Policy advisory opinion on Meta’s cross-check program

I. Executive summary

In October 2021, following disclosures about Meta's cross-check program in the Wall Street Journal, the Oversight Board accepted a request from the company to review cross-check and make recommendations for how it could be improved. This policy advisory opinion is our response to this request. It analyzes cross-check in light of Meta's human rights commitments and stated values, raising important questions around how Meta treats its most powerful users. As the Board began to study this policy advisory opinion, Meta shared that, at the time, it was performing about 100 million enforcement attempts on content every day. At this volume, even if Meta were able to make content decisions with 99% accuracy, it would still make one million mistakes a day. In this respect, while a content review system should treat all users fairly, the cross-check program responds to broader challenges in moderating immense volumes of content.

According to Meta, making decisions about content at this scale means that it sometimes mistakenly removes content that does not violate its policies. The cross-check program aims to address this by providing additional layers of human review for certain posts initially identified as breaking its rules. When users on Meta's cross-check lists post such content, it is not immediately removed as it would be for most people, but is left up, pending further human review. Meta refers to this type of cross-check as "Early Response Secondary Review" (ERSR). In late 2021, Meta broadened cross-check to include certain posts flagged for further review based on the content itself, rather than the identity of the person who posted it. Meta refers to this type of cross-check as "General Secondary Review" (GSR).

In our review, we found several shortcomings in Meta's cross-check program. While Meta told the Board that cross-check aims to advance Meta's human rights commitments, we found that the program appears more directly structured to satisfy business concerns. The Board understands that Meta is a business, but by providing extra protection to certain users selected largely according to business interests, cross-check allows content which would otherwise be removed quickly to remain up for a longer period, potentially causing harm. We also found that Meta has failed to track data on whether cross-check results in more accurate decisions, and we expressed concern about the lack of transparency around the program.

In response, the Board made several recommendations to Meta. Any mistake-prevention system should prioritize expression which is important for human rights, including expression of public importance. As Meta moves towards improving its processes for all users, the company should take steps to mitigate the harm caused by content left up during additional review, and radically increase transparency around its systems.

Key findings

The Board recognizes that the volume and complexity of content posted on Facebook and Instagram pose challenges for building systems that uphold Meta's human rights

commitments. However, in its current form, cross-check is flawed in key areas which the company must address:

Unequal treatment of users. Cross-check grants certain users greater protection than others. If a post from a user on Meta's cross-check lists is identified as violating the company's rules, it remains on the platform pending further review. Meta then applies its full range of policies, including exceptions and context-specific provisions, to the post, likely increasing its chances of remaining on the platform. Ordinary users, by contrast, are much less likely to have their content reach reviewers who can apply the full range of Meta's rules. This unequal treatment is particularly concerning given the lack of transparent criteria for Meta's cross-check lists. While there are clear criteria for including business partners and government leaders, users whose content is likely to be important from a human rights perspective, such as journalists and civil society organizations, have less clear paths to access the program.

Delayed removal of violating content. When content from users on Meta's cross-check lists is identified as breaking Meta's rules and while undergoing additional review, it remains fully accessible on the platform. Meta told the Board, that, on average, it can take more than five days to reach a decision on content from users on its cross-check lists. This means that, because of cross-check, content identified as breaking Meta's rules is left up on Facebook and Instagram when it is most viral and could cause harm. As the volume of content selected for cross-check may exceed Meta's review capacity, the program has operated with a backlog which delays decisions.

Failure to track core metrics. The metrics that Meta currently uses to measure cross-check's effectiveness do not capture all key concerns. For example, Meta did not provide the Board with information showing it tracks whether its decisions through cross-check are more or less accurate than through its normal quality control mechanisms. Without this, it is difficult to know whether the program is meeting its core objectives of producing correct content moderation decisions, or to measure whether cross-check provides an avenue for Meta to deviate from its policies.

Lack of transparency around how cross-check works. The Board is also concerned about the limited information Meta has provided to the public and its users about cross-check. Currently, Meta does not inform users that they are on cross-check lists and does not publicly share its procedures for creating and auditing these lists. It is unclear, for example, whether entities that continuously post violating content are kept on cross-check lists based on their profile. This lack of transparency impedes the Board and the public from understanding the full consequences of the program.

The Oversight Board's recommendations

To comply with Meta's human rights commitments and address these problems, a program that corrects the most high-impact errors on Facebook and Instagram should be structured substantially differently. The Board has made 32 recommendations in this area, many of which are summarized below.

As Meta seeks to improve its content moderation for all users, it should prioritize expression that is important for human rights, including expression which is of special public importance. Users that are likely to produce this kind of expression should be prioritized for inclusion in lists of entities receiving additional review above Meta's business partners. Posts from these users should be reviewed in a separate workflow, so they do not compete with Meta's business partners for limited resources. While the number of followers can indicate public interest in a user's expression, a user's celebrity or follower count should not be the sole criterion for receiving additional protection. If users included due to their commercial importance frequently post violating content, they should no longer benefit from special protection.

Radically increase transparency around cross-check and how it operates. Meta should measure, audit, and publish key metrics around its cross-check program so it can tell whether the program is working effectively. The company should set out clear, public criteria for inclusion in its cross-check lists, and users who meet these criteria should be able to apply to be added to them. Some categories of entities protected by cross-check, including state actors, political candidates and business partners, should also have their accounts publicly marked. This will allow the public to hold privileged users accountable for whether protected entities are upholding their commitment to follow the rules. In addition, as around a third of content in Meta's cross-check system could not be escalated to the Board as of May-June 2022, Meta must ensure that cross-checked content, and all other content covered by our governing documents, can be appealed to the Board.

Reduce harm caused by content left up during enhanced review. Content identified as violating during Meta's first assessment that is high severity should be removed or hidden while further review is taking place. Such content should not be allowed to remain on the platform accruing views simply because the person who posted it is a business partner or celebrity. To ensure that decisions are taken as quickly as possible, Meta should invest the resources necessary to match its review capacity to the content it identifies as requiring additional review.

II. Request from Meta

1. The Oversight Board first became aware of cross-check in 2021 when deciding its case on the suspension of former US President Donald Trump's accounts. Although Meta did not mention cross-check in its initial referral or materials sent to the Board, it described the cross-check program in response to a Board question about any different treatment the account may have received. As part of its May 2021 decision, the Board made two recommendations relevant to the cross-check program:

- "Produce more information to help users understand and evaluate the process and criteria for applying the newsworthiness allowance, including how it applies to influential accounts."

- "The company should also clearly explain the rationale, standards and processes of the cross-check review, and report on the relative error rates of determinations made through cross-check compared with ordinary enforcement procedures."

2. In September 2021, the Wall Street Journal revealed documentation produced by former employee and company critic Frances Haugen. The Journal's reporting described cross-check as exempting Meta's most influential users from normal content moderation processes. The Independent reported that Frances Haugen said the company had "lied" to the Board about cross-check "repeatedly" during the Trump case. Internal Meta documentation published by the Journal revealed that some of its employees considered cross-check's 'whitelisting' practices "not publicly defensible." Similarly, according to the Journal, users benefiting from the cross-check system at the time were given a 24-hour "self-remediation" window to edit or remove violating content and thus avoid any Meta-imposed penalties.

3. On September 21, 2021, following the Wall Street Journal articles, the Board called on Meta to commit to transparency about the system. The following day, Meta held a briefing with the Board on cross-check. The Board concluded that "the team within Facebook tasked to provide information has not been fully forthcoming in its responses on cross-check. On some occasions, Facebook failed to provide relevant information to the Board, while in other instances, the information it did provide was incomplete."

4. Shortly after the Board called for greater transparency on cross-check, Meta submitted this policy advisory opinion request. After briefly summarizing the system, Meta described cross-check as a program that "provides additional levels of review for certain content 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." Meta defines false positives as the mistaken removal of content that does not violate the content policies that establish what is allowed on Facebook and Instagram.

5. Meta posed the following three questions to the Board:

   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?

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

   What criteria should Facebook use to determine who is included in ER Secondary Review and prioritized as one of the many factors by our cross-check ranker in order to help ensure equity in access to this system and its implementation?

6. The Board accepted Meta's request on October 21, 2021. Following this acceptance, the Board sent Meta questions. The Board asked Meta 74 questions. 58 were answered fully, 11 were answered partially, and five were not answered. Meta took months to respond to some of these questions.

7. The Board also received 87 public comments related to this policy advisory opinion: nine from Asia Pacific and Oceania, two from Central and South Asia, 12 from Europe, three from Latin America and the Caribbean, three from the Middle East and North Africa, three from Sub-Saharan Africa, and 55 from the United States and Canada. To read public comments submitted for this policy advisory opinion please click here. In addition, the Board held four regional workshops focused on the cross-check program.

8. Based on its analysis of this information, independent research, and stakeholder engagement, the Board now answers Meta's questions and provides its assessment of the cross-check system. Meta also told the Board it has made significant changes to the cross-check program over the past year. The Board understands these changes to be, at least in part, an effort to respond to public criticisms of the program. The Board's explanation of the program and its analysis of it is based on how Meta states the program is currently functioning. However, at times the Board references its understanding of past practices as they inform likely areas of recurrent risk.

9. The Board explored whether the program serves in practice to address and mitigate adverse impacts according to Meta's human rights responsibilities. This analysis, grounded in international human rights standards and Meta's stated values and commitments, implicates important questions of how Meta treats its most influential and powerful users, permits content to flow across its platforms, and provides information to the public about its actions.

III. Meta's cross-check system

Meta's explanation of why it uses cross-check

10. Facebook and Instagram users create billions of pieces of content each day. Meta is constantly moderating content; or screening, evaluating, and taking action on it based on the company's content policies. On Facebook, these policies are the Community Standards, and on Instagram, they are the Community Guidelines.

11. According to Meta, moderating content at this scale presents challenges, and its human reviewers and automated systems sometimes mistakenly remove content that does not violate Meta policies. Meta refers to these decisions as false positives. False negatives are a form of under-enforcement and refer to content that violates Meta policies but is not determined to be violating on review. Under-enforcement also includes violating content that is not detected by automated or human

reviewers, and system design choices that allow violating content to remain visible after a first review.

	Review determined content does not violate the Community Standards	Review determined content violates the Community Standards
Content does not violate the Community Standards	True Negative	False positive (over-enforcement)
Content violates the Community Standards	False negative (under-enforcement)	True positive

12. The cross-check system only addresses over-enforcement, or false positives. Through this system, Meta delays taking any enforcement action on select content initially identified as violating to allow for possible additional review with the aim of avoiding false positives.

13. Meta described cross-check as a mistake-prevention strategy that allows it to balance protecting users' voice from false positives with the need to quickly remove violating content. As part of the policy advisory opinion request, Meta highlighted the inclusion of "journalists reporting from conflict zones and community leaders raising awareness of instances of hate or violence," as well as civic actors where "users have a heightened interest in seeing what their leaders are saying."

14. The system further includes users that Meta describes as "business partners." These partners have dedicated points of contact at Meta. According to the company, these users include "health organizations, news publishers, entertainers, musicians, artists, creators and charitable organizations." The Board understands that this category includes users that are likely to generate money for the company, either through formal business relationships or because they draw users to the platform and keep them engaged there. The Board understands that "business partners" likely also include major companies, political parties and campaigns, and celebrities.

15. Meta told the Board that it adds "business partners" to cross-check to prevent mistaken deletions that limit the ability of users and advertisers to reach their audience and customers, and the economic and reputational impact that such errors may cause the company. For these users, Meta aims to avoid "negative experiences for both Facebook's business partners and the significant number of users who follow them."

16. Meta stated that it prefers under-enforcement compared to over-enforcement of cross-checked content, as “in the current business landscape maximizing the benefit of cross-check (preventing false positives) is generally considered to be more important than minimizing the cost of cross-check [i.e., views of violating content]. This is due to the perception of censorship.” The Board interprets this to mean that, for business reasons, addressing the “perception of censorship” may take priority over other human rights responsibilities relevant for content moderation.

How cross-check works

17. Meta’s ordinary content moderation processes apply to most users. When content is identified as violating Meta’s content policies, Meta takes an enforcement action. This includes content deletion and the application of warning screens, depending on the type of policy violation. Some violations can also lead to account-level penalties, such as suspension and termination. However, in some cases, content receives a different treatment, as is the case of the cross-check system.

18. Meta uses the term cross-check to refer to a false positive-prevention program. It provides for additional layers of review for content before enforcement action is taken. Escalation-only content policies, which can only be applied by specialized teams at Meta, may be applied during this enhanced review. These policies include the newsworthiness and spirit of policy allowances and all rules that Meta has determined require additional context to enforce. Cross-check review processes are triggered under two sets of circumstances.

19. First, cross-check provides guaranteed additional **human review of content** by specific entitled entities whenever they post content that is identified as requiring enforcement under Meta content policies. Meta calls this **Early Response Secondary Review** or **ERSR**. An "entity” is anything on Facebook or Instagram that can post content, such as Facebook pages, Facebook profiles, and Instagram accounts. Entities can represent individual people and groups or organizations. Meta creates and maintains lists of entities it has decided are entitled to receive the benefits ERSR provides. This means that if any entitled entity posts content that is identified as violating the Community Standards or Guidelines, it will not be removed according to the procedures that apply to regular users, but instead will be sent for extra levels of review. Because ERSR is based on lists, only certain pre-selected users receive this benefit.

20. The second part of the cross-check system provides additional review of certain content identified as violating Meta policies, regardless of the identity of the user who posted it. Meta calls this **General Secondary Review**, or **GSR**. Whenever any piece of content posted by any entity on the platform is identified as violating a Meta policy, by a human or automation, Meta uses an automated process called ‘cross-check ranker’ to instantaneously analyze various factors and determine if the content should be sent for additional review, and how it should be prioritized within

a queue of other content awaiting the same type of review. According to Meta, because this system is based on the characteristics of the content, content posted by any user on Facebook or Instagram is eligible to be selected for GSR. GSR was implemented in 2021, and the Board understands that, to some extent, it was developed and implemented across the platform in response to criticism of ERSR, including the Haugen revelations.

21. The initial detection of content in both types of cross-check that may trigger a review can happen either proactively, through automated Meta systems after the content is posted, or reactively, following user reports. The enforcement actions that may trigger cross-check review include content deletion and the application of warning screens, depending on the type of policy violation. As most content policy violations can lead to account-level penalties, such as suspension and termination, these types of enforcement are also impacted. Cross-check applies across Facebook and Instagram, except for some content types (e.g., reels, podcasts) that are not currently eligible for the program. According to Meta, "10% of organic content that is otherwise subject to integrity enforcement is not eligible for cross-check review today."

22. During the time after content eligible for cross-check (through GSR or ERSR) is identified for enforcement but before it is subject to the additional review process, this content remains fully accessible on the platform, even if the first assessment is that the content violates the Community Standards or Guidelines.

23. The Board understands that if Meta had more moderators available, more content in cross-check review queues would receive an additional human review. However, Meta has chosen to only guarantee additional human review for content that goes through ERSR, the system for entitled entities. Meta has not invested the resources needed for all content identified through GSR to receive additional human review. While the review paths for these two mechanisms differ, as described below, should any reviewer at any stage of the process find that content does not violate Meta policies, the review process ends, and the content remains on the platform.

Early Response Secondary Review (ERSR)

24. Meta states that it includes entities on ERSR lists by assigning them a "tag" that correlates to the nature and sensitivity of the entity. Specific tags correspond to different ERSR lists. Meta states that it applies an ERSR tag to entities corresponding to the following categories: (1) civic and government; (2) significant world events; (3) media organizations, businesses, communities and creators, including advertisers; (4) historically over-enforced; (5) legal and regulatory or entities for which erroneous action may present legal risk to Meta, for example in the context of ongoing litigation; (6) entities whose content is under review, meaning cases where action by any reviewer would undermine ongoing deliberation or would present risk to Meta. According to Meta, beyond the factors that it uses to determine whether an entity fits within any of the categories stated above, such as advertisement

spending or history of enforcement, entitlement to ERSR is also determined by an assessment of the impact a potential enforcement mistake would have on the company in terms of the level of company leadership that would be involved in finding a solution. In other words, a key rationale for ERSR is to avoid provoking people who have the means to engage senior-level executives directly or create public controversy those executives might need to remedy.

25. Meta informed the Board that it is currently consolidating and updating its ERSR lists. Previously, Meta's lists corresponded to the level of escalation that would be required to enforce content policies against a particular entity. According to Meta, all entities currently entitled to ERSR are now subject to the same review process. This process may include discretionary escalation to the highest levels of the company.

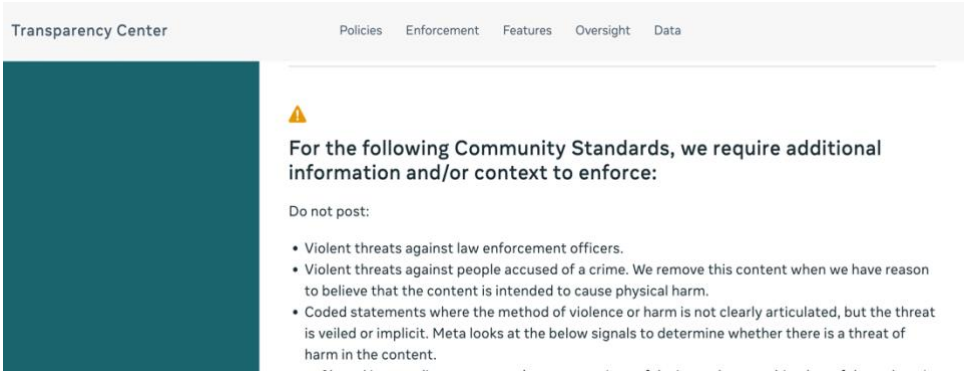
26. Meta told the Board that during the second quarter of 2022 it established general criteria for adding and removing entities from ERSR lists, and new processes for periodic audits and internal oversight. Meta did not provide details on these processes, and what actions might trigger the re-evaluation and removal of an entity. Meta did explain that, in general, the tags that place an entity on an ERSR list will expire after a year, and in theory the entitled entities would need to be assessed and tagged anew. According to Meta, this logic generally covers entities in the following categories: legal and regulatory; significant world events; media organizations; businesses, communities, and creators; historically over-enforced; and entities escalated for higher context review. Meta noted two exceptions to the one-year expiration rule. First, tags for entities in the civic and government category do not have a default expiration. Second, tags for entities in the other categories mentioned above may be given shorter ERSR entitlement at Meta's discretion.

27. Whenever a piece of content by any of the entitled entities is marked for enforcement by automated or human review, no enforcement action is taken, and the content is instead sent for **enhanced review by a human moderator**. This first level of enhanced review is done by what Meta refers to as a "**Regional Market Team**," a team within Meta. This team includes both Meta employees and hired contractors who have additional contextual and language knowledge about a specific geographic market. If a Market Team reviewer determines the content is non-violating, the process ends, and the content remains on the platform.

28. However, if the Market Team reviewer finds the content violates Meta's policies, the content remains on the platform while it is escalated further to what Meta calls the "**Early Response Team**" for another review. According to Meta, this team has "deeper policy expertise and the ability to factor in additional context."

29. The Early Response Team is also allowed wider discretion than other Meta content moderators and can apply content policies that "require additional information or context to be enforced." Meta often marks these content policies with a yellow exclamation point within each Community Standard, as shown below. For example,

at the end of Facebook's Violence and Incitement Community Standard, Meta prohibits "violent threats against law enforcement officials." According to Meta, the determination of whether to keep up or remove content that may violate these context-specific parts of the policy may only be made by a team that is allowed to factor in additional context, like the **"Early Response Team**."



30. The **Early Response Team** may also apply what Meta calls its "newsworthiness" and "spirit of the policy" allowances, which allow otherwise violating content to remain on the platform because Meta finds it is in the public interest or finds that, even though it violates the letter of a policy, it does not violate the intent of the policy. The Board also believes this discretion extends to the application of account-level penalties. However, as disclosed by Meta, the **Early Response Team** does not have language or regional expertise and it relies on translations and contextual information provided by the relevant Regional Market Team to assess the content.

31. At the time of the Board's briefings with Meta, approximately 0.01% of all content identified as needing enforcement under a Meta policy was escalated through cross-check to reviewers who may apply these contextual policies and allowances. Content posted by users on ERSR lists is guaranteed to reach those reviewers before any enforcement action: it may not be removed or have a warning screen applied by automated review, at-scale human reviewers, or **Market Team** reviewers. During the entire time the cross-checked content is awaiting its final determination, it remains on the platform, where users are free to like and share it.

32. Once the content is reviewed by the **Early Response Team**, if it is found violating, Meta may take the corresponding enforcement action, such as removing the content or applying a warning screen. However, Meta may also escalate the decision further. The Board understands that the escalation procedures at this phase are broadly discretionary. If the Early Response Team finds that the content "is an extreme edge case interpretation of [Meta's] policies" or it "presents a significant risk to the company or the community, and/or there is disagreement among internal stakeholders on how to respond," the Early Response Team may perform an additional review in conjunction with other Meta teams. According to Meta, these escalated reviews "include the views and input of the Content Policy subject matter experts (SMEs) and the local Public Policy, Comms, and Legal teams"

and may include input from other teams. After that review, it could even be escalated further to company leadership before receiving any enforcement action.

33. Additionally, Meta told the Board that if "the issue has significant service blocking, legal, regulatory, or safety risk, or where [it has] limited time to make a decision, [the Early Response Team] will on rare occasions escalate a decision directly to global senior leadership." Meta stated it assesses liability risk, urgency, geopolitical impact, service blocking risk and disagreement between internal teams as factors to escalate these decisions.

34. In summary, a piece of content posted by an entity on an ERSR list may receive up to five reviews before it is subject to enforcement, even if reviewers repeatedly find that it violates Facebook or Instagram rules and escalate it along the cross-check pathways:

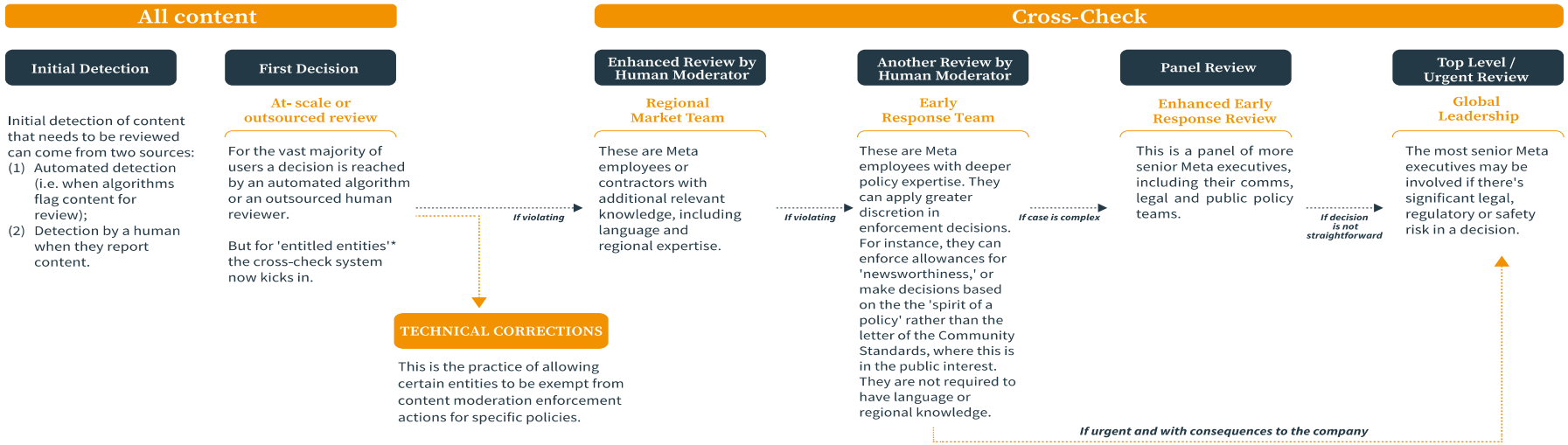
1. Initial review by automation or a human reviewer that identifies content for enforcement based on Meta's policies.
2. Regional Market Team review.
3. Early Response Team review. This is the first review that can authorize enforcement against the content. This team may request an enhanced Early Response Team review that incorporates other teams or pass directly to Global Leadership review.
4. Enhanced Early Response Team review with subject matter experts, Public Policy, Communications and Legal Teams.
5. Global Leadership review. This is a discretionary escalation from the Early Response Team based on the severity of consequences to the company.

   The process stops and the content remains on the platform if it is found non-violating at any stage of review.

How Cross-Check Works

ERSR: Early Response Secondary Review

Available only for "entitled entities" within special categories.
Intended to correct mistaken over-enforcement.

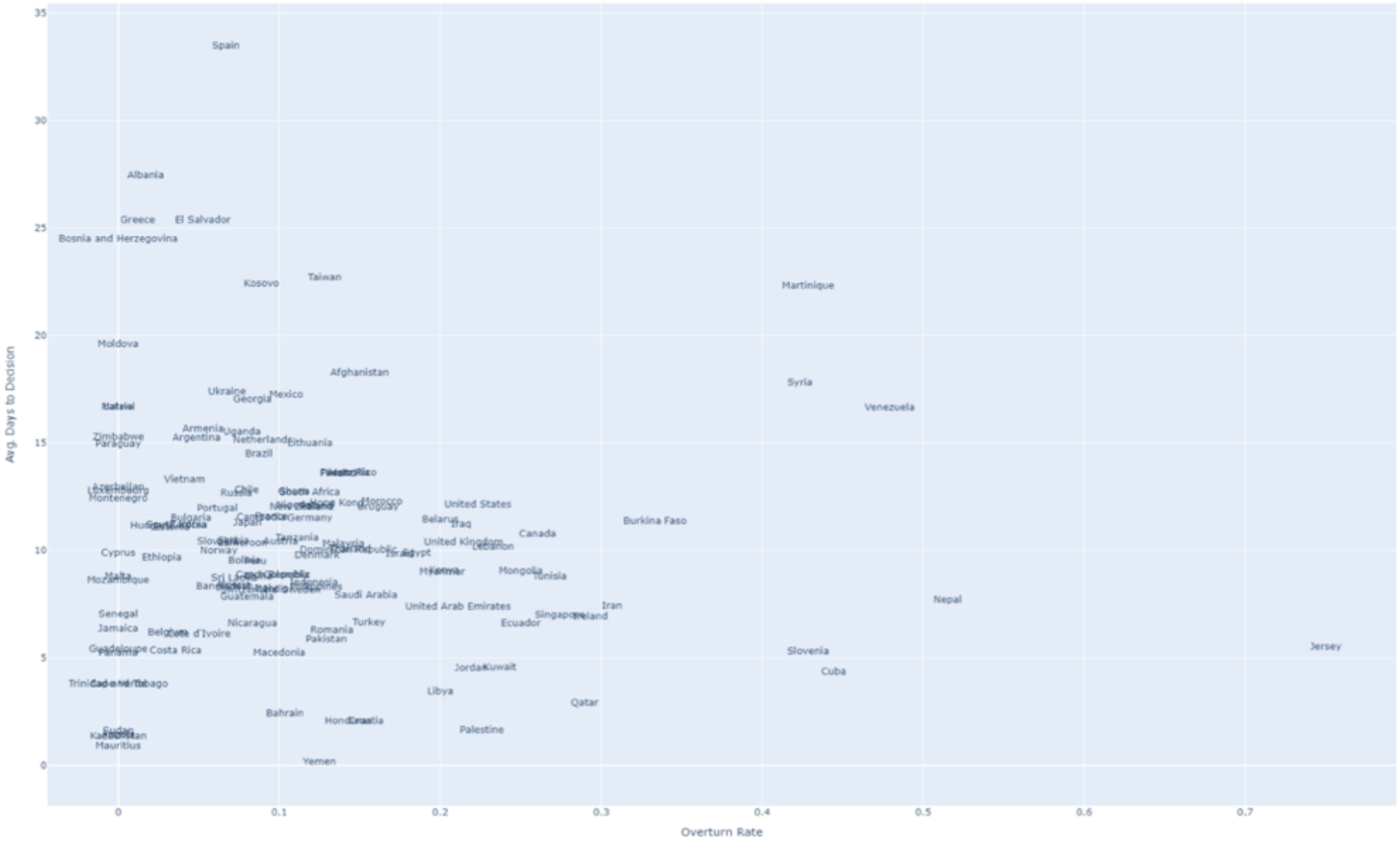


***Categories, as defined by Meta, of entitled entities that benefit from Early Response Secondary Review**

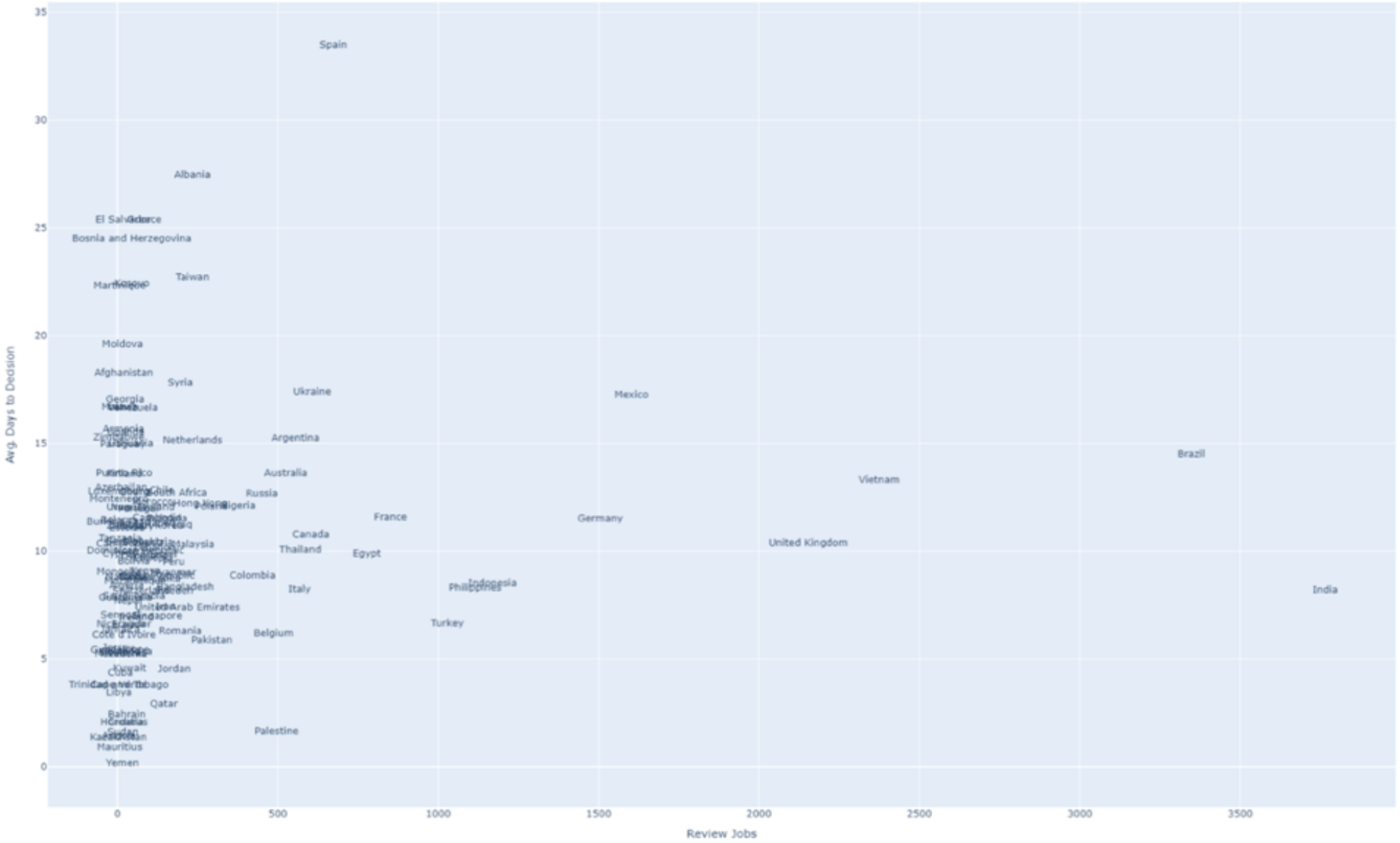
**Entities related to escalation responses or high-risk events*
**Entities included for legal compliance purposes*
**High-visibility public figures and publishers*
**Marginalized populations*
**Civic entities*

Content remains accessible throughout these reviews until the final removal decision is made.

35. The ERSR pathway may take several days. According to Meta, its internal goal for cross-check review by the Market Team ranges from 12 to 120 hours depending on the severity of the potential violation. In practice, Meta stated the mean time to final decision for Early Response Secondary Review is more than five days. For content posted by users in the United States, Meta states that it takes "approximately 12 days on average for a decision." Other countries have even slower mean resolution times. For instance, the mean time-to-decision for Afghanistan and Syria is about 17 days. In the information Meta provided the Board, the longest time a piece of content remained in the Early Response Secondary Review queue was 222 days. Meta provided several graphs with this data to the Board, displaying the mean time to decision between March 2021 and February 2022, with the overturn rate and the number of jobs, or pieces of content reviewed, and different countries.



(1) Scatterplot of average days to decision vs. overturn rate by country



(2) Scatterplot of average days to decision vs. number of jobs by country (excl. USA)

36. Meta states that content that spends more time awaiting review has been designated as low severity under its "violation severity framework." This scheme ranks content based on the specific Community Standard that the first review indicates it violates. Meta's framework ranks each Community Standard according to the potential harm that policy violations may cause, a determination that Meta has said it made based on the company's research. For example, it considers hate speech more harmful than spam, with potential hate speech prioritized ahead of spam in the ERSR queue.

37. That said, in September 2021, the Wall Street Journal reported that Brazilian soccer star Neymar posted non-consensual intimate imagery of another person on his Facebook and Instagram accounts. According to reporting by The Guardian, the video was online for over a day, and "an internal review of the Neymar posts found that the video was viewed 56 million times on Facebook and Instagram before removal" despite representing a clear violation of Meta content policies. According to Meta, the reason for the prolonged accessibility of this violating content was a "delay in reviewing the content due to a backlog at the time."

38. A core metric that Meta told the Board it uses to justify the cross-check system and assess how well it is working is the "overturn rate." This is the percentage of content that is found non-violating during cross-check review, reversing the initial determination and preventing the enforcement of content that Meta's rules allow. Meta provided several different figures to the Board about its overturn rate for ERSR content. According to Meta, for different time periods over the last year the overturn rate ranged from 30% to 90%. When the overturn rate is low, ERSR is keeping more content ultimately found violating on the platform during the multiple layers of cross-check review. When the overturn rate is high, ERSR is preserving more non-violating content from mistaken enforcement.

39. According to Meta, "most views happen when content is fresh, so speed in reviewing decisions and removing content quickly is crucial in preventing harm." Therefore, violating content that is subject to ERSR remains accessible on the platform throughout the period during which it is likely to receive the vast majority of its views.

General Secondary Review (GSR)

40. The second mechanism that Meta says forms part of its cross-check system is **General Secondary Review** (GSR). Whereas ERSR applies to all content posted by specific entitled entities, GSR may apply to any content posted on the platform, regardless of the poster, based on an algorithmic determination.

41. GSR is a relatively new system. In the fall of 2021, when Facebook whistleblower Frances Haugen disclosed information about cross-check, the Board understands her to have been referring to its previous iteration, which was based entirely on the

entity posting the content. The company now includes GSR within cross-check as well. Meta stated that GSR was available for content posted by some users in 2021 and fully implemented for content posted by all users in early 2022.

42. To decide what content to send through the GSR pathway before taking an enforcement action, Meta uses an algorithm (i.e., a set of rules that need to be followed by a computer for a specific task) called cross-check ranker. This algorithm is based on the following features: "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." Entity sensitivity is therefore a factor in both systems, although in ERSR it is the key factor and in GSR it is one factor among many. Meta stated that it has considered including additional factors and expects to do so in the future.

43. According to Meta, content must satisfy two conditions to be eligible for GSR. First, it needs to have been identified for enforcement (i.e., violating a Community Standard or Guideline) by automation or human review. Second, it must be marked by the cross-check ranker as high priority. If both conditions are met, the content is not enforced immediately and instead is placed in a queue for additional human review by a **Regional Market Team**. These are the same Market Teams that also perform the first enhanced review of content posted by ERSR-entitled entities.

44. The Market Teams are unable to review all content that is guaranteed review under ERSR and all the content that is placed in a queue for possible review under GSR. Because entitled entities on ERSR lists are guaranteed review, the Market Teams must first dedicate reviewer capacity to this content. With any remaining capacity, the relevant Market Team reviews the algorithmically identified GSR content. The Market Teams also review certain content outside the cross-check program, among other tasks they must prioritize.

45. Therefore, even though GSR content may be highly prioritized by the cross-check ranker algorithm as meriting additional review because it may have been identified as a likely false positive, the Market Team may not have capacity to review it. In some cases, if there is no review capacity at the Market Team level, and Meta has chosen to make available its outsourcing capacity, some GSR content may be sent for that additional review to outsourced human reviewers. If the GSR content is reviewed by a Market Team reviewer, in most cases that decision is final. If the content is found violating, it is generally subject to enforcement (e.g., removed or warning screen applied). If it is found not violating, it remains on the platform. However, if the **Early Response Team** has any additional capacity after its obligations to review all ERSR content prior to its possible removal, that team may review highly-prioritized GSR content that a Market Team reviewer has found violating before Meta proceeds to enforcement.

46. Similarly to the ERSR process, content that has been originally assessed as violating the Community Standards and placed in the GSR queue remains on the platform

while awaiting additional review. However, unlike ERSR, content in the GSR queue pending review will not remain on the platform indefinitely. Content that is not reviewed eventually "times out" of the GSR queue. When this happens, Meta reverts to its initial enforcement decision without further review. This means the action that would have been applied, such as removal or a warning screen, is applied on a delayed basis without additional review. If reviewers do not reach a specific piece of content in the GSR queue, it will stay on the platform for between two and four days before Meta removes it from the review queue and applies the enforcement action. At the same time, cross-check ranker continually identifies newer and more highly prioritized content, suspends enforcement on the content, and adds it to the GSR queue.

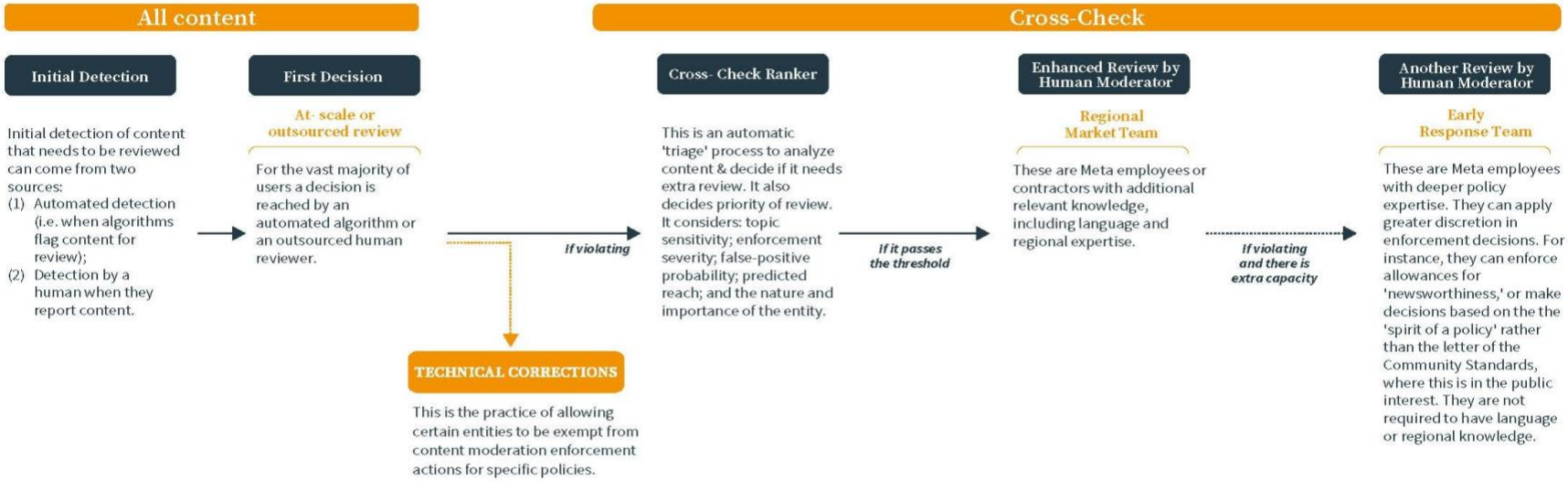
47. The overall effectiveness of GSR is limited by Meta's choice of how much reviewer capacity to provide to this type of review in each of its markets. The majority of GSR content is reviewed by an outsourced reviewer, a Market Team reviewer, or times out of the system. This means most GSR content never reaches the **Early Response Team** and, therefore, will never reach a level of review where contextual analysis, escalation-only policies, and policy allowances can be applied.

48. Meta also calculates the overturn rate for content that receives cross-check review through the GSR pathway. Meta provided the Board with different rates for this figure over the past year. At the time of the Board's briefings with Meta in February 2022, the overturn rate for General Secondary Review was about 80%. Meta later provided new information to the Board, stating that between March 2022 through May 2022, the overturn rate was about 70%. While there was also variation, these figures varied less over time. Most GSR content initially identified as violating is found to not violate any Meta policy with secondary review. As content times out of the GSR queue, it is highly likely that Meta is enforcing a significant number of false positives identified by cross-check ranker.

How Cross-Check Works

GSR: General Secondary Review

Can apply to any piece of content by any user.
Implemented fully in 2022.



Content remains accessible throughout these reviews until the final removal decision is made.

Cross-check and reported exemptions from enforcement

49. The Wall Street Journal reporting described cross-check as a system to exempt “VIP users from the company’s normal enforcement." Meta disclosed to the Board that it does have a system that blocks some enforcement actions outside of the cross-check system. Meta refers to this practice as “technical corrections,” and public reporting has described it as “allowlisting” and “whitelisting.”

50. “Technical corrections” are automatic exceptions to content policy enforcement. This means they override almost all automated or human reviewer attempts to apply an enforcement action for a preselected set of content policy violations. Every piece of content identified for enforcement is automatically checked to see if any “technical corrections” apply.

51. If the content is protected by a correction, it will be exempt from that specific enforcement. As explained by Meta, a “technical correction” applies only to a specific entity for a specific policy violation and does not serve to bar enforcement for other policy violations. At the time of the Board’s briefings with Meta, it stated that it applied about a thousand **technical corrections** per day. Meta did not disclose how many and what type of entities have benefited from a “technical correction.”

52. If the content is not protected by any correction, it is then checked for cross-check eligibility. At that point, Meta’s normal cross-check processes to identify if the user is an ERSR entitled entity or if the content is prioritized by cross-checker ranker for the GSR queue apply.

53. Meta first stated that it primarily applies "technical corrections” to "two violation type groups (spam/inauthentic behavior and impersonation).” Meta later confirmed that as of September 21, 2022, there are four active “technical corrections” and that this might also change over time.

54. Meta told the Board that “a limited number of ‘technical corrections’ remain, and [Meta] recognize[s] an ongoing need for them.” According to Meta, such “corrections help [Meta] prevent enforcement mistakes on content or entities that are highly unlikely to violate our policies and direct human review resources where needed most.”

55. Meta acknowledged shortcomings about its past technical corrections practices. Meta told the Board that the “lack of governance over practices in the past, [...] inadvertently resulted in some entities not receiving many enforcement actions.” Meta stated that “different teams could apply different corrections to the same entity in a way that, when combined, resulted in the entity and its content not receiving a wide variety of enforcement actions.” Meta stated that because this practice was “the inadvertent result of a decentralized system, [Meta] [is] taking

steps to ensure that there is a governance structure around the use of cross-check lists."

Cross-check in the context of government requests to remove content

56. When governments request that Meta remove content, Meta may remove the content because it violates company content policies. It also may remove or "geoblock" content for legal reasons, limiting its accessibility in certain areas. Meta has told the Board that it adds entities to cross-check ERSR lists to protect them from erroneous actions that may present legal risk to Meta, for example in the context of ongoing litigation.

57. According to Meta, government requests to remove content are addressed by specialized teams that may enforce on content immediately, regardless of whether it was posted by an ERSR entitled entity or could have been highly prioritized by cross-check ranker. In other words, removals resulting from government requests supersede cross-check privileges.

IV. Framework for Board analysis

International human rights standards

58. On March 16, 2021, Meta announced a Corporate Human Rights Policy, where it outlines its commitment to respect rights in accordance with the UN Guiding Principles on Business and Human Rights (UNGPs). The UNGPs, endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of businesses. These rights include, "at minimum, [...] those expressed in the International Bill of Human Rights," (Principle 12).

59. As a global corporation committed to the UNGPs, Meta should respect international human rights standards wherever it operates and address any adverse human rights impacts (Principle 11). This also means that Meta should "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts" (Principle 13).

60. The UNGPs also establish that businesses should carry out human rights due diligence to assess actual and potential impacts and act upon their findings (Principle 17). To do that effectively, businesses should monitor qualitative and quantitative indicators and incorporate input from impacted stakeholders (Principle 20).

61. Through its cases, the Board assesses the human rights impacts of specific enforcement decisions. When these cases reveal that Meta is causing a negative impact, or may not be taking steps to identify, monitor, and limit negative impacts more broadly, the Board makes appropriate corrective recommendations. In a

policy advisory opinion, the Board focuses directly on Meta's policy choices, including development and enforcement processes, to assess whether the company is upholding its commitment to respect rights under the UNGPs.

62. Applied to cross-check, the Board explored whether the program serves in practice to address and mitigate adverse human rights impacts according to Meta's responsibilities. The Board also closely examined the metrics Meta uses to determine the effectiveness of the program, and what that suggests about the company's objectives.

63. In its analysis, the Board finds that a wide array of rights may be impacted by the cross-check program. Freedom of expression, which includes the right to seek and receive information (Article 19, International Covenant on Civil and Political Rights; General Comment 34, 2011, para. 11), may be enhanced to the extent that cross-check serves to limit enforcement against content that does not violate platform policies. This results in positive impact for the posting user and those who wish to access their content.

64. The Board also notes that cross-check could, in theory, serve to ensure that those who face particular barriers to exercise their right to freedom of expression benefit from the additional layer of protection that the program may provide. Targeted mass reporting of non-violating content, for example, could be inhibited by a false-positive mistake-prevention system.

65. However, these positive effects may be limited if the system is designed primarily to protect or prioritize the expression of people who are already powerful. The Board also notes that the cross-check program raises non-discrimination concerns, as certain entities are afforded additional protection.

66. Further, the cross-check program's protection of violating content may contribute to an environment that inhibits expression from those who may be targeted by that violating content. The range of violating content that may be left on the platform for additional time could severely impact a variety of human rights, and the consequences may vary depending on the affected users' situations. Adverse human rights impacts will likely be felt more acutely by individuals and groups who face marginalization and discrimination.

67. The Board's analysis accounts for these standards. Its policy recommendations also acknowledge the limitations of Meta's ability to moderate content at scale. If Meta's moderation more accurately moderated the content of all users, it would not need special programs based on entitled entities to help advance its respect of human rights.

Meta's values

68. International human rights standards set parameters on Meta's policies and practices. Within those standards, however, social media companies may adopt different rights-respecting approaches. Meta's values should guide the company's discretionary decisions.

69. Meta has stated that it has five values that influence the development of enforcement of its content policies on Facebook and Instagram. These values are "Voice," "Authenticity," "Privacy," "Safety," and "Dignity." According to Meta, "Voice" is the company's "paramount" value. The Board finds that cross-check, and a false-positive mistake-prevention system in general, primarily engages "Voice," "Privacy," "Safety," and "Dignity."

70. A false positive mistake prevention system that keeps content on the platform that does not violate Meta's policies contributes to Facebook and Instagram as places for expression. Conversely, to the extent that a false positive mistake-prevention system keeps violating and harmful content on the platform and facilitates its reach, it may negatively impact the "Voice," "Safety," "Privacy," and "Dignity" of others. To the extent that the system privileges the speech of some relative to others by delaying and decreasing the probability of enforcement, this unequal treatment implicates Meta's value of "Dignity," which relates to the expectation that Meta will treat all users fairly. The company should ensure that its systems are structured to consider the full range of Meta values.

V. Assessment of cross-check system

71. At the time of the Board's briefings with Meta, it performed about 100 million enforcement attempts on content every day. At this volume, even if Meta were able to make content moderation decisions with 99% accuracy, it would still make one million mistakes every day. Meta's content moderation mistakes include over-enforcement and under-enforcement, meaning that Meta both removes non-violating content and fails to remove violating content.

72. In this respect, Meta's use of cross-check responds to broader challenges in moderating immense volumes of content. The Board agrees that within this challenging context, Meta needs mechanisms to address both false positives and false negatives. Meta has a responsibility to address these larger problems in ways that benefit all users and not just a select few, however. Any decisions related to delaying or exempting enforcement actions for either certain users or certain pieces of content should align with Meta's human rights responsibilities and its stated values. Cross-check, both as it previously operated and in its current form, fails to do that.

73. The Board notes that Meta has made improvements to this system, both before referring this request to the Board and during the time the Board has been assessing cross-check. However, several aspects of the cross-check system do not

align with Meta's responsibility to identify and mitigate negative human rights impacts or uphold the company's values. These include:

- A broad scope to serve multiple and contradictory objectives that enables visibility and virality for violating content.
- The unequal access to discretionary policies and enforcement.
- That program enrollment may exceed capacity.
- The failure to track core metrics to assess the program and make improvements.
- The lack of transparency and auditability about its functioning.

74. Despite significant public concern about the program, Meta has not effectively addressed problematic components of its system. In this section, the Board highlights several of these problems. In the sections that follow we make a series of recommendations to Meta to outline how a mistake-prevention system could better comply with the company's commitments.

    Broad scope to serve multiple and contradictory objectives that enables visibility for violating content

75. Meta told the Board that Early Response Secondary Review exists to "protect voice, [and] enhance transparency and community trust." Meta further called attention in its request to the Board to its inclusion of journalists and community leaders in cross-check. The company highlighted that cross-check ensures that voice is preserved in a variety of important scenarios:

- "Members of marginalized communities who re-share violating hate speech targeted at them in order to raise awareness about or condemn it, which have been mistakenly removed for violating our hate speech policies."
- "Journalists who report in conflict zones where designated organizations are active, whose content has been mistakenly removed for violating our Dangerous Organizations and Individuals policies."
- "Health-related nudity, such as post-mastectomy reconstruction or breastfeeding photographs, which have been mistakenly removed for violating our nudity policies."

76. In a meeting with the Board, when asked about negative impacts that might ensue without ERSR, Meta officials stated that one issue, for example, is that it could impede communication and the flow of information in a crisis such as a natural disaster or political upheaval. These points of emphasis in Meta's stated rationale for the system contrast strikingly with how the system operates.

77. The Board shares Meta's concern about wrongly removing non-violating content posted by people drawing attention to human rights violations, working to promote women's health, and other public interest reporting. In fact, the Board's decisions have addressed such mistakes. Meta identifies these cases as "enforcement errors" only after the Board brings those cases to the company's attention. Examples

include the *Wampum belt* decision (2021-012-FB-UA), concerning the incorrect removal of an Indigenous artist's expression countering hate after multiple erroneous human review decisions; the Board's *Mention of the Taliban in news reporting* decision (2022-005-FB-UA) about the incorrect removal of a news outlet's post reporting on a designated organization; and the *Breast cancer symptoms and nudity* decision (2020-004-IG-UA), concerning the incorrect automated removal of a post that should have benefited from the health-related exception to Meta's adult nudity policies.

78. While Meta focuses on at-risk voices posting non-violating content when describing the program, Meta also stated that the cross-check program serves a core business function, as it serves an "important role in managing Facebook's relationships with many of [its] business partners." Relatedly, the cross-check tag sensitivity framework, which underpins both the "entity sensitivity" factor for GSR ranking and ERSR tags, is directly linked, among other factors, to the degree of reputational and internal backlash that is anticipated if particular content is removed in error. For example, Meta characterizes the risk of "escalation at the highest levels (CEO, COO)" as corresponding to a cross-check "extremely high severity" tag. Correlating highest priority within cross-check to concerns about managing business relationships suggests that the consequences that Meta wishes to avoid are primarily business-related and not human rights-related.

79. In order to assess how Meta prioritizes entities within cross-check, the Board repeatedly requested that Meta share its Early Response Secondary Review list for the Board's analysis. Meta did not provide the Board with this list. The Board cannot fully assess the degree to which the company is meeting its human rights responsibilities under the program or the profile of the entities that are guaranteed enhanced review if it does not know how the program is being implemented and precisely who benefits from it. Meta argued that providing a list of users subject to cross-check would violate the company's legal obligations in relation to user privacy. Based on legal advice, the Board believes, and has pointed out to Meta, that these concerns could have been mitigated and more extensive disclosures provided.

80. Almost five months after the Board first requested this information, Meta provided the Board a list with limited aggregate data about each listed entity on the current Early Response Secondary Review list. Specifically, Meta only disclosed the type of entity (e.g., Instagram user, Facebook page), their associated country and language, as self-selected by the entity, and whether or not Meta considers the entity "civic" and a "partner." Not all information was provided for each category of entity. For example, a quarter of the listed Instagram entities did not select a specific country or language in their profile settings and are also not considered either a civic actor or Meta partner.[50] This means that, for these entities, Meta only disclosed the existence, but not the identities or characteristics, of a group of Instagram users benefitting from cross-check.

81. This limited disclosure impairs the Board's ability to carry out its mandated oversight responsibilities. Meta's description of the "civic" category, for example, includes state actors, elected officials, "civic influencers" and candidates to public office, among others. Similarly, the "partner" category spans news organizations, celebrities, artists and more. The Board cannot evaluate, for example, the degree to which journalists, rights defenders and dissidents in particular countries are granted the same protection for their expression as the state actors who are enrolled in ERSR under program policy.

82. Meta has told the Board that it has no comprehensive system in place to systematically assess which journalists, human rights defenders or civil society figures in a particular geography should be subject to ERSR. Inclusion of such users on the list is based on decentralized decisions by Meta staff described by the company as "internal experts with high market knowledge." This raises the risk that significant gaps and inconsistencies exist in terms of who is afforded the added layers of protection for expression that cross-check ERSR provides.

83. Journalists posting content from conflict contexts, political opposition seeking elected office, celebrities posting a wide range of content, and business partners posting content to sell goods pose fundamentally different risk profiles from a free expression and human rights perspective. Given Meta's problems moderating content at scale, within current limitations user-generated content should be subject to different rights-oriented prioritization. Meta has described a system that does not include strategies or tactics to ensure that the individuals and the expression most needing protection receive it in the near term, with the ultimate goal of providing better content moderation for all.

84. Under ERSR, should content from any entitled entity be identified as violating and flagged for additional review, such content, regardless of risk profile, remains on platform during its period of peak virality in the aftermath of immediate posting. This is significant for two reasons. First, viral content spreads quickly on and across platforms. Second, once something is posted by an entity that has a large reach, the content will inevitably be recorded and reshared individually by users even if the original post is deleted. This means that accounts that benefit from the ERSR cross-check system may upload violating content and know that it can attain far reach even if it is violating.

85. Although the Board notes that Meta stated it has a system to prioritize high severity ERSR content for review, this content still remains on the platform until all necessary reviews are completed, sometimes for significant periods. For example, in the Neymar case, it is difficult to understand how non-consensual intimate imagery posted on an account with more than 100 million followers would not have risen to the front of the queue for rapid, high-level review if any system of prioritization had been in place. Given the serious nature of the policy violation and the impact on the victim, this case highlights the need for Meta to adopt different approaches for content pending review and shorten review timelines.

86. Delayed enforcement of violating content is a significant source of harm under the cross-check program. According to Meta's own research, user views of violating content because of cross-check are due to "incorrect overturns, and the delay of enforcement of non-overturns for which enforcement is slowed due to the secondary review process." The company acknowledges that affording additional protection to some privileged users' content may confront other users with violating content such as hateful speech or harassing posts.

87. Content automatically granted ERSR is different from content identified and sent for GSR. On the one hand, as noted above, the percentage of ERSR content ultimately found violating seems to vary. During time periods when the overturn rate is low, a key flaw in the system is its failure to ensure the prompt removal of violating content.

88. On the other hand, most GSR content is consistently ultimately found non-violating. For this system, the overturn rate seems to reveal that there are greater over-enforcement issues at scale, and that secondary review is mostly permitting non-violating content to remain accessible. The Board thus notes that to the extent that GSR preserves more expression, its impact is limited by the capacity constraints Meta imposes.

89. In sum, the Board finds that while Meta characterizes cross-check as a program to protect vulnerable and important voices, it appears to be more directly structured and calibrated to satisfy business concerns. While the Board understands that Meta is a business and should be able to design policies that meet business concerns, these same policies should not be characterized as serving as human rights risk mitigation measures if they do not meet that objective. Additionally, if Meta's business design choices negatively impact human rights, it should identify and then prevent, mitigate, or cease those negative impacts through program improvements.

Unequal access to discretionary policies and enforcement

90. Cross-check is designed to subject some content to more nuanced moderation decisions, determining whether any exception or specialized policy might apply to decline enforcement. According to Meta, "if content that was cross-checked is escalated for further review, it may then be subject to a decision based on [...] context-specific policies." Cross-check enables human review by the "Early Response Team" which the Board believes can grant exceptions in enforcement, both relating to the specific content and penalties against the entity itself. Content reviewed through ERSR is guaranteed to reach this team before possible removal, and content reviewed through GSR has a higher chance of reaching this team.

91. Meta has repeatedly told the Board and the public that the same set of policies apply to all users. Such statements and the public-facing content policies are

misleading, as only a small subset of content reaches a reviewer empowered to apply the full set of policies.

92. Entitlement to Early Response Secondary Review therefore provides a significant benefit to the user. It means that more of the content they choose to post is more likely to remain on the platform. In the case of non-violating content, it is protected from mistaken removal. In the case of violating content, it is allowed to remain on the platform during peak viewership before a later removal.

93. The Board also believes that, in addition to applying content policies with more discretion, content reviewed on escalation may benefit from decisions to not apply account restrictions that would be applied under normal procedures. In general, content policy violations correspond to "strikes" against an account, which in turn correspond to specific consequences. According to Meta's Transparency Center, strikes lead to increasingly long periods of time where accounts cannot post content. For serious or repeated strikes, Meta will disable an account.

94. The Board inquired about the discretionary application of policies and enforcement consequences. Meta responded that it does "not have statistically significant data distinguishing between penalties applied to cross-check versus non-cross-checked entities" and is "not aware of and [has] not located research or analysis" addressing these possible discrepancies. Given that cross-check may exempt users from account-level consequences, the Board is troubled that the company has either chosen not to track and analyze this information or has failed to disclose it to the Board.

95. According to public reporting in The Guardian, after Neymar posted violating content, he "was not subject to the normal Facebook procedure for someone who posts unauthorized nude photos, which is to have their account deleted." This example was revealed through whistleblower disclosures, and it is not clear how widespread such practices may be. The Board also asked Meta to confirm the account-level restrictions it applied in this case. The company ultimately disclosed that the only consequence was content removal, and that the normal penalty would have been account disabling. The Board notes that Meta later announced it signed an economic deal with Neymar for him to "stream games exclusively on Facebook Gaming and share video content to his more than 166 million Instagram fans."

96. Unequal access to escalated reviews as well as policy exceptions is particularly concerning given the lack of objective or transparent criteria for inclusion on Early Response Secondary Review lists. As noted above, it is not clear how Meta ensures that those most likely to be subject to over-enforcement or facing challenges to exercising their rights to freedom of expression are given this additional protection. The Board is concerned that those often most at risk, including journalists and human rights defenders, who may report on dangerous organizations or document graphic abuses, are those least likely to be proactively added to such lists given the investment that would be required to find those people across the globe.

97. On the other hand, Meta explained to the Board that it has a dedicated team charged with ensuring that all eligible entities representing government officials and organizations are enrolled in ERSR. Criteria to include "businesses, media organizations and creators" also seem clearer. According to Meta, one criterion, for example, is a specific amount of spending or revenue generated by an entity across Meta's "family of apps," although the amount may vary over time.

98. The Board is also concerned that in operating cross-check, Meta focuses disproportionate attention on more lucrative markets, instead of focusing on contexts with greater risks to human rights, including freedom of expression. At the time of the Board's briefing with Meta, 42% of content reviewed through the Early Response Secondary Review pathway originated from the United States or Canada. Similarly, 20% of all the entities on ERSR lists at that time correspond to those two countries. In contrast, according to Meta, just 9% of "monthly active people" on Facebook were from the United States and Canada. This data shows that users based in the United States and Canada have disproportionate access through Early Response Secondary Review to specialized review pathways that guarantee access to the full set of Meta policies, analysis of context, and likely the possibility of non-standard account penalties for violating content.

99. This disparity is correlated to the fact that "average revenue per person" in the US and Canada is the highest in the world, at around three times larger than in Europe and about 12 times larger than in Asia-Pacific. These facts highlight the financial incentives that shape how ERSR operates and reinforce concerns of equity. Through the design of cross-check, users in lucrative markets with heightened risk of public relations implications for Meta enjoy greater entitlement to protection for their content and expression than those elsewhere.

100. In addition, for GSR, the cross-check ranker prioritizes content according to factors like "topic sensitivity" that potentially require automated assessment of the language of the content. The Board is concerned that Meta does not prioritize training its automated processes on less-spoken languages and less lucrative markets. Limited investment in moderation in these languages limits the ability of algorithms to identify topics in such content. This suggests that users in these markets, including the Global South, may be disadvantaged when assessed for GSR cross-check eligibility. Similarly, Meta disclosed that "a group of languages are reviewed by non-native speakers using our translation and slurs highlighting tools." This reinforces the Board's concern that cross-check does not benefit all users equally, even through GSR.

Program enrollment exceeds capacity

101. ERSR and GSR eligibility persistently exceeds the human review capacity Meta allocates to the cross-check program. The mismatch between the volume of content that is designated for enhanced review through these systems and the

inadequate human resources allocated to the task represents a critical flaw in the system.

102. Meta told the Board that it "never intended to operate with a consistent backlog of cases, though operational capacity constraints and increasing volumes have led to a backlog in Early Response Secondary Review. [...T]hat backlog consists of content we have assessed as likely being low severity." Notwithstanding Meta's statement that it did not intend to maintain a continuous backlog, the company has failed to assign sufficient human resources to meet the content moderation needs of these programs. Moreover, as noted above, not all content subject to delayed enforcement is low severity.

103. Limited human review capacity has different but related consequences for Early Response Secondary Review and General Secondary Review. For ERSR, capacity shortfalls mean that content will remain on the platform during the time period it is most likely to accrue views. As this content remains on the platform until it receives enhanced review, content posted by high-profile ERSR users that violates Meta policies remains on the platform during its period of highest viewership. While Meta may attempt to review content that may cause greater harm first, it is not clear that it does so consistently, and this still reflects a design decision to provide automatic protection to entities selected based largely on commercial criteria.

104. For General Secondary Review, limited capacity may lead to two consequences. First, the Early Response Team may fill its time with ERSR content, as that content must be reviewed to apply any enforcement action. The Early Response Team therefore often does not have availability to review GSR content, and GSR content does not reach this critical level of review where policies requiring additional context and discretion may be applied. Second, limited capacity at the Market Team review level means that more GSR content times out of the queue before review and is removed by default. As the majority of this content appears to consistently be non-violating, this means that a key consequence of limited capacity for General Secondary Review content is that Meta removes more content that is likely non-violating.

105. These flaws compound the disparities in treatment of different users on the platform. Privileged users enrolled in ERSR have more chances to be reviewed by a moderator who may apply context to uphold their content, have a greater range of policy exceptions that can apply to uphold their content, and benefit from a system where even violating content is guaranteed viewership for some period of time. Ordinary users whose content might have access to GSR review, by contrast, have more limited opportunities for review of their content, are more likely to be subject to content policies without contextual review or the applications of policy exceptions, and, as content times out, are more likely to have non-violating content removed. This system has serious implications for the values of "Voice," "Dignity," "Privacy," and "Safety" that Meta claims to pursue.

Failure to track core metrics to assess the program and make improvements

106. The Board evaluated the metrics Meta uses to justify and evaluate the cross-check program. The metrics that Meta currently uses do not capture all key concerns and do not seem to have provoked changes when shortcomings were identified. Additionally, Meta is failing to monitor and set goals on a broad enough set of metrics to give a full picture of how the program operates and establish targets for improvement accordingly.

107. As discussed above, one metric Meta calculates is the overturn rate, or percentage of content that is subject to cross-check and ultimately found non-violating, despite the initial identification as violating by automation or human review. In Meta's words, "overturn rate is the efficacy rate of the cross-check system." According to information it provided the Board, Meta "want[s] [the overturn] percentage to be high. If none of the decisions were overturned through cross-check, that would mean [it was] cross-checking the wrong content."

108. Even though Meta has stated that the overturn rate should be high, Meta continues to provide the greatest protections to Early Response Secondary Review users. According to the figures Meta provided the Board, this rate varies significantly. Providing this protection to content without a consistently high overturn rate suggests Meta may be, according to its own goals, cross-checking the wrong content.

109. Content moderation in large volumes is marked by both over and under-enforcement. Meta focuses on the prevalence of violating content as its main public metric to assess how effective its moderation efforts are at removing harmful content. This includes content that flows through the cross-check system. Meta calculates prevalence by estimating the percentage of all views of content on Facebook or Instagram that were views of violating content. The use of prevalence as its general success metric may encourage Meta to further automate the removal of content and limit context-based enforcement to ensure low prevalence across the platform, without proper mechanisms to prevent wrongful deletions of content at scale. The consistently high overturn rate on GSR, for example, supports that inference.

110. The Board notes that Meta did not provide the Board with information showing that it tracks data about the accuracy of decisions made through its cross-check system. This means that even though the program is supposed to ensure accurate content moderation decisions, it does not appear that Meta is tracking whether its decisions via cross-check pathways are more or less accurate than decisions made through its normal scaled quality control mechanisms. Accuracy data would be a key indicator of the possible influence of non-content policy concerns on moderation decisions made in cross-check. By measuring success based on overturn rate only, Meta is not considering whether the ultimate decisions are the correct decisions.

111. Additionally, Meta has stated that the Regional Market Teams and the Early Response Team are specialized, having a particular set of skills, training and access to internal tools that allows them to make cross-check-level moderation decisions. However, as described above, at certain points in both the ERSR and GSR pathways, decisions may be made by contracted reviewers. These reviewers do not have the same access or training as Meta employees. If the goal of the cross-check program is to produce the most policy accurate decisions for entitled entities and important content, then measuring accuracy of cross-check decisions in general, but across reviewer types in particular, should be a basic tenet to understand if the operational design is working as intended.

112. Additionally, as an objective of cross-check is to protect important content most at risk of over-enforcement, Meta should focus on additional methods to identify such content. Meta disclosed that while it is actively working on understanding and mitigating over and under-enforcement for specific populations and problem areas, it still "needs to centrally define which populations are under/over-enforced. Pending such an effort we do not have a good way of creating a before the fact definition."

Lack of transparency and auditability of the program and its functioning

113. Lastly, the Board is concerned about the limited information Meta has provided the public and its users about this program. This policy advisory opinion resulted from Meta's failure to disclose to the Board key information about this program in the context of its deliberation on a case about a prominent user subject to cross-check.

114. Currently, Meta does not inform users that they are subject to ERSR, the entity-based mechanism in cross-check. It also does not inform users when they report content posted by a cross-checked entity. The company also provides limited transparency on the complex secondary review processes cross-checked content benefits from.

115. In addition, Meta does not share publicly its procedures for ERSR list creation and its auditing framework. The Board does not know, for example, whether entities that continuously post violating content are kept on Early Response Secondary Review lists based on their profile. Meta has given no indication that violation history or frequency is a factor in creating or maintaining Early Response Secondary Review lists. The lack of transparency regarding auditing impedes the Board and the public from understanding the full consequences of the cross-check system.

Conclusions about cross check

116. The Board acknowledges that a mistake-prevention system could be a useful safeguard against the improper removal of important content. However, if cross-check does not target such expression and permits severely violating content to remain on the platform, the program creates negative human rights impacts that

Meta is not monitoring or mitigating. The Board thus concludes that cross-check is currently neither designed nor implemented in a manner that meets Meta's human rights responsibilities and company values.

117. In its case decisions, the Board looks to the three-part test in Article 19 of the ICCPR, evaluating whether restrictions on expression meet requirements of legality, legitimate aim, and necessity and proportionality.

118. Legality refers to whether the rules are clearly and accessibly communicated. The existence, purpose and nature of the system is opaque in ways that cannot be justified given the significant effects cross-check has on the exercise of fundamental rights. Content policies presented as globally applicable that can only be applied with additional context at escalation, including through cross-check, are misleading.

119. Legitimate aim refers to whether restrictions are targeted at objectives specified in Article 19, including to respect the rights of others and protect national security, public order, and public health. The metrics through which the company measures the effectiveness of its enforcement systems suggest that its motivations are substantially focused on business reasons.

120. Necessity and proportionality refer to whether restrictions on expression are the least intrusive way to meet the legitimate aim. Here, the Board reiterates its concerns about inequitable access to the benefits of cross-check. Meta maintains clear processes to determine some of its users are entitled entities, such as state actors and business partners. Without clear criteria for other users who are likely to post content with significant human rights value, the program less clearly benefits others, including members of marginalized and discriminated-against groups. Meta is also not collecting and monitoring information about whether this program produces more accurate results in practice. Lastly, through cross-check, Meta defaults to leave content identified as violating on its platforms. As a policy matter, Meta is setting aside what it has determined is a proportionate response at scale for some content, often based only on economic or public relations concerns.

121. To comply with Meta's human rights responsibilities and company values, a system to prevent over-enforcement should be structured substantially differently than it is at present.

VI. Enforcement recommendations

122. In response to the questions posed by Meta, here the Board provides recommendations on entity-based mistake-prevention systems and dynamically determined content-based mistake-prevention systems. Meta has a responsibility to address its content moderation challenges in ways that benefit all users and not just a select few. However, given the focus of this policy advisory opinion, the Board

focuses here on limited-scope mistake-prevention systems.

Entity-based mistake-prevention system governance recommendations

123. Any system based on entity eligibility, such as Early Response Secondary Review, should be carefully designed, subject to oversight, and continuously monitored. This should ensure that it meets its stated purposes and evaluates externalities and unintended consequences it may cause. Such a system should protect users who are likely to post expression that is particularly important from a human rights perspective.

124. It is critical that Meta be clear about its objectives and tailor its systems narrowly to meet those objectives. It should also avoid providing protection for expression that violates its content policies or human rights commitments. Additionally, given that certain users may benefit from additional protections and avenues for expression, the company should provide the public with robust information about these processes so that they can adequately evaluate the information and opinions they see on the platform.

Users that should be included in entity-based mistake prevention systems

125. Meta states that the categories of inclusion for its entity-based mistake-prevention system cover "civic and government," "significant world events," "media," "historically over enforced," and "marginalized communities," "businesses," "creators," "entities escalated for review," and "legal and regulatory."

126. These broad categories require additional sorting and specification. In light of Meta's human rights commitments and stated values, if the company opts to operate an entity-based false-positive prevention system, there are certain categories of users that *should* be provided such protection, users that *may* be provided this protection, and users that *should not* be provided such protections given the human rights risks they pose.

127. First, entities that *should* be included are those who are likely to produce expression that it is important from a human rights perspective, including on matters of public importance. This benefits not only those users, but those who wish to access the information they share.

128. These users should include, for example, people whose content runs a high risk of over-enforcement, journalists and media organizations, public officials and candidates for office, and other civic actors including human right defenders and advocates for marginalized communities. In this respect, the Board views a list-based system as a proxy for providing additional protections to critical expression, and not protection based simply on the identity of the speaker. The Board recognizes that Meta maintains various lists of entities to which it affords greater protection, including its journalists' registry and its roster of "trusted partners" from

civil society. These existing, vetted entities could form one source from which the company might build an objective, global, human rights standards-based system accessible to all those whose expression satisfies the criteria for inclusion.

129. Second, entities that *may be* included may be based on company priorities and may include users with commercial importance and business partners. This might include advertisers, businesses with pages or groups that are at risk of over-enforcement, users who pose a special reputational risk to the company, or other users with a commercial relationship with Meta.

130. Third, there are entities that *should not* be included in any entity-based mistake prevention system that delays all enforcement. These include entities and users that repeatedly create or share content that violates Meta policies or terms of service. Meta's current account-level enforcement system, based on strikes and penalties, could be leveraged for the purposes of implementing this rule. Should users included due to commercial importance frequently post violating content, they should not continue to benefit from a system that delays enforcement. Meta has a responsibility to identify such users and exclude them from systems that provide their violating content additional visibility. While the number of followers could be a legitimate proxy for the degree of public interest in user's expression, a user's celebrity or follower count should not be the sole criterion for an entity-based mistake prevention system.

131. Meta's inclusion of all entities in the same system places them in direct competition for limited review resources. Meta should prioritize adequately resourcing mistake-prevention systems that mitigate human rights harms. In this context, Meta should ensure that content with human rights or public interest implications is reviewed in a timely fashion by skilled reviewers with the ability to take further context into consideration, regardless of whether the content came from entity-based or content-based pathways.

132. The Board recommends that Meta take steps to either use separate pathways or create prioritization mechanisms to differentiate between users that should be included due to Meta's human rights responsibilities and users that are included due to commercial priorities, given their different risk profiles. Businesses, for example, might be more likely to have content identified as violating spam rules, as they may rapidly post commercial content. Users with a large follower count may post on important matters of public interest but may similarly post violating content.

Decision makers should be qualified and empowered to make rights-respecting decisions

133. Consistent with the Board's recommendations throughout, Meta should prioritize its mistake-prevention secondary review workflows according to risk profile and human rights value.

134. The content posted by entities that Meta *should* include based on human rights concerns should be reviewed by teams with context and language expertise. This review pathway, including its escalation paths, should be devoid of business considerations. Meta should take steps to ensure this team does not report to public policy or government relations teams or those in charge of relationship management with any affected users.

135. The path dedicated to resolving issues explicitly related to Meta's business priorities could address, for instance, ads enforcement, spam rules, feature limits and behavioral issues. An example of behavioral issues is a business page being wrongly penalized for uploading pictures at a much faster rate than a normal profile. Either through lower prioritization or separation into a different workflow, these reviews should not displace resources targeted at human rights mitigation.

136. The Board notes that the Early Response Team, which is permitted to apply policy exceptions and interpret context, does not require its reviewers to have cultural or linguistic expertise. According to Meta, it makes decisions based on notes provided by Regional Markets Teams. Meta itself acknowledged that "relying on translations is imperfect." In this context, the Board urges Meta to ensure cultural and linguistic expertise at these levels of review. Meta should consider incorporating employees with cultural and linguistic expertise from at-risk regions into these teams and developing procedures to include staff with such expertise in decision making.

Guidance to create and govern lists for entity-based mistake-prevention systems

137. Meta should establish clear and public criteria for entity-based mistake-prevention eligibility. These criteria should differentiate between users whose expression merits additional protection from a human rights perspective, including information in the public interest, and users included for business reasons. For example, Meta currently defines one cross-check category as "Media Organizations, Businesses, Communities and Creators." This category includes "health organizations, news publishers, entertainers, musicians, artists, creators, and charitable organizations." Criteria this broad are insufficient. Meta should also develop criteria based on patterns of violating or undesirable behavior on the platform to avoid granting protections to harmful users.

138. Meta should add entities to mistake-prevention systems only once the process is objective, well-governed, and transparent. All entities that are proposed to be added to a list should be made aware of the possibility and should be given the option to decline inclusion if they so desire. Those who choose to be included should be required to review Meta's content rules and re-commit themselves to following them. While the Board views cross-check as providing benefits to included users, Meta should operate based on principles of user consent.

139. Clear public criteria should also provide a basis for users who qualify to proactively seek inclusion on such lists. Meta should establish a process whereby users can

apply for over-enforcement mistake-prevention protections should they meet the company's articulated criteria. State actors should be eligible to be added or apply based on these criteria and terms but given no other preference.

140. In addition to meeting public criteria, the process for inclusion, regardless of whether a user or Meta initiates the process, should involve: (1) a requirement to review Meta's content policy and an additional, explicit, commitment to follow them; (2) an acknowledgement of the program's particular rules; and (3) a system to inform users proactively of changes to Meta's content policies to facilitate awareness and compliance.

141. Meta sometimes works with civil society through its 'trusted partner' program and other stakeholder engagement initiatives to gather information about entities that should be considered for protection. The Board recommends that Meta strengthen its engagement with civil society for the purposes of list creation. Users should be able to nominate others that meet the public criteria, as long as the nominees may decline inclusion. This is particularly urgent in countries where the company's limited presence does not allow it to identify candidates for inclusion independently.

142. List creation, and particularly this engagement, should be run by specialized teams, independent from teams whose mandates may pose conflicts of interest, such as Meta's public policy teams. To ensure criteria are being met, specialized staff, with the benefit of local input, should ensure objective application of inclusion criteria. Public policy teams often interact with and lobby government actors, creating unavoidable conflicting incentives. While they may nominate candidates, they should not be decision makers.

143. Meta told the Board that currently a single company employee may decide to add entities to a particular cross-check list, and there is no required review of those decisions. Going forward, the company should have an established process for objective, criteria-based review of all entities that will receive additional benefits. At least two people on different teams should be involved to finalize inclusion on any list-based protection, and individuals with personal or business relationships with nominated entities should not be decision makers.

Guidance to maintain and audit lists for entity-based mistake prevention systems

144. In addition to establishing clear criteria for entry to a mistake-prevention protection program, Meta should establish clear criteria and processes for audit and removal. Should entities no longer meet eligibility criteria, they should be removed.

145. Meta told the Board that its new proposed governance structure includes rules to add and remove entities from the lists; tag expiration rules; periodic audit procedures; and an oversight structure. Meta also disclosed, however, that there were exceptions to some of these rules, such as "Civic and Government" entities not

having default expiration periods. Meta also shared that it is currently auditing a limited subset of entities on Early Response Secondary Review as it moves towards a more simplified list structure.

146. The Board recommends that Meta require at least yearly review of all included entities in any mistake-prevention system that provides benefits to such entities. There should also be clear protocols to shorten that period where warranted. Similar to its recommendations on initial inclusion in any list-based system, the Board recommends that at least two people with separate reporting structures participate in internal audits.

147. Meta should also ensure clear removal criteria for any list-based protection program. One criterion should be the amount of violating content posted by the entity. These could, for example, be based on a "three-strikes" policy, unless Meta has established a harsher penalty for the violation(s) at issue (e.g., Non-Consensual Intimate Imagery account removal). Such a system should give entities warnings and then remove them from cross-check when they accrue their final strike, regardless of whether the violation merits removal from the platform as a whole. Entities should be able to appeal removal and reapply in the future.

148. Lastly, the Board emphasizes that, while internal audit procedures are a step in the right direction, internal auditing without external oversight falls short. External audits, by the Board or another third party (e.g., researchers or civil society), are required in order to assess whether a mistake-prevention system mitigates negative human rights impacts. While the Board acknowledges serious privacy and safety concerns with external auditing, the Board believes that Meta can take mitigating steps to anonymize and aggregate data to address these concerns.

Some entities receiving additional protection should be publicly marked

149. The Board has repeatedly called on Meta to inform users and the public about its policies and practices. Any entity-based mistake-prevention system should provide all users on the platform with clarity about how Meta applies its rules. Currently, users do not know if they are enrolled in ERSR. Additionally, users viewing and reporting content posted by users enrolled in ERSR are not informed that the content may be subject to special review procedures.

150. The Board recommends that some categories of entities protected by the system should have their accounts publicly marked. These categories include all state actors and political candidates, all business partners, all media entities, and all other public figures included because of the commercial benefit to the company in avoiding false positives. This will allow the public to hold privileged users accountable for whether protected entities are upholding their commitment to follow the rules and hold Meta accountable for adhering to the publicly announced parameters of the program.

151. The Board identifies several risks in publicly identifying users who are enrolled in a false positive mistake-prevention program. First, there could be additional adversarial risk from users attempting to gain control of accounts with special protections, knowing that violating content will remain on the platform for a period of time. Second, some categories of users may face harassment or other attacks if they are perceived as maintaining a relationship or receiving special protection from the company.

152. However, the Board finds these risks can be mitigated, and the benefits outweigh these potential harms. First, Meta should invest any necessary resources to enhance account protection for users subject to a mistake-prevention system. Meta has experience providing extra layers of protection for journalists and other categories of users. Such procedures could be adapted for use in any future entity-based mistake-prevention system. While adversarial risk is real, it is not insurmountable in this context. Although the Early Response Secondary Review list is currently not public, many users already assume that high-profile accounts are included in the cross-check program.

153. Meta should not identify beneficiaries who are human rights defenders, entities included because they are subject to historical over-enforcement, and those included because they are at risk of harm, although they should be able to opt-in to identification. Clear criteria for inclusion and separation of the program for different objectives will facilitate this process.

154. Lastly, when users report content posted by an entity publicly identified as benefiting from additional review, reporting language should make it explicit that special procedures will apply, explaining the steps and potentially longer time for resolution.

Content-based mistake-prevention system governance recommendations

155. While an entity-based system should include users who are likely to produce expression that merits additional protection from a human rights perspective and users who may be at particularly high-risk for erroneous over-enforcement, a content-based system seeks to protect such content directly without regard to who posted it.

Content that should be selected and prioritized for content-based mistake prevention systems

156. According to Meta, its General Response System "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)."

157. The most heavily weighted factors for the ranking algorithm are topic sensitivity and entity sensitivity. As discussed above, entity sensitivity is, among other factors, directly related to the degree of internal escalation a mistake would cause. In this respect, Meta's cross-check ranker also prioritizes content that might cause economic or reputational damage, an objective already served by ERSR. Although GSR may have been designed to respond to some criticisms of the former exclusively entity-based cross-check system, this suggests that the company continues to prioritize expression based on the speaker and not the importance of the expression.

158. The Board agrees that universal eligibility for a false-positive mistake-prevention system is a positive step. However, such a system should prioritize identifying content that is not also targeted by an entity-based system. It should provide enhanced protection based upon a human rights rationale. While Meta may give some additional protection where over-enforcement might threaten its business interests, similarly to list-based systems, it should not do so at the expense of its human rights commitments.

159. An algorithmic ranker for a false positive prevention system could, for example, prioritize content based on the types of decisions that are hard for automation and human moderators at scale (e.g., historically over-enforced speech or speech by marginalized communities). In conjunction, the algorithm could prioritize the review order of this content based on the severity of the possible violation, the likelihood of being a false positive, and the likelihood of virality.

160. The Board recommends that to increase the impact of a content-based false positive prevention system, Meta should consider reserving a minimum amount of review capacity by teams that can apply all content policies (e.g., the Early Response Team). It further should analyze the content receiving additional review for insights as to where Meta's systems are resulting in the most high-impact mistakes, and prioritize review resources accordingly.

Technical corrections

161. Meta explained that "technical corrections" completely prohibit any enforcement for a specific policy violation on a particular entity. There may be business reasons to provide such protection to an extremely select set of entities, but any such system has the potential to create great risk of exempting entities that post violating content from content moderation enforcement. If such a system is used, it should be subject to the highest level of internal and external scrutiny. "Technical corrections" exempt certain entities from certain enforcement and are rightly understood as an "allowlist" or "whitelist," however limited its scope may be.

162. All recommendations regarding list-based programs, such as clear and firm criteria, cross-team review processes to grant any exemption, and audit processes to maintain exemptions apply here. Additionally, exemption should be prohibited for

content that Meta classifies as a high-severity violation. Meta should conduct periodic audits for all enforcement actions that are blocked by such exemptions. If, as Meta states, this is about a thousand actions per day, it should have the capacity to do so. This audit, with information on the scope and accuracy of the program, should be included in Meta's quarterly transparency reports.

163. Finally, the company should proactively and periodically search for unexpected or unintentional exemptions that may linger from previous iterations of this program. In its decisions, the Board repeatedly notes cases where Meta has inadvertently failed to update or maintain systems, and the consequences of such gaps in governance on an exemption system could be critical.

General mistake-prevention system governance recommendations

164. Beyond the broad governance changes to how a list-based and content-based mistake-prevention system should be established and audited, the Board also recommends that the procedures within such a system focus on harm mitigation and be subject to continuous monitoring for learning and improvement.

Harm mitigation following identification of violating content

165. As the company itself recognizes, a core cause of harm in Meta's false-positive mistake-prevention system results from delayed enforcement of violating content during the time period in which it is most likely to be viewed. As stated above, Meta itself identifies that the biggest drivers of users seeing violating content from cross-checked users or content on its platforms are "incorrect overturns, and the delay of enforcement of non-overturns for which enforcement is slowed due to the secondary review process." The Board urges Meta to take steps to mitigate those harms.

166. First, Meta should take measures to ensure that additional review decisions are taken as quickly as possible. Investments and structural changes should be made to expand the review teams so that reviewers are available and working in relevant time zones whenever content is flagged for any enhanced human review.

167. Second, the Board recommends that Meta use methods aside from defaulting to no enforcement action for pieces of content subject to enhanced review. This could include using the least intrusive means, for example, downranking, slowing the virality, hiding, or temporarily removing the content. Establishing different prioritization or pathways for content and entities of different nature should facilitate Meta applying different consequences to different types of content.

168. Content identified as violating during Meta's first assessment that is high severity, for example according to Meta's framework, should be removed or hidden pending review, and not permitted to remain on the platform accruing views simply because the posting user is a business partner or celebrity. The difference between

enforcement options, such as removal, hiding, and downranking, should be based on violation severity. Meta's framework, in theory, is designed to account for the likelihood of near-term harm, and whether the content has been identified as particularly likely to be an enforcement error. If content is hidden on these grounds, a notice indicating that it is pending review should be provided to users in its place.

169. Third, Meta should not operate these programs at a backlog. Maintaining a content queue for review that exceeds capacity means that content that may be violating will remain on the platform for an extended period of time. Delaying any enforcement while taking weeks to reach a decision results in functionally exempting entitled entities from the rules.

170. Meta should invest the resources necessary to match its review capacity to the content it identifies as requiring additional layers of review. This does not, however, mean that it should have the algorithm select less content. The consequence of Meta failing to build sufficient review capacity should not be delaying content from enforcement or outright mistaken deletions made by at-scale systems or reviewers. Meta has devised processes to prioritize review and ensure that its workforce has a continual stream of content to review. Given that GSR review currently results in a consistently high overturn rate, the Board believes that more content would benefit from this review.

171. Fourth, Meta should not automatically prioritize entity-based secondary review and make a large portion of the algorithmically selected content-based review dependent on extra review capacity.

Ensuring appeal availability

172. Meta informed the Board that it does not provide appeal or review opportunities consistently across all types of content. Appeals for content subject to the cross-check program seem to suffer from the same inconsistency.

173. The Board understands that providing appeals on content that has already reached the highest level of analysis within the company may be unnecessary, given that an appeal would replicate those pathways. However, the Board is concerned that some content may not be receiving appeal eligibility, despite not reaching those highest levels. The Board believes that Meta could have and should have provided more clarity on this point when asked repeatedly by the Board.

174. Additionally, the Board is particularly concerned about this confusion as it relates to appeals eligibility to bring cases to the Board, both for users to restore their own cross-checked content and to report the content of other users that benefit from cross-check. In fact, according to Meta, "for the months of May and June 2022, an average of 35% of the content in the cross-check system [...] could not be escalated to the Oversight Board." Users included on Early Response Secondary Review lists are among the users with the highest reach on the platform. This situation may be

depriving the Board of some of the most critical content moderation cases on Facebook and Instagram.

175. As a first step, Meta must provide clarity regarding appeals eligibility in general, and ensure that content that does not reach the highest level of review is able to be appealed internally. Second, Meta must guarantee that it is providing an opportunity to appeal to the Board for all content the Board is empowered to review under its governing documents, regardless of whether the content reached the highest levels of review within Meta.

Learning and improvement

176. To meet its human rights responsibilities, Meta should monitor its activities which impact rights on a periodic basis. The results of these reviews should guide Meta in making improvements to its policies and practices to minimize human rights harms. In this case, Meta maintains a variety of metrics related to the cross-check program that already show where the company should be making improvements. The Board believes that Meta should also provide the public with information about how this system is functioning, both to meet transparency responsibilities and to hold itself accountable for improvement.

177. First, Meta already maintains an overturn rate for its entity-based system (Early Response Secondary Review) and for its content-based system (General Secondary Review). Meta should use trends in overturn rates to inform whether to default to the original enforcement within a shorter time frame or what other enforcement action to apply pending review. If overturn rates are consistently low for particular subsets of policy violations or content in particular languages, for example, Meta should continually calibrate how quickly and how intrusive an enforcement measure it should apply.

178. Second, Meta told the Board that it has conducted post-mortem analysis exercises after Meta's "risk assessment" team has identified risk areas or there was an event that the company saw as a failure. The Board recommends that these and other reviews be conducted regularly on cross-check, based on internal risk assessments that pressure-test the system at the key points outlined in this policy advisory opinion.

179. Third, Meta disclosed that one of the categories it uses for Early Response Secondary Review is "historically over-enforced entities." This means that the company has already identified entities where Meta acknowledges it is unable to enforce its policies consistently and effectively. In addition to providing such entities access to over-enforcement mistake-prevention programs, Meta should use this data to inform how to improve its enforcement practices at scale. Meta should measure over-enforcement of these entities and it should use that data to help identify other over-enforced entities. Reducing that metric should be an explicit and high-priority goal for the company.

180. There are additional metrics Meta should develop and monitor to better align mistake-prevention strategies with human rights standards. For example, Meta should establish new metrics to quantify the impact of leaving violating content on the platform. In particular, the company should calculate the number of views a piece of content that is ultimately removed accumulates while pending review because of mistake-prevention mechanisms. Meta should determine a baseline for this metric and report on goals to reduce it.

181. Meta also disclosed that it also takes steps to address some issues related to under-enforcement. These include "classifiers to detect content that likely violates our policies; user reports that identify potentially violating content; human review sweeps where our teams review potentially violating content; High Risk Early Review Operations (HERO), a system where humans review content predicted to go viral; and reporter appeals, where users who report violating content may appeal [Meta's] decision."

182. The Board notes that efforts outside of at-scale automated enforcement and appeals have a narrow scope. Additionally, some of these initiatives compete for resources with cross-check. For example, HERO review is done by the market teams, which also must devote capacity to cross-check. HERO also only affords reviews to content expected to go viral. The Board agrees that high-reach content may cause more harm but believes it should be accompanied by efforts to improve moderation comprehensively. Meta should continue to invest in early detection and warning systems; and hiring and embedding people with local and language expertise in their trust and safety, content review operation, and mistake-prevention system list-creation efforts.

VII. Transparency recommendations

183. The Board has made a series of recommendations about how Meta should design and govern any false-positive mistake-prevention program. Meta's human rights responsibilities also mean it should provide transparency to the public about these programs. Transparency reports should contain comprehensive data so that users and the public understand how the program is functioning and what its consequences on public discourse may be. In addition to the described metrics, the Board recommends Meta include:

a. Overturn rates for false-positive mistake-prevention systems, disaggregated according to design choices and enforcement teams (e.g., Markets, Early Response, contractors, etc.) For example, the Board has recommended that Meta create separate streams for different categories of entities or content based on their expression and risk profile. The overturn rate should be reported for any entity-based and content-based systems, and categories of entities or content included.
b. The total number and percentage of escalation-only policies applied due to false-positive mistake prevention programs relative to total enforcement decisions.

c. Average and median time to final decision for content subject to false-positive mistake-prevention programs, disaggregated by country and language.
d. Aggregate data regarding any lists used for mistake-prevention programs, including the type of entity and region.
e. Rate of erroneous removals (false positives) on all reviewed content, including the total amount of harm generated by these false positives measured as the predicted total views on the content (i.e., over-enforcement).
f. Rate of erroneous keep-up decisions (false negatives) on content, including the total amount of harm generated by these false negatives, measured as the sum of views the content accrued (i.e., under-enforcement).

184. The Board has previously recommended that Meta disclose error rates in general, but also that it should "report on the relative error rates of determinations made through cross check compared with ordinary enforcement procedures." The Board believes that Meta's focus on prevalence, while useful in certain specific contexts, does not provide the right incentives to the company or the right tools for the public to understand how Meta's content moderation ecosystem is functioning.

185. Meta told the Board that it is "currently investing in an aggregate topline metric measurement that allows us to understand false positives through the entire system and are working to build this metric which we hope to share externally in our transparency reporting. This metric would be the counter metric to our false negative measurement that is currently reported through prevalence metrics." This is a step in the right direction and the Board urges Meta to complete this work as soon as possible.

186. In addition to the metrics emphasized in previous sections, which serve both to benchmark improvement and to provide information, Meta should further provide basic information in its Transparency Center regarding the functioning of any mistake-prevention system it uses that identifies entities or users for additional protections. The Board understands the potential for user adversarialism to attempt to bypass enforcement, and Meta may choose to summarize some points of its enforcement practices. The current level of transparency is inadequate and not justified by fear of adversarial risk.

187. More generally, the Board notes that providing greater transparency to external researchers, in particular access to data, is an essential component of oversight for mistake-prevention systems. Throughout the stakeholder engagement conducted for this analysis, the Board heard concern about Meta seeking to limit its current data access programs for external parties. Considering that systems like cross-check require making complex trade-offs, independent researchers could provide Meta with valuable insights on the impacts of its choices. The Board believes Meta should institute a pathway for external researchers to gain access to non-public data about the cross-check system that would allow them to understand the program more fully through public-interest investigations and provide their own recommendations for improvement. While mitigation measures to protect user

privacy must be taken, Meta could and should allow for greater understanding of how its platforms work.

VIII. Annex with recommendations and measures of implementation

The Board made multiple recommendations to Meta in its policy advisory opinion. This annex pairs those recommendations with measures of implementation to monitor Meta's progress. Meta should provide information about its implementation work in its quarterly reports on the Board. Additionally, Meta should convene a biannual meeting of high-level responsible officials to brief the Board on its work to implement the policy advisory opinion recommendations.

#	**Recommendation**	**Measures of implementation**
Entity-based mistake-prevention governance		
1	Meta should split, either by distinct pathways or prioritization, any list-based over-enforcement prevention program into separate systems: one to protect expression in line with Meta's human rights responsibilities, and one to protect expression that Meta views as a business priority that falls outside that category.	Meta provides the Board with information detailing how both inclusion and operation are split for these categories of entities. Meta publicizes the details about these systems in its Transparency Center. *Enforcement*
2	Meta should ensure that the review pathway and decision-making structure for content with human rights or public interest implications including its escalation paths, is devoid of business considerations. Meta should take steps to ensure that the team in charge of this system does not report to public policy or government relations teams or those in charge of relationship management with any affected users.	Meta provides the Board with information detailing the decision-making pathways and teams involved in the content moderation of content with human rights or public interest implications. *Enforcement*
3	Meta should improve how its workflow dedicated to meet Meta's human rights responsibilities incorporates context and language expertise on enhanced review, specifically at decision making levels.	Meta provides the Board with information detailing how it has improved upon its current process to include language and context expertise at the moment that context-based decisions and policy exceptions are being considered. *Enforcement*

4	Meta should establish clear and public criteria for list-based mistake-prevention eligibility. These criteria should differentiate between users who merit additional protection from a human rights perspective and those included for business reasons.	Meta releases a report or Transparency Center update detailing the criteria for list-based enhanced review eligibility for the different categories of users the program will enroll. *Transparency*
5	Meta should establish a process for users to apply for over-enforcement mistake-prevention protections should they meet the company's publicly articulated criteria. State actors should be eligible to be added or apply based on these criteria and terms but given no other preference.	Meta implements a publicly and easily accessible, transparent application system for any list-based over-enforcement protection, detailing what purposes the system serves and how the company assesses applications. Meta includes the number of entities that successfully enrolled in mistake-prevention through application, their country and category each year in its Transparency Center. *Enforcement*
6	Meta should ensure that the process for list-based inclusion, regardless of who initiated the process (the entity itself or Meta) involves, at minimum: (1) an additional, explicit, commitment by the user to follow Meta's content policies; (2) an acknowledgement of the program's particular rules; and (3) a system by which changes to the platform's content policies are proactively shared with them.	Meta provides the Board with the complete user experience for onboarding into any list-based system, including how users commit to content policy compliance and how they are notified of policy changes. *Enforcement*
7	Meta should strengthen its engagement with civil society for the purposes of list creation and nomination. Users and trusted civil society organizations should be able to nominate others that meet the criteria. This is particularly urgent in countries where the company's limited presence does not allow it to identify candidates for inclusion independently.	Meta provides information to the Board on how the company engages with civil society to determine list-based eligibility. Meta provides data in its Transparency Center, disaggregated by country, on how many entities are added as a result of civil society engagement as opposed to proactive selection by Meta. *Enforcement*
8	Meta should use specialized teams, independent from	Meta provides the Board with internal documents detailing which teams handle

	political or economic influence, including from Meta's public policy teams, to evaluate entities for list inclusion. To ensure criteria are met, specialized staff, with the benefit of local input, should ensure objective application of inclusion criteria.	list creation and where they sit in the organization. *Enforcement*
9	Meta should require that more than one employee be involved in the final process of adding new entities to any lists for false positive mistake-prevention systems. These people should work on different but related teams.	Meta provides the Board with information detailing the process by which new entities are added to lists, including how many employees must approve inclusion and what teams they belong to. *Enforcement*
10	Meta should establish clear criteria for removal. One criterion should be the amount of violating content posted by the entity. Disqualifications should be based on a transparent strike system, in which users are warned that continued violation may lead to removal from the system and or Meta's platforms. Users should have the opportunity to appeal such strikes through a fair and easily accessible process.	Meta provides the Board with information detailing the threshold of enforcement actions against entities at which their protection under a list-based program is revoked, including notifications sent to users when they receive strikes against their eligibility, when they are disqualified, and their options for appeal. It should also provide the Board with data about how many entities are removed each year for posting violating content. *Enforcement*
11	Meta should establish clear criteria and processes for audit. Should entities no longer meet the eligibility criteria, they should be promptly removed from the system. Meta should review all included entities in any mistake-prevention system at least yearly. There should also be clear protocols to shorten that period where warranted.	Meta provides the Board with data on the amount, type of entity, and reason for removal from entity lists as a result of audits, along with a timeline for conducting audits periodically. *Enforcement*
List transparency		
12	Meta should publicly mark the pages and accounts of entities receiving list-based protection in the following categories: all state actors and political candidates, all business partners, all media actors, and all other public figures included because of the	Meta marks all entities in these categories as benefiting from an entity-based mistake prevention program and announces the change in its Transparency Center. *Transparency*

	commercial benefit to the company in avoiding false positives. Other categories of users may opt to be identified.	
13	Meta should notify users who report content posted by an entity publicly identified as benefiting from additional review that special procedures will apply, explaining the steps and potentially longer time to resolution.	Meta provides the Board with the notifications for users that report content from users identified as benefiting from additional review and confirm global implementation and data that shows these notifications are consistently shown to users. *Enforcement*
14	Meta should notify all entities that it includes on lists to receive enhanced review and provide them with an opportunity to decline inclusion.	Meta provides the Board with (1) the notifications sent to users informing them of their inclusion in a list-based enhanced review program and offering them the option to decline; and Meta (2) publicly reports annual numbers in its Transparency Center on the amount of entities, per country, that declined inclusion. *Enforcement*
Enhanced review and prioritization		
15	Meta should consider reserving a minimum amount of review capacity by teams that can apply all content policies (e.g., the Early Response Team) to review content flagged through content-based mistake-prevention systems.	Meta provides the Board with documentation showing its process of consideration of this recommendation and the rationale for its decision on whether to implement it and publishes this justification to their Transparency Center. *Enforcement*
16	Meta should take measures to ensure that additional review decisions for mistake-prevention systems that delay enforcement are taken as quickly as possible. Investments and structural changes should be made to expand the review teams so that reviewers are available and working in relevant time zones whenever content is flagged for any enhanced human review.	Meta provides the Board with data that demonstrates a quarter-over-quarter reduction in time-to-decision for all content receiving enhanced review, disaggregated by category for inclusion and country. *Enforcement*
17	Meta should not delay all action on content identified as potentially severely violating and	Meta updates its Transparency Center with its new approach to enforcement action during the time when content

	should explore applying interstitials or removals pending any enhanced review. The difference between removal or hiding and downranking should be based on an assessment of harm, and may be based, for example, on the content policy that has possibly been violated. If content is hidden on these grounds, a notice indicating that it is pending review should be provided to users in its place.	receives enhanced review and provides the Board with information detailing the enforcement consequence it will apply based on content-specific criteria. Meta shares with the Board data on the application of these measures and their impact *Enforcement*
Resourcing		
18	Meta should not operate these programs at a backlog. Meta should not, however, achieve gains in relative review capacity by artificially raising the ranker threshold or having its algorithm select less content.	Meta provides the Board with data that demonstrates a quarter-over-quarter reduction in total backlogged content and amount of days with a backlog for cross-check review queues. *Enforcement*
19	Meta should not automatically prioritize entity-based secondary review and make a large portion of the algorithmically selected content-based review dependent on extra review capacity.	Meta provides the Board with internal documents detailing the distribution of review time and volume between entity-based and content-based systems. *Enforcement*
20	Meta should ensure that content that receives any kind of enhanced review because it is important from a human rights perspective, including content of public importance, is reviewed by teams that can apply exceptions and context.	Meta provides the Board with information that shows the percentage of content receiving review by teams that can apply exceptions and context because it has been posted by an entitled entity or because it has been identified algorithmically as meriting enhanced review disaggregated by mistake-prevention system (e.g., GSR vs. ERSR). *Enforcement*
Automatic bars to enforcement ('technical corrections')		

21	Meta should establish clear criteria for the application of any automatic bars to enforcement ('technical corrections'), and not permit such bars for high severity content policy violations. At least two teams with separate reporting structures should participate in granting technical corrections to provide for cross-team vetting.	Meta publishes the number of entities currently benefitting from a "technical correction" on an annual basis, with indication of what content policies are barred from enforcement. *Enforcement*
22	Meta should conduct periodic audits to ensure that entities benefitting from automatic bars to enforcement ('technical corrections') meet all criteria for inclusion. At least two teams with separate reporting structures should participate in these audits to provide for cross-team vetting.	Meta provides information to the Board on its periodic list auditing processes. *Enforcement*
23	Meta should conduct periodic multi-team audits to proactively and periodically search for unexpected or unintentional bars to enforcement that may result from system error.	Meta publishes information annually on any unexpected enforcement bars it has found, their impact, and the steps taken to remedy the root cause. *Enforcement*
Procedural fairness		
24	Meta should ensure that all content that does not reach the highest level of internal review is able to be appealed to Meta.	Meta publishes information on the number of content decisions made through enhanced review pathways that were not eligible for appeal. This yearly data, disaggregated by country, should be broken down in a way that explains what, if any, percentage of the content did not get an appeal because it reached global leadership review. *Enforcement*
25	Meta must guarantee that it is providing an opportunity to appeal to the Board for all content the Board is empowered to review under its governing documents, regardless of whether the content reached the highest levels of review within Meta.	Meta publicly confirms all content covered under the Board's governing documents are receiving Oversight Board appeal IDs to submit a complaint to the Board, providing documentation to demonstrate where steps have been taken to close appeal availability gaps. Meta creates an accessible channel for users to achieve

		prompt redress when they do not receive an Oversight Board appeal ID. *Enforcement*
Learning and improvement		
26	Meta should use the data it compiles to identify "historically over-enforced entities" to inform how to improve its enforcement practices at scale. Meta should measure over-enforcement of these entities and it should use that data to help identify other over-enforced entities. Reducing over-enforcement should be an explicit and high-priority goal for the company.	Meta provides data to the public that shows quarter-over-quarter declines in over-enforcement and documentation that shows that the analysis of content from "historically over-enforced" entities is being used to reduce over-enforcement rates more generally. *Enforcement*
27	Meta should use trends in overturn rates to inform whether to default to the original enforcement within a shorter time frame or what other enforcement action to apply pending review. If overturn rates are consistently low for particular subsets of policy violations or content in particular languages, for example, Meta should continually calibrate how quickly and how intrusive an enforcement measure it should apply.	Meta provides the Board with data detailing the rates at which queued content remains up or is taken down, broken out by country, policy area, and other relevant metrics, and describes changes made on an annual basis. *Enforcement*
Improving program accountability		
28	Meta should conduct periodic reviews of different aspects of its enhanced review system, including content with the longest time to resolution and high-profile violating content left on platform.	Meta publishes the results of reviews to the cross-check system on an annual basis, including summaries of changes made as a result of these reviews. *Transparency*
29	Meta should publicly report on metrics that quantify the adverse effects of delayed enforcement as a result of enhanced review systems, such as views accrued on content that was preserved on the platform as a result of mistake-prevention systems but	Meta includes one or more key metrics demonstrating the negative consequences of delayed enforcement pending enhanced review mechanisms in the Community Standards Enforcement Report, along with goals to reduce these metrics and progress in meeting those goals.

	was subsequently found violating. As part of its public reporting, Meta should determine a baseline for these metrics and report on goals to reduce them.	*Transparency*
30	Meta should publish regular transparency reporting focused specifically on delayed-enforcement false-positive prevention systems. Reports should contain data that permits users and the public to understand how these programs function and what their consequences on public discourse may be. At minimum, the Board recommends Meta include: a. Overturn rates for false-positive mistake-prevention systems, disaggregated according to different factors. For example, the Board has recommended that Meta create separate streams for different categories of entities or content based on their expression and risk profile. The overturn rate should be reported for any entity-based and content-based systems, and categories of entities or content included. b. The total number and percentage of escalation-only policies applied due to false-positive mistake-prevention programs relative to total enforcement decisions. c. Average and median time to final decision for content subject to false-positive mistake-prevention programs, disaggregated by country and language.	Meta releases annual transparency reporting including these metrics. *Transparency*

	d. Aggregate data regarding any lists used for mistake-prevention programs, including the type of entity and region. e. Rate of erroneous removals (false positives) versus all reviewed content, including the total amount of harm generated by these false positives measured as the predicted total views on the content (i.e., over-enforcement) f. Rate of erroneous keep-up decisions (false negatives) on content, including the total amount of harm generated by these false positives, measured as the sum of views the content accrued (i.e., under-enforcement)	
31	Meta should provide basic information in its Transparency Center regarding the functioning of any mistake-prevention system it uses that identifies entities or users for additional protections.	A section is added to the Transparency Center explaining its array of mistake-prevention systems (the Board understands the potential for user adversarialism to attempt to bypass enforcement, and Meta may choose to summarize some points of its enforcement practices). *Transparency*
32	Meta should institute a pathway for external researchers to gain access to non-public data about false-positive mistake-prevention programs that would allow them to understand the program more fully through public-interest investigations and provide their own recommendations for improvement. The Board understands that data privacy concerns should require stringent vetting and data aggregation.	Meta discloses a pathway for external researchers to obtain non-public data on false positive mistake-prevention programs. *Transparency*