Evaluating Platform Election-Related Speech Policies

The blog post was updated on Sept. 10, 2020 to reflect Facebook and Pinterest's policy changes on Sept. 3 2020 and Twitter's policy changes on Sept. 11, 2020. We have since made this blog post a PDF to preserve a policy archive and show the evolution that has occurred during this short period of time.

Since publishing this blog post on Aug. 18, 2020 we have seen significant changes to Facebook, Twitter, and Pinterest's election-related policies. These changes include new policies that address a specific type of election-related content that we previously highlighted as important to election integrity: content that delegitimizes election results. Additionally, these three platforms have clarified language around *how* these policies will be applied — either through labeling or removing the content. These updates are steps in the right direction and we hope platforms that lack policies to address election-related content will find these news policies inspiring.

Election-related misinformation and disinformation on social media can have a significant impact on political behavior and on citizens' trust in election outcomes. There is a growing consensus among the social media platforms, e.g., Facebook, Twitter and YouTube, that election-related misinformation and disinformation requires special attention — and, in some cases, action — from content, trust and safety, and policy teams. In recent months several platforms have updated their policies and expanded their definitions of election-related content: in December 2019, Facebook expanded its voter and/or census interference policy; in May 2020, Twitter introduced a Civic Integrity Policy; on Aug. 5, TikTok released updated policies on combating election interference; and Nextdoor and YouTube updated policies on election misinformation and election-related content, respectively, on Aug. 13.

Confronting the ambiguity and complexity around what constitutes misleading election-related content is an ongoing challenge. Even when platforms agree about the general principles at stake, the specifics can be difficult to parse: a post that would be considered "voter suppression" by one platform, for example, might be deemed "misleading information" by another. In this blog post we examine the election-related policies of 14 different platforms and assess the extent to which they address the threats stemming from election-related misinformation and disinformation. We also look at how these policies cover a specific type of content that has already been, and is likely to continue to be, especially problematic in the 2020 election: false claims aimed at delegitimizing election results. Finally, we briefly discuss the challenges of applying these policies in practice. The attached PDF contains a detailed break-down of how each platform's policies fit into our framework and how we arrived at our conclusions.

Key Takeaways:

- Overall, we find that few platforms have comprehensive policies on election-related content as of August 2020. Moreover, the terms platforms use to describe problematic content — such as "misrepresentation" and "misinformation" — are not in clear alignment, with platforms sometimes using different terms to describe similar election-related issues.

- We have defined three core categories of election-related misinformation and disinformation — Procedural Interference, Participation Interference and Fraud. Using this framework, we classify all election-related policies from six major platforms and rate the platform policies as "None," "Non-comprehensive" or "Comprehensive." We found that only a handful of platforms have comprehensive policies.

- A fourth, broader category of election-related content, which aims to delegitimize election results on the basis of false claims, poses significant problems for all of the platforms we analyzed. None of these platforms have clear, transparent policies on this type of content, which is likely to make enforcement difficult and uneven.

- In addition, some platforms have carved out exceptions for "newsworthy" content or posts "that are of public interest." Such exceptions must be based upon detailed guidance and are not a reason for the platforms to avoid using their own free expression rights to provide fact-checks and counter-messaging.

- Enforcing policies that prohibit "misrepresentations" or "misleading" content of various kinds will require platforms to know the facts on the ground. Policies seeking to address falsifiable delegitimizing content will require platforms to make difficult judgment calls about political speech. There is no panacea for election-related misinformation and disinformation. But clear, transparent policies will make it easier to mitigate threats to the election and will improve user confidence in enforcement decisions.

Four Categories of Election-Related Misinformation

The first step to mitigating the impact of election-related misinformation and disinformation is understanding the current policy landscape: what election-related policies are in place at the social media companies? What shared concepts are these policies based on? What potential vulnerabilities remain? While it is clear that the process of crafting policies to confront misleading election-related content is ongoing — platforms are still actively developing these policies as the election approaches — broad areas of agreement have emerged about what kinds of content need to be addressed.

Based on our analysis of the major platforms' policies, we have defined four core categories of election-related content. Three of these, listed below, concern specific harms related to voting procedures and voter suppression:

- **Procedural Interference**: Misleading information about the actual election procedures. Content directly related to dates and components of the voting process that prevents people from engaging in the electoral process. For example:
    - Content that misleads voters about how to correctly sign a mail-in ballot.
    - Content that encourages voters to vote on a different day.
- **Participation Interference**: Related to voter intimidation. Content that deters people from voting or engaging in the electoral process. For example:
    - Content that affects the desire or perceived safety of voters engaging in the electoral process.
    - Misleading information about the length of lines at a polling station, to deter in-person voting.
    - Misleading information about a supposed COVID-19 outbreak in an area, to deter in-person voting.
    - Allegations of a heavy police presence or ICE agents at polling stations.
- **Fraud:** Content that encourages people to misrepresent themselves to affect the electoral process or illegally cast or destroy ballots. For example:
    - Offers to buy or sell votes with cash or gifts.
    - Allegations that pets are mailed ballots to cast a vote.

As we show in the attached document, these categories are adequately covered by the larger platforms' policies, and broad agreement about the lines between permissible and impermissible election-related content makes it easier to classify and address content falling into these categories. However, there is a fourth category of election-related content that poses other, more difficult content-moderation issues and, to complicate matters further, is not generally addressed by the platforms' policies. This is content aiming to delegitimize election results on the basis of false or misleading claims.

Addressing this type of content effectively will require very precise, granular definitions. There is a spectrum of possible claims intended to delegitimize the election — some authentic, and some disingenuous — and deciding which kinds of claims have the potential to cause specific harm is difficult. While broad claims about the legitimacy of the election, such as "it's all rigged" or "the system is broken," are a part of normal political discourse, such claims become highly problematic when they are made on the basis of misrepresentation — such as a misleading video or photo. The former is something Americans might see and hear every day; the latter has the potential to cause political instability if it is combined with coordinated inauthentic behavior or goes viral. In the chart below we lay out a matrix of potential scenarios to understand these claims and their potential to cause harm:

Scenario	**Type**	**Message**
Scenario 1	Generic, Non-falsifiable Claim	"The election is rigged!"
Scenario 2	More Specific, Non-falsifiable Claim	"The election is rigged because I heard a postal worker has been bought off by party X to burn all party Y ballots."
Scenario 3	Specific, Falsifiable Claim with Purported Evidence	"The election is rigged! Here is a video!" [Video that is real, but taken out of context, or manipulated, through artificial means or cut to change the intent/message/outcome of the original video]

Figure 1: The three scenarios represent different statements with varying levels of "evidence" to support each claim. In the section that follows, the scenarios are tested against the platforms' policies to see if they provide clear action on the content.

This should not be taken to mean that there are clear lines here, or that addressing issues stemming from these kinds of claims is as simple as applying a rubric. This area of election-related content, more than others that can be more clearly delineated, will require "judgment calls," even with specific, granular definitions in place. We discuss this problem below in the context of a comparison of platform policies.

Comparing Platform Policies

How do popular platforms address these broad categories of election-related content? That answer ranges from comprehensively to not at all. We compared the community guidelines of 14 platforms: Facebook, Twitter, YouTube, Pinterest, Nextdoor, Parler, Gab, Discord, WhatsApp, Telegram, Snapchat, TikTok, Reddit and Instagram. Our choice to look at these platforms in particular was guided by two criteria: 1) platforms that are the most popular U.S. social media platforms by user base or 2) platforms that market themselves as political forums.

We rated each platform's policies as either "None," "Non-comprehensive" or "Comprehensive," depending on how specifically it addresses the content type (see the attached PDF for a detailed assessment of each platform's election-related policies):

- **None**: The platform has no explicit policy or stance on the issue.
- **Non-comprehensive**: Policy in this category contains indirect language, or uses broad "umbrella" language, such that it is not clear what type of election misinformation and disinformation the policy covers. This is also reserved for policies that give one detailed example such that they cover some, but not all, of a subject.
- **Comprehensive**: Policy in this category uses direct language and is clear on what type of election misinformation and disinformation the policy covers. It also sufficiently covers the full breadth of the category.

Our first comparison of the policies compared platforms across the three core categories detailed above.

Six platforms address these three categories in some manner: Facebook, Twitter, YouTube, Pinterest, Nextdoor and TikTok. The scores for these platforms can be found in the chart below. Seven platforms — Parler, Gab, Discord, WhatsApp, Telegram, Snapchat and Reddit — do not have election-related policies at all. Instagram's policies are ambiguous since it is not clear whether Facebook (which owns Instagram) applies the same policies uniformly across both platforms.

	Procedural Interference	Participation Interference	Fraud
Facebook	Comprehensive	Comprehensive	Comprehensive
Twitter	Comprehensive	Comprehensive	Non-comprehensive
YouTube	Comprehensive	Non-comprehensive	Non-comprehensive
Pinterest	Comprehensive	Comprehensive	Comprehensive
Nextdoor	Comprehensive	None	None
TikTok	Non-comprehensive	None	None

*Figure 2: UPDATE: This chart has been edited on Sept. 11, 2020 to reflect Pinterest's policy changes on Sept. 3, 2020. The changes are shown in red. For Procedural Interference, Pinterest's policy changed from "None" to "Comprehensive" and for Participation Interference and Fraud Pinterest's policy changed from "Non-Comprehensive" to "Comprehensive." The evaluation of the policies by platform was informed by their community guidelines and standards linked here: **Facebook**, **Twitter**, YouTube, Pinterest, Nextdoor and TikTok. For more detail about each platform's policies and justification for each evaluation, see the attached PDF at the end of the blog post.*

These three categories refer to specific types of election interference — for example, instructing users to vote on the wrong day, falsely alleging that there are legal repercussions for voting, and attempting to buy or sell votes. Though comprehensiveness of these policies isn't a guarantee of their effectiveness or of their consistent enforcement, we have nonetheless chosen to focus on this aspect for our initial assessment as a way to help inform our outreach to platforms and escalation channels in our monitoring work in the months to come.

Our second platform comparison analyzed how platforms handle a less clear-cut category of content: delegitimization of election results. Imagine, for example, an Instagram Story on Nov. 4 showing a person shredding ballots and alleging that the election was stolen — only the video was shot in 2010 in Ukraine. There is significant potential for harm if there are a large number of such false claims, or if they go viral before they can be fact-checked. At the same time, there must be a high threshold for "falsifiability" here, because broad claims of illegitimate results are a normal part of political discourse.

Below is our analysis of how platform policies address various scenarios, listed in the matrix in the previous section, related to the delegitimization of election results. We do not believe that all of these posts, especially ones similar to the non-falsifiable Scenario 1, should be necessarily taken down, limited or otherwise acted upon. We do believe, however, that it is important for the platforms to provide predictability as to what their policies intend to cover. As is evident, no platform currently has fully comprehensive policies to address the spectrum of content that could be classified in this category. Scenario 4 has been added to this analysis as of Sept. 11, 2020.

Scenario 1: “The election is rigged!”
Scenario 2: “The election is rigged because I heard a postal worker has been bought off by party X to burn all party Y ballots.”
Scenario 3: “The election is rigged! Here is a video!” [Video that is real but taken out of context, or manipulated, through artificial means or cut to change the intent/message/outcome of the original video
Scenario 4: Candidate declares victory before authoritative media outlets call the election.

Delegitimization of Election Results				
	Scenario 1: Generic, Non-falsifiable Claim	**Scenario 2: More Specific, Non-falsifiable Claim**	**Scenario 3: Specific, Falsifiable Claim with Purported Evidence**	**Scenario 4: Specific Falsifiable Claim from Candidate**
Facebook	Non-comprehensive	Comprehensive: labeled	Comprehensive: labeled	Comprehensive: labeled
Twitter	Comprehensive: labeled or removed	Comprehensive: labeled or removed	Comprehensive: labeled or removed	Comprehensive: labeled or removed
YouTube	None	None	Comprehensive: video will be removed.	None
Pinterest	Non-Comprehensive	Comprehensive: removed	Comprehensive: removed	Non-Comprehensive
Nextdoor	None	Non-Comprehensive	Non-Comprehensive	None
TikTok	None	Non-Comprehensive	Non-Comprehensive	None

*Figure 3: Figure 3: UPDATE: This chart has been updated as of Sept. 11, 2020 to reflect policy changes from Twitter and Pinterest, and scenario 4 was added. The changes to the chart are shown in red. The evaluation of the policies by platform was informed by their community guidelines and standards linked here: **Facebook**, **Twitter**, **YouTube**, **Pinterest**, **Nextdoor** and **TikTok**. For more detail about each platform’s policies and justification for each evaluation, see the attached PDF at the end of the blog post.*

Already, posts from blue-checkmark users and highly visible politicians have made claims about whether the election will be “rigged” or “fraudulent.” As this table shows, companies have been reluctant to wade into this murky category of political speech. Without some platforms leading the way with comprehensive, transparent policies, it can be difficult to imagine more platforms creating such policies. It is possible, and necessary, to construct such policies that make it harder for actors to exploit this gap in policy. However, the success of these policies also includes a crucial component — how these policies are enforced and what exceptions, if any, apply to certain judgement calls on highly visible content.

Policy Enforcement

Platforms have a variety of moderation responses at their disposal for addressing election-related content that violates their policies. Since 2016, Facebook, for example, has used a framework called Remove, Reduce, Inform. As its name suggests, the policy removes content that violates the platform's policies, reduces the reach of problematic content that doesn't violate its policies but can still be harmful, and informs users with additional information on the content they share and see on the platform. While other platforms have different rubric terms, these three types of intervention are common across the industry. The different enforcement options are also informed by the gravity of the infringement, the nature of the account posting the content, and prior infringements made by the account posting the content.

One critical piece to the discussion on enforcement is an established precedent — sometimes a result of deliberate policy, sometimes what seems to be ad hoc — that some content can be exempt from enforcement if the platform believes "the public interest in seeing it outweighs the risk of harm." Facebook has codified this exemption in its "newsworthiness" clause, which allows content that would otherwise be in violation of Facebook's policies to stand "if it is newsworthy and in the public interest." Likewise, Twitter has created a "public-interest exception" exempting content that would otherwise violate its policies "if it directly contributes to understanding or discussion of a matter of public concern." To date, these policies have been enforced inconsistently, and both candidates and citizens have been left wondering where the lines will be drawn as we approach Election Day. Inconsistent enforcement and vague guidelines have raised questions around the motivations for each platform, and whether proprietary political concerns could influence content policy decisions. Those questions weaken the ability for the platforms to act decisively against election disinformation and can be best dispelled with regular, predictable interpretations of these important exceptions.

Conclusion

Crafting comprehensive policies and enforcing them in a clear and transparent manner is critical. There are and will be legitimate concerns about voter safety and election integrity during this unprecedented election; users will post about COVID-19 hotspots, long lines, or other intimidating scenarios that may deter citizens from voting. In practice, determining the line between posting “misleading” content and airing grievances could be a challenge — e.g., when a user shares an opinion or a prediction. In a recent Facebook post, Facebook CEO Mark Zuckerberg acknowledged the challenge of maintaining a public space for these conversations while also combating the spread of misinformation. Difficulties striking this balance are not new. However, this does not diminish the stakes of getting it right. Content that uses misrepresentation to disrupt or sow doubt about the larger electoral process or the legitimacy of the election can carry exceptional real-world harm.

Combating online disinformation requires action to be taken quickly before content goes viral or reaches a large population of users predisposed to believe that disinformation. Quick and effective action will require the platforms to make decisions against a well-documented framework and for those decisions to be enforced fairly — without adjusting those actions due to concerns about the political consequences.

We believe platforms can be more specific about when misinformation becomes a falsifiable statement and in which situations they will act, and that none of these companies should shy away from using their own free expression rights to provide fact-checks or counter-messaging when appropriate.