October 2021

Policy Advisory Opinion Request: Cross-Check System

TABLE OF CONTENTS

Summary

Dear Members of the Oversight Board:

Facebook writes to request a Policy Advisory Opinion ("PAO") from the Oversight Board regarding the operation of Facebook's cross-check system, which is used to help ensure that enforcement decisions about our Community Standards are made accurately and with additional levels of human review. As set forth below, Facebook has been making changes to cross-check to ensure that we are protecting users' voice while also promoting authenticity, safety, privacy, and dignity on our platform. The changes Facebook is considering are borne out of our experience with prior iterations of cross-check, and we look forward to receiving the board's views on these changes.

We understand the board requested documents Frances Haugen shared with the *Wall Street Journal* about cross-check. We anticipate sharing the documents related to this request that are not subject to the attorney-client privilege once we have completed our review of them. We also will provide answers to the follow-up questions from our 22 September 2021 cross-check briefing, as well as additional questions, as part of this PAO process. Finally, the information below represents our current plans for the cross-check system. We will continue to update the board and provide it with supplemental, relevant information.

Facebook's Policy Advisory Opinion Request: Cross-Check System

I. Issue Statement and Policy Advisory Opinion Request

A. Issue Statement

Facebook users create billions of pieces of content each day. Moderating content at this scale presents challenges, including tradeoffs between important values and goals. We seek to quickly review potentially violating content, and remove it if it violates our policies. But we must balance this goal against the risk of "false positives" (erroneous removal of non-violating content[1]) to protect users' voice.

To balance these considerations, Facebook implemented the cross-check system to identify content that presents a greater risk of false positives and provide additional levels of review to mitigate that risk. Cross-check provides additional levels of review for certain content[2] that our internal systems flag as violating (via automation or human review), with the goal of preventing or minimizing the highest-risk false-positive moderation errors that might otherwise occur due to various factors such as the need to understand nuance or context. While cross-check provides additional levels of review, reviewers apply the same Community Standards[3] that apply to all other content on Facebook.[4]

The cross-check system plays a crucial function in helping to protect human rights. For instance, the cross-check system includes entities and posts from journalists reporting from conflict zones and community leaders raising awareness of instances of hate or violence. Cross-check reviews take into account the context that is helpful to action this content correctly. Cross-check reviews may also apply to civic entities, where users have

[1] Throughout the PAO we refer to the "removal" of content, which we are using to describe integrity actions more generally. These can also include, for example, the use of warning screens or removal of pages.
[2] Throughout the PAO, we refer to "content" that is reviewed through our cross-check system. We also use cross-check to review other actions such as removing a page or profile.
[3] Cross-check also applies to Instagram. Where we reference "Community Standards," it is inclusive of the Instagram Community Guidelines.
[4] For certain Community Standards, sometimes referred to as "escalation only policies," we require additional information and/or context to enforce. (We have previously briefed the board and staff on these policies.) If content that was cross-checked is escalated for further review, it may then be subject to a decision based on these context-specific policies.

a heightened interest in seeing what their leaders are saying.

In addition, cross-check serves an important role in managing Facebook's relationships with many of our business partners. Incorrectly removing content posted by a page or profile with a large following, for instance, can result in negative experiences for both Facebook's business partners and the significant number of users who follow them. We also apply cross-check to some very large Groups, where an error can impact hundreds of thousands or millions of users. Cross-check does not exempt Facebook's business partners or Groups from our content policies, but it does sometimes provide additional levels of review to ensure those policies are applied accurately.

Facebook has invested significant resources to improve cross-check over the last several years, with an increased focus beginning in 2020. While we have made progress in improving the system, there are still a number of difficult and impactful decisions where we seek the Oversight Board's guidance. Those decisions focus on how Facebook should balance our goals of removing content that violates our policies, on the one hand, while ensuring that it continues to foster open communication and free expression, on the other hand.

B. Questions

In light of these concerns, we seek the board's guidance on the following questions:

1. Because of the complexities of content moderation at scale, how should Facebook balance its desire to fairly and objectively apply our Community Standards with our need for flexibility, nuance, and context-specific decisions within cross-check?

2. What improvements should Facebook make to how we govern our Early Response ("ER") Secondary Review cross-check system[5] to fairly enforce our Community Standards while minimizing the potential for over-enforcement, retaining business flexibility, and promoting transparency in the review process?

3. What criteria should Facebook use to determine who is included in ER Secondary Review and prioritized as one of many factors by our cross-check ranker[6] in order

[5] As described below, our ER Secondary Review cross-check system relies on lists of users and entities whose content receives additional cross-check review if flagged as potentially violating.
[6] As described below, the cross-check ranker is a new system that ranks and prioritizes content for potential cross-check review based on false-positive risk using a set criteria.

to help ensure equity in access to this system and its implementation?

To aid the Oversight Board in responding to these inquiries, Facebook includes below an overview of the historical and current cross-check system, along with the further changes we are considering making going forward. Facebook also provides a summary of internal research we have conducted regarding the risk of "false positive" content moderation on the platform. We have included this internal research as attachments.

II. Cross-Check System

A. Overview of Cross-Check

Facebook and Instagram users post billions of pieces of content each day. Even with thousands of dedicated reviewers around the world, it is not possible to manually review every piece of content that potentially violates our Community Standards. The vast majority of violating content that we remove is proactively detected by our technology before anyone reports it. When someone posts on Facebook or Instagram, our technology checks to see if the content may violate the Community Standards. In many cases, identification is a simple matter. The post either clearly violates our policies or it doesn't. But in other cases, the content is escalated to a human reviewer for further evaluation.

Our primary review systems use technology to prioritize high-severity content, which includes "viral" content that spreads quickly. When the systems flag content for escalation, our reviewers make difficult and often nuanced judgment calls about whether content should remain on the platform. While we always aim to make the right decisions, we recognize that false positives do occur and some content is set for removal for violating Facebook's policies when it actually does not. Facebook has therefore invested in mistake prevention to further review false positives and mitigate them. Cross-check is one of these mistake-prevention strategies.

In response to the Oversight Board's prior recommendation regarding former President Trump, we described the cross-check system as follows:

> *Facebook's review teams are trained to ensure that their content decisions are accurate and consistent, based on the policies outlined in the Facebook Community Standards or Instagram Community Guidelines. This is especially important when people widely share potentially violating content on Facebook or Instagram, and we*

> *endeavor to make the right decision on this content due to the number of people who could see it.*
>
> *In these instances, we may employ additional reviews for high-visibility content that may violate our policies—for example, reporting from a war zone with graphic imagery that a closely-followed news source shares. This process, which we refer to as cross-check, means that our review teams will assess this content multiple times.*
>
> *These additional reviews are a supplemental safeguard to ensure we're accurately taking action on potentially violating content that more people see. It also helps us verify that when content violates our policies, including from public figures or popular Pages, we consistently remove it.*[7]

Cross-check is a system used to help ensure that enforcement decisions are made accurately and with additional levels of human review. If during cross-check a reviewer confirms that content violates our Community Standards, we enforce those policies and address the violating content accordingly. Depending on the complexity of the content, we may apply multiple levels of review, including in rare instances review by leadership. If the final reviewer determines that the content at issue does not violate our Community Standards, the reviewer can "overturn" the initial action and leave the content on the platform.

B. Historical Cross-Check Practices

We first implemented the system now known as cross-check in 2013. The details of the system have evolved over the years and, where possible, we have provided dates and date ranges explaining when these changes occurred.

To determine what content or entities received cross-check review, our teams identified and compiled lists of users or entities perceived to have higher associated risk with false positive actions against them. "False positive risk" refers to the risk of incorrect enforcement against content or entities that do not actually violate our Community Standards. To determine which users and entities were associated with a higher false

[7] *See* "Reviewing high-visibility content accurately," https://transparency.fb.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/ (last accessed 15 October 2021).

positive risk, our teams applied a variety of criteria, including the type of user or entity[8] (*e.g.*, an elected official, journalist, significant business partner, human rights organization), the number of followers, and the subject matter of the entity.

When users or entities identified on those lists posted content or took actions that our systems flagged as potentially violating our policies, we would add the content or entity to a queue for cross-check review. In general, from approximately 2013 until 2020, content or entities flagged as potentially violating our policies would be reviewed in the order they were added to the queue.

Beginning in 2020, we made changes so that most content in the queue was prioritized using a risk framework, which assigned a level of false-positive risk that could result if Facebook incorrectly removed that content. This risk framework generally relied on three factors: (1) the sensitivity of the entity, (2) the severity of the alleged violation, and (3) the severity of the potential enforcement action. Based on those factors, the content would be assigned one of three tiers of review: low (reviewed by contract reviewers), medium (reviewed by our markets team who have specialized regional expertise), and high (reviewed by our markets team and Early Response team who have deeper policy expertise and the ability to factor in additional context). Within those review tiers, the content in the queue was then prioritized by potential policy violation severity. Our reviewers would examine the content, confirm whether it violated our policies, and if so, enforce those policies. This review could include escalating levels of review, based on the potential complexity of the issue, including in rare instances, review by leadership. If the final reviewers concluded that the content did not violate our policies, the content would remain on the platform.

While content was pending in a cross-check review queue, the content at issue generally remained on the platform. We would, however, notify certain impacted users that their content was flagged as potentially violating our policies and—with the exception of high-severity violations that were removed immediately—allow them a 12- to 48-hour window to "self-remediate" (*i.e.*, remove the content themselves). If the user failed to self-remediate and the cross-check review determined that the content violated our policies, we would take an enforcement action and apply a strike against that user. (As we explain in our Transparency Center in response to a board recommendation, the accrual of strikes lead to restrictions on creating content and using our products.) If the user self-remediated within the window, however, we would not apply a strike against that

[8] Entity is a general term for where content could originate or appear, such as a user account, page, or group.

user. We ended the self-remediation window as of May 2021 because of equity and legitimacy concerns.

The lists for cross-check review sometimes have been used to intervene in enforcement systems outside of the cross-check review process with the aim of preventing potentially incorrect enforcement actions against entities appearing on those lists. Historically the practice has been referred to as "allow-listing." We have been reviewing and adjusting this practice and plan to discuss this during our briefings for the PAO with the board.

C. Current Cross-Check Practices

As with all of our policies and processes, we continually look for ways to improve and we are constantly making changes. Earlier this year, we conducted another holistic analysis of our historical cross-check practices and identified additional opportunities to improve the system. We have since implemented a number of changes to address these considerations and believe we have made significant progress. One structural change is that the cross-check system is now made up of two components: "General Secondary Review" and "Early Response (ER) Secondary Review." While we will continue to use the list-based approach described above for inclusion in ER Secondary Review for a percentage of certain users and entities, with General Secondary Review, we are in the process of ensuring content from *all users and entities* on Facebook and Instagram are eligible for cross-check review based on a dynamic prioritization system called "cross-check ranker."

General Secondary Review involves contract reviewers and people from our markets team who perform a secondary review of content and entities that may violate our policies before an enforcement action is taken. This review does not rely solely on the identity of a user or entity to determine what content receives cross-check review. The cross-check ranker ranks content based on false positive risk using criteria such as topic sensitivity (how trending/sensitive the topic is), enforcement severity (the severity of the potential enforcement action), false positive probability, predicted reach, and entity sensitivity (based largely on the compiled lists, described above). Our research has determined that these criteria are important factors for identifying content that poses the highest false-positive risk. The cross-check ranker is already used for the majority of cross-check reviews today.

ER Secondary Review is similar to the legacy cross-check system. To determine which content or entities receive ER Secondary Review, we continue to maintain lists of users and entities whose enforcements receive additional cross-check review if flagged as potentially violating the Community Standards. We have, however, added controls to

that process of compiling and revising these lists. Prior to September 2020, most employees had the ability to add a user or entity to the cross-check list. After September 2020, while any employee can *request* that a user or entity be added to cross-check lists, only a designated group of employees have the authority to make additions to the list. We are also considering annual audits of cross-check lists, exploring ways to include time limits and periodic re-verification requirements for inclusion, and improving our governance structure to include additional analysis and controls in place to define the list of users and entities eligible for this review.

In recent months, Facebook reviews an average of several thousand cross-checked jobs per day, with a large majority completed in General Secondary Review.[9] ER Secondary Review now makes up the minority of these daily reviews. We anticipate a continued shift in the number of cross-check review jobs being the result of General Secondary Review prioritization through the end of 2021 and into 2022.

If a piece of content is from an individual or entity that is included as part of ER Secondary Review, it is typically first reviewed by the markets team. The Early Response team will then review to confirm whether the content is violating. In general, if the markets team finds that the content does not violate our policies, the Early Response team will not review. If a piece of content is from an individual or entity that is prioritized by the cross-check ranker, contractors or the markets team typically review it, unless there is additional Early Response team capacity to review. As with legacy cross-check, high complexity issues may receive additional review, including in rare instances review by leadership. If the final review finds that it violates our Community Standards, we remove it. If our reviews find that it does not violate, we leave it up.

As of October 1, 2021, approximately 550,000 users and entities have actions that require some form of ER Secondary Review based on inclusion on the lists described above. Examples of users and entities eligible for ER Secondary Review include, but are not limited to:

- ***Entities related to escalation responses or high-risk events.*** Currently, there is an informal process in place where teams preparing for a high-risk event identify entities at high risk of over-enforcement. For instance, if a user's controversial content is going viral (*e.g.*, live video of police violence), we may identify that user for ER Secondary Review to prevent erroneous removal.

[9] Relative to the millions of pieces of content being flagged and actioned for violating our Community Standards daily, this is a small proportion.

- ***Entities included for legal compliance purposes***. We use ER Secondary Review in certain instances to comply with legal or regulatory requirements.

- ***High-visibility public figures and publishers***. We identify entities for ER Secondary Review because over-enforcement may result in a negative experience for a large segment of users.

- ***Marginalized populations.*** We identify human rights defenders, political dissidents, and others who we believe may be targeted by state-sponsored or other adversarial harassment, brigading, or mass reporting in order to protect against these attacks.

- ***Civic Entities***. We follow objective criteria and the expertise of our in-region policy teams to identify politicians, government officials, institutions, organizations, advocacy groups, and civic influencers. We include these entities for ER Secondary Review in order to prevent mistakes that would limit non-violating political speech and inadvertently impact discussion of civic topics like elections, public policy, and social issues. We aim to ensure parity across a country's civic entities—for example, if we include a national cabinet ministry in ER Secondary Review, we would include all ministries in that country's government in ER Secondary Review.

We are currently reviewing how to improve the criteria for identifying entities who should receive ER Secondary Review. For instance, we are exploring evolving our criteria in areas such as the number of followers, the number of previous false positive enforcements, legal/regulatory requirements, as well important political/societal issues.

Although we have made significant improvements to the cross-check system, we are still exploring ways to further ensure that this system appropriately balances our goals of removing content that violates our Community Standards while ensuring that we minimize our enforcement mistakes that have the greatest impact. We will continue to share updates as we refine cross-check and further reduce our false-positive content enforcement rate. We welcome the Oversight Board's guidance and look forward to answering any further questions you have.

III. Research on Cross-Check

To inform the initial development of the cross-check ranker, we interviewed fourteen internal stakeholders across the operations, policy, and product teams, seeking to assess

the various risks of over-enforcement. We chose internal stakeholders due to the complications of explaining how enforcement works, but are considering external engagement in the future. The Facebook stakeholders we interviewed believe false-positive enforcement decisions are riskier when they could contribute to the perception that Facebook is intentionally limiting free expression. These concerns are at their most significant when the content at issue relates to political topics, social justice issues, counter-speech, and any "borderline" situations where confirming a policy violation, if one exists, requires contextual decision-making. The stakeholders also believe that false-positive mistakes are riskier when they potentially involve more significant negative experiences to a user's ability to use our services, even if they are not immediate. And mistakes that result in enforcement action against an entire page are generally perceived as more severe than those that result in removal of only a single piece of content.

In our interviews, we also found that the perceived vulnerability and value of the entity in question, the sensitivity of the topic, and the severity of the potential enforcement were some of the most important factors in assessing the negative experience related to over-enforcement. Stakeholders also focused on the legitimacy of the enforcement action—including whether a particular mistake-prevention effort helps to mitigate any impact of Facebook's overall content moderation efforts—and its impact on the perception that Facebook is inappropriately limiting a user's voice.

We have already incorporated some of these research findings into our current cross-check system, and the cross-check ranker is driven in significant part by the factors we identified in consultation with our stakeholders. As discussed above, the cross-check ranker considers topic sensitivity, enforcement severity, false-positive probability, predicted reach, and the nature and importance of the entity in prioritizing content for cross-check review. We are continuing to explore how to incorporate factors such as perceived legitimacy and impact on voice into the ranker as well. We will continue to dedicate additional research and resources to improving the cross-check system, and we look forward to the Oversight Board's recommendations on how we can best ensure that this system reflects Facebook's core principles and aims.