The Center for Democracy & Technology welcomes the opportunity to provide comments on Policy Advisory Opinion 2021-02, regarding Meta’s policy on cross-checking the moderation decisions for certain high-profile or influential accounts. Below, we respond to two of the questions posed by the Oversight Board in its call for public comment.

**Whether a cross-check system is needed and if it strengthens or undermines the protection of freedom of expression and other human rights.**

In general, yes, a cross-check system can provide an important opportunity for close evaluation of the ways in which Facebook and Instagram’s rules apply to powerful, high-profile, or at-risk users. Content moderation systems will always be prone to error, given the sheer volume of content uploaded to online services and the complexity of evaluating human communication against a single set of rules. Having a risk-based model that helps to prioritize certain content for additional review is, in general, a useful safeguard that can help avoid erroneous enforcement without obliging the user to take on the burden of appealing the company’s decisions.

Meta’s cross-check program, however, has suffered from two fundamental flaws: its application only to high-profile individuals and its focus exclusively on evaluating “false positives”. In combination, those flaws have led to the creation of essentially two tiers of Facebook/Instagram usage, where already powerful or influential users were more likely to have their speech remain on the service than regular users who did not benefit from the cross-check system. More influential users were shielded from false-positive removal of their content, including in circumstances where Meta services’ policies were unclear or the determination of whether to remove content was a close call, while regular users were not.

Meta has begun to address this disparity by making cross-checking for false positives available to all users, in the General Secondary Review program. The kinds of errors that Meta’s moderators or technical moderation tools may make will not be unique to high-profile or powerful users,¹ and the harms of over-removal are not felt only by those with the largest audiences. It will be crucial for Meta to continuously evaluate that program to understand how it is being applied to and experienced by users from a variety of different backgrounds and who post a variety of types of

---

content. This evaluation should include consultation with civil society organizations that represent the interests of users in different regions and cultural contexts, so that Meta can better understand the consequences and real toll of over-removal on regular users and whether the cross-check ranker is effectively identifying content that should undergo a General Secondary Review. Meta should make improvements to the General Secondary Review program based on this evaluation and also commit sufficient personnel and resources to ensure that the Secondary Review program provides a meaningful check for all users and does not function as a fig leaf that justifies the cross-check program for powerful users.

In addition to this broadening of the scope of the false-positive review in its General Secondary Review program, Meta should also implement a false-negative review for its high-profile users (in what is now called the “Early Response Secondary Review” system). The focus on false-positive review in its cross-check system has shielded Meta from a risk of high-profile public criticism for overbroad content removal, but it does not address the risk of abuse of Meta’s services by high-profile individuals. As discussed below, the issue of false-negative decisions is a vital element of how Meta moderates content posted by high-profile individuals and has significant consequences for the overall fairness of the content moderation on Meta services and the societal impact of high-profile individuals’ speech.

_Cross-check is designed to be a “false positive” prevention mechanism. What are the checks and balances, if any, this system should contemplate to mitigate the risks of “false negatives” [erroneous lack of action on violating content]?_

Meta should incorporate a check for false-negative results, or erroneous decisions not to remove or otherwise action content, into its Early Response Secondary Review process. According to Meta, the 660,000 users and entities who currently undergo ER Secondary Review represents users in the following categories: “elected official, journalist, significant business partner, human rights organization”, as well as those with a large number of followers or those who post about sensitive topics. The ER Secondary Review list presents a useful starting point for Meta to more carefully consider how high-profile and powerful individuals may be abusing their services and causing real-world harm, even if there will not be perfect overlap between users facing a high risk of false-positive results and those facing a high risk of false-negative results.

For example, as CDT discussed in our comments to the Oversight Board in its consideration of the suspension of Donald Trump’s account, there are certain categories of policy violation in

---


3 E. Llansó, _CDT’s Comments to Facebook Oversight Board on 2021-001-FB-FBR (Case Regarding Suspension of Trump’s Account)_ , Center for Democracy & Technology (Feb. 11, 2021)
which false-negative moderation decisions pose a heightened risk to public safety. Specifically, in consideration of the risk of incitement to violence, we urged the Oversight Board to consider the six-part threshold test from the Rabat Plan of Action, which states that “the speaker’s position or status in society” is a necessary element to the assessment of whether a speaker’s statement is likely to incite violence.\footnote{Report of the United Nations High Commissioner for Human Rights on the expert workshops on the prohibition of incitement to national, racial or religious hatred, United Nations, A/HRC/22/17/Add.4 at 11 (Jan. 11, 2013) \url{https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf}.} We urged that Facebook should “develop or improve their internal escalation process for moderation decisions about content posted by political leaders. . . . Given the risk and intensity of political violence around the world, Facebook should develop specialized teams within its Trust & Safety structure that have the expertise to assess the risk that statements by political leaders will incite violence.”\footnote{CDT’s Comments to Facebook Oversight Board on 2021-001-FB-FBR, supra n.3.}

With the cross-check system, Meta already has a significant amount of the infrastructure in place necessary to perform this sort of check—not just to evaluate whether political leaders’ speech is being removed unnecessarily, as under the current ER Secondary Review process, but to verify whether other reported content complies with the services' policies against incitement, and to include a deeper contextual assessment by specialized staff. In other words, Meta should adapt the ER Secondary Review process from serving as one that privileges the speech of the powerful and transform it into a process that considers the full context that surrounds powerful speakers, including when that context demonstrates that their speech is more likely to incite offline violence.

Incitement to violence by political leaders is not the only category of content and user that should undergo ER Secondary Review for false-negative results. Meta will need to carefully develop its estimation of what content or activity poses a high risk of false-negative decisions in its moderation system. This could potentially encompass many kinds of users and content: for example, Meta may find a generally high rate of false negatives in content in certain languages, or from certain sub-populations, where the real meanings of posts are not well-understood by moderators and where technical tools are not adequately trained. These are important dynamics to understand for Meta’s content moderation system overall, but may yield an overwhelming number of accounts potentially subject to false-negative review under the ER Secondary Review System. Accordingly, CDT recommends that, in addition to general characteristics of posts or accounts that lead to a high risk of false negatives, Meta consider the risk that a false-negative decision poses to the public or to vulnerable groups and communities, as opposed to the risk such decisions pose to the user who posted the content.\footnote{In general, there seems to be a low risk to the user who posted content of a false-negative result, as the outcome in such cases is that content the user wished to post on the service remains available on the service. The risks of a failure to accurately enforce the service’s content policies fall largely on other users or the general public.}
Thus, CDT recommends that Meta adapt its risk framework to consider, in particular, the risk to the public or to vulnerable groups and communities of false negatives for specific topics/content policy categories, as well as the risks that certain users or entities pose by virtue of their role in society (e.g. political leader) or the reach of their posts. These two factors—content policy category and role of the user—are both crucial to evaluating the risk of harm that a false-negative decision poses to other users or to the general public. The failure to remove a politician’s post of their own nude body, for example, is not likely to cause significant harm to other users, but the failure to remove a politician’s post that has been flagged as non-consensually shared intimate imagery could have significant ramifications for the individual depicted.

Certain categories of content, such as incitement to violence, carry a very high inherent risk associated with false-negative decisions, while others, such as disinformation, may require an initial situational analysis to establish the degree of risk that false-negative decisions pose. This situational analysis should include an assessment of imminence and scope of the potential harm to others. For example, false information about the mechanics of voting, posted by a political figure the day before an election, carries a high false-negative risk, in that the consequences of a false negative decision to remove the content could be experienced immediately and widely by a relevant population. But the same false information about voting procedures, shared by the same politician, 6 months away from an election, may not raise the same imminent risk to the population and so may not need to be prioritized as highly for false-negative review.

There is a risk, in developing such a system, that individuals will abuse it by flagging content with erroneous categories in order to subject that content to an additional layer of scrutiny, in hopes of it being (erroneously) removed. Meta should not let concerns about such abuse prevent it from incorporating false-negative reviews into their ER Secondary Review system, but it should build safeguards against such abuse into its processes. For example, and in addition to prioritizing certain user roles and content types for review, Meta should consider whether its frontline moderators can include clear signals in their initial evaluation of content that the reason given for flagging content bears no rational relation to the underlying content.

* * *

In its current form, the ER Secondary Review component of the cross-check system is a one-way ratchet of permissiveness for some of the most powerful users of Meta services. It creates two classes of Facebook and Instagram users, where already influential speakers have extra guarantees that they will be able to speak freely and use their “Voice” on these services, while everyday users have experienced the risk of false-positive moderation decisions without those safeguards.
While Meta’s move towards making the General Secondary Review system applicable to all users addresses part of this problem, Meta’s current policies nevertheless skew towards permissiveness for powerful speakers while failing to consider the heightened threats to other users and the public that these same speakers can sometimes pose. Meta is already applying special procedures to these accounts, which include not only politicians and influential speakers but also Meta’s business partners. Meta should commit to reviewing non-removal decisions for these accounts with the same level of care and scrutiny that it is already applying to removal decisions. It should also incorporate what it learns about the nature of false-positive and false-negative decisions into improvements in the policies, technology, and procedures of its content moderation system. This would help ensure that Meta’s standards and processes are geared not towards working in the interests of the already-powerful, but instead working in the public interest.

____________________
For more information, please contact Emma Llansó, Director of CDT’s Free Expression Project, ellanso@cdt.org.