrity policies prohibit content that aims to "mislead voters about the time, place, means and eligibility to vote, including false claims that could materially discourage voting, including those disputing the validity of vote by mail or encouraging others to interfere with democratic processes." Despite this policy, YouTube declined to state that Big Lie content is prohibited.

YouTube says it removes violative content but that it relies on an automated flagging system and user flags to moderate content. It does not commit to staffing up to better enforce election-integrity policies, instead relying on the free labor of its users, along with its automated systems, to do the heavy lifting.

On YouTube, users who violate the rules get three strikes before they are removed from the platform, but the strike count resets every 90 days. Finally, our critique of Google's efforts to moderate content across languages (see above) applies to YouTube as well.

YouTube's advertising policies mirror Google's disappointing approach. One minor improvement over Google's policy is that YouTube prohibits content that is technically manipulated or doctored to mislead users if it "may pose a serious risk of egregious harm." YouTube fails to provide examples of what may qualify to meet this standard. However, YouTube's page explaining how it aims to bolster disclosure of other AI-generated content is fairly helpful.

YouTube is silent on whether it reduces the visibility of election content flagged for violating rules and awaiting review. Nor does it share any commitment to expand transparency and researcher access to APIs.

YouTube does commit to applying its policies equally to all users, including VIP accounts. This is a dramatic departure from its past practices and we'll be monitoring to ensure it lives up to this commitment.

Pinterest — Insufficient

Pinterest issued a response marked "privileged and confidential" late in the evening on April 22, 2024. To honor this we are not linking to it here.

The company claims that the platform is actively engineering a different kind of experience for users, focused less on virality or politics and more on giving users "ideas to create a life they love." The company commits to a year-round policy of acting against civic misinformation, including lies about elections. But it does not reference enforcement related to Big Lie mis- and disinformation. Pinterest plans to promote nonpartisan election information in search results ahead of certain elections, giving users a link to vote.org as a destination for further information.

As elections near, Pinterest will limit search recommendations for election-related content in places like the home feed, related Pins, notifications or "more ideas" within a board.

The company has not allowed political-campaign ads since 2018 and will not monetize content served in response to election-related searches during election seasons in countries where Pinterest shows a search advisory, as discussed above. That means for the months surrounding the U.S. elections (and certain other global elections), Pinterest won't show any ads when a user searches for common election-related search terms like candidate names, "vote" and "election campaigns."

Pinterest does not commit to bolstering staffing for critical teams, such as trust and safety or other content-moderation roles. It also does not reference in-language moderation or staffing across non-English content. It does commit to treating all users equally, without giving special treatment to VIP users or candidates.

The company provides little in the way of transparency for the public or researchers, beyond mentioning its mandated compliance reporting for the Digital Services Act in the E.U. It also mentioned some statistics that leave more questions than answers. "In Q2 2023, 93% of Pins deactivated for violation of [Pinterest's] civic misinformation policy were seen by fewer than 10 users before they were deactivated." These details raise questions about the visibility and virality of the additional 7% of deactivated Pins.

Reddit — Partial

Reddit provides few details in response to the six demands from our letter. The company claims that the "democratic, community-based structure of our platform [...] differs significantly from the governance structures of the other platforms to which you've directed your letter."

However, the company does not commit to enforcing or removing claims legitimizing the Big Lie or other election misinformation. Reddit makes no claims about increasing staff on platform-integrity teams.

Reddit prohibits political ads outside of the United States. Within the United States, it restricts political ads at the federal level, and “only once the campaign has first done a live AMA with the Reddit community.” Reddit prohibits political ads featuring deepfakes and requires ads containing synthetic content to include a clear disclosure. Humans review ads on Reddit for compliance. Over the course of the election cycle this year, Reddit promises to share authoritative civic information about voting processes through on-platform means like curated AMA series with election authorities.

The company claims that all rules apply equally to all platform users, but fails to describe moderation or other enforcement across languages.

With respect to calls for enhanced transparency, Reddit provides noncommercial academic researchers access to Reddit data free of charge through its API.

Rumble — Fail
Rumble, a notorious haven for lies and extremism, failed to respond to the letter.

Snap — Partial
Snap responded on April 19, 2024. It was the first company to reply and the most responsive to our demands. For instance, Snap states that it promotes factual election content and expressly prohibits false information, threats and calls for violence,. But the company did not explicitly reference policies related to Big Lie misinformation.

Unlike many of its competitors, Snap claims to remove — as opposed to label or downrank — content that violates its policies. It also claims to have language capabilities commensurate with all countries in which it operates. But it fails to go into detail about how many languages and how many in-language moderators it employs. Snap claims to limit AI-generated content that seeks to undermine civic processes and discourage voters and indicates that My AI, Snap's chatbot, has been trained to refrain from issuing opinions on political candidates. Notably, Snap claims that all political ads on its site undergo human-review fact checking and vetting for misleading AI before posting. Snap purports that VIP accounts receive the same treatment as other accounts.

Snap was silent on whether it reduces the distribution of election content that has been flagged and is awaiting review. It also fails to go into detail about transparency efforts and access to data for researchers and journalists. Snap does not promise to label all political advertisements containing AI, but only to vet for "misleading" AI.

TikTok — Partial
We received TikTok's response on the morning of April 23, 2024, though TikTok claimed to have sent it the previous day.

TikTok committed to ensuring that its platform is not used to sow misinformation in ways that reduce the integrity of civic processes and institutions. Notably, however, it did not commit to expelling Big Lie content and it did not specifically promise to staff up trust and safety teams. It did share the amount of money the company plans to spend on trust and safety ($2 billion) and the number of trust and safety staff it is currently employing (40,000). These figures would be more helpful if provided with comparative context about past spends and staffing statistics.

TikTok states that it moderates content in 50 languages but does not provide details about which specific languages it covers. TikTok does not run political ads, so several of the civil-society letter's demands to prevent lies and hate in political ads are moot. TikTok claims to elevate authoritative information about elections, and claims to label "synthetic" content. It claims it does not permit "manipulated content that could be misleading" — presumably this includes deepfakes. The platform purports that it does not recommend unsubstantiated content, and says that users "may" receive labels warning them to think twice about posting unsubstantiated content.

Unlike many of the other respondents, TikTok provides conflicting details about how it promises to treat VIP user accounts. In one sentence it claims to treat VIPs the same as other accounts; however, in the following sentences, TikTok clarifies that "because of the role these public interest accounts play in civic processes and civil society, we enforce different account restrictions in keeping with our commitment to human rights and free expression."

TikTok claims to be expanding access to API, but that process is still underway and the details provided were hazy.

Twitch — Fail

Twitch failed to respond to the letter. This is troubling given previous reports that Twitch provides a breeding ground for white supremacists and other far-right extremists.

X — Fail

X acknowledged receipt of the letter on April 9, 2024, but it never responded other than to confirm receipt. X has joined the underbelly of the internet: The company has gutted its staff and its content-moderation policies since Elon Musk assumed ownership and has reinstated thousands of previously banned accounts. The company just recently upgraded its auto-response to press inquiries from a poop emoji to a "We will get back to you soon" infinite loop. The company's infamous owner is himself responsible for regularly spreading dangerous conspiracy theories and hate speech.

RECOMMENDATIONS

Taken together, the responses from the platforms are wholly insufficient and demonstrate a lack of seriousness across the industry about the precarious state of elections around the globe. The responses also fail to acknowledge the companies' respective roles in destabilizing democracy.

Free Press calls for swift adoption and implementation of the policies and practices outlined in the letter from civil-society groups. We are also calling for public disclosure from the platforms on timelines for implementation of said policies and practices, and a much more robust investment in trust and safety and election integrity around the globe and across languages.

There is much for tech and social-media companies to do: First and foremost, they must reinstate and bolster the efforts they have retreated from over the last 18 months. Despite the goodwill of many inside of these companies, past election cycles have shown that tech and social-media executives will act in the public good only if they are met with external oversight and persistent public pressure. We therefore recommend the following:

:

01 Increased media coverage and scrutiny of social-media company policies and practices, including whether they are living up to their stated promises to the public.

02 App-store providers like Apple and Google should examine whether the most negligent of these companies — namely Discord, Rumble, Twitch and X — are in compliance with the providers' terms of service.

03 Advertisers should call on their social-media advertising partners to increase commitments to election-integrity efforts like the ones outlined in the letter from civil-society groups.

04 Investors and shareholders should ask these companies hard questions about their role in destabilizing democracy, and demand greater accountability efforts inside the company.

05 Trust and safety team members within these companies should continue to advocate for the development, honing and enforcement of policies that protect democracy and public safety.