Submission to the Facebook Oversight Board: Policy on Cross-checks

14 January 2022

Authored by: *[in alphabetical order] Anamika Kundu, Digvijay Singh, Divyansha Sehgal and Torsha Sarkar*

Edited by: Arindrajit Basu

Centre for Internet and Society

Whether a cross-check system is needed?

Recommendation for the Board: The Board should investigate the cross-check system as part of Meta's larger problems with algorithmically amplified speech, and how such speech gets moderated.

Explanation: The issues surrounding Meta's cross-check system are not an isolated phenomena, but rather a reflection of the problems of algorithmically amplified speech, as well the lack of transparency in the company's content moderation processes at large. At the outset, it must be stated that the majority of information on the cross-check system only became available after the media reports published by the Wall Street Journal. While these reports have been extensive in documenting various aspects of the system, there is no guarantee that the disclosures obtained by them provides the complete picture regarding the system. Further, given that Meta has been found to purposely mislead the Board and the public on how the cross-check system operates, it is worth investigating the incentives that necessitate the cross-check system in the first place.

Meta claims that the cross-check system works as a check for false positives: they "employ additional reviews for high-visibility content that may violate our policies." Essentially they want to make sure that content that stays up on the platform and reaches a large audience, is following their content guidelines. However, previous disclosures have proven policy executives have prioritized the company's 'business interests' over removing content that violates their policies; and have waited to act on known problematic content until significant external pressure was built up, including in India. In this context, the cross-check system seems less like a measure designed to protect users who might be exposed to problematic content, and more as a measure for managing public perception of the company.

Thus the Board should investigate both how content gains an audience on the platform, and how it gets moderated. Previous whistleblower disclosures have shown that the mechanics of algorithmically amplified speech, which prioritizes engagement and growth over safety, are easily taken advantage of by bad actors to promote their viewpoints through artificially induced virality. The cross-check system and other measures of content moderation at scale would not be needed if it was harder to spread problematic content on the platform in the first place. Instead of focusing only on one specific system, the Board needs to urge Meta to re-evaluate the incentives that drive content sharing on the platform and come up with ways that make the platform safer.

Meta's Obligations under Human Rights Law

Recommendation for the Board: The Board must consider the cross-check system to be violative of Meta's obligations under the International Covenant of Civil and Political Rights (ICCPR). Additionally, the cross-check ranker must be incorporated with Meta's commitments towards human rights, as outlined in its Corporate Human Rights Policy.

Explanation: Meta's content moderation, and by extension, its cross-check system, is bound by both international human rights law as well as the Board's past decisions. At the outset, The system fails the three-pronged test of legality, legitimacy and necessity and proportionality, as delineated under Article 19(3) of the International Covenant of Civil and Political Rights (ICCPR). Firstly, this system has been "scattered throughout the company, without clear governance or ownership", which violates the legality principle, since there is no clear guidance on what sort of speech, or which classes of users, would deserve the treatment of this system. Secondly, there is no understanding about the legitimacy of aims with which this system had been set up in the first place, beyond Meta's own assertions, which have been countered by evidence to the contrary. Thirdly, the necessity and proportionality of the restriction has to be read along with the Rabat Plan of Action, which requires that for a statement to become a criminal offense, a six-pronged test of threshold is to be applied: a) the social and political context, b) the speaker's position or status in the society, c) intent to incite the audience against a target group, d) content and form of the speech, e) extent of its dissemination and f) likelihood of harm. As news reports have indicated, Meta has been utilizing the cross-check system to privilege speech from influential users, and in the process, have shielded inflammatory, inciting speech that would have otherwise qualified the Rabat threshold. As such, the third requirement is not fulfilled either.

Additionally, Meta's own Corporate Human Rights Policy commits to respecting human rights in line with the UN Guiding Principles on Business and Human Rights (UNGPs). Therefore, the cross-check ranker must incorporate these existing commitments to human rights, including:

- The right to freedom of expression:, UN Special Rapporteur on freedom of opinion and expression report A/HRC/38/35 (2018); Joint Statement of international freedom of expression monitors on COVID-19 (March, 2020).
  - The Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression addresses the regulation of user-generated online content.
  - The Joint Statement issued regarding Governmental promotion and protection of access to and free flow of information during the pandemic.

- The right to non-discrimination: International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), Articles 1 and 4.
  - Article 1 of the ICERD defines racial discrimination.
  - Article 4 of the ICERD condemns propaganda and organisations that attempt to justify discrimination or are based on the idea of racial supremacism.
- Participation in public affairs and the right to vote: ICCPR Article 25.
- The right to remedy: General Comment No. 31, Human Rights Committee (2004) (General Comment 31); UNGPs, Principle 22.
  - The General Comment discusses the nature of the general legal obligation imposed on State Parties to the Covenant.

- Guiding Principle 22 states that where business enterprises identify that they have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes.

Meta's obligations to avoid political bias and false positives in its cross-check system.

Recommendation for the Board: The Board must urge Meta to adopt and implement the Santa Clara Principles on Transparency and Accountability to ensure that it is open about risks to user rights when there is involvement from the State in content moderation. Additionally, the Board must ask Meta to undertake a diversity and human rights audit of its existing policy teams, and commit to regular cultural training for its staff. Finally, the Board must investigate the potential conflicts of interest that arise when Meta's policy team has any sort of nexus with political parties, and how that might impact content moderation.

Explanation: For the cross-check system to be free from biases, it is important for Meta to come clear to the Board regarding 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. It also needs to disclose to the Board in which particular situations it uses the system and in which it does not. Principle 4 under the Foundational Principles of the Santa Clara Principles on Transparency and Accountability in Content Moderation encourage companies to realize the risk to user rights when there is involvement from the State in processes of content moderation and asks companies to makes users aware that: a) a state actor has requested/participated in an action on their content/account, and b) the company believes that the action was needed as per the relevant law. Users should be allowed access to any rules or policies, formal or informal work relationships that the company holds with state actors in terms of content regulation, the process of flagging accounts/content and state requests to action.

The Board must consider that erroneous lack of action (false positives) might not always be a system's flaw, but a larger, structural issue regarding how policy teams at Meta functions. As previous disclosures have proven, the contours of what sort of violating content gets to stay up on the platform has been ideologically and politically coloured, as policy executives have prioritized the company's 'business interests' over social harmony. In such light, it is not sufficient to simply propose better transparency and accountability measures for Meta to adopt within its content moderation processes to avoid political bias. Rather, the Board's recommendations must focus on the structural aspect of the human moderator and policy team that is behind these processes. The Board must ask Meta to a) urgently undertake a diversity and human rights audit of its existing team and its hiring processes, b) commit to regular training to ensure that their policy staffs are culturally literate in the socio-political regions they work in. Further, the Board must seriously investigate the potential conflicts of interest that happen when regional policy teams of Meta, with nexus to political parties, are also tasked with regulating content from representatives of these parties, and how that impacts the moderation processes at large.

Finally, in case decision 2021-001-FB-FBR, the Board made a number of recommendations to Meta which must be implemented in the current situation, including: a) considering the political context while looking at potential risks, b) employment of specialized staff in content moderation while evaluating political speech from influential users, c) familiarity with the political and linguistic context d) absence of any interference and undue influence, e) public explanation regarding the rules Meta uses when imposing sanctions against influential users and f) the sanctions being time-bound.

Transparency of the cross-check system.

Recommendation for the Board: The Board must urge Meta to adopt and implement the Santa Clara Principles on Transparency and Accountability to increase the transparency of its cross-check system.

Explanation: There are ways in which Meta can increase the transparency of not only the cross-check system, but the content moderation process in general. The following recommendations draw from The Santa Clara Principles and the Board's own previous decisions:

Considering Principle 2 of the Santa Clara Principles: Understandable Rules and Policies, Meta should ensure that the policies and rules governing moderation of content and user behaviors on Facebook are **clear, easily understandable, and available in the languages** in which the user operates.

Drawing from Principle 5 on Integrity and Explainability and from the Board's recommendations in case decision 2021-001-FB-FBR which advises Meta to"*Provide users with accessible information on how many violations, strikes and penalties have been assessed against them, and the consequences that will follow future violations*", Meta should be able to **explain the content moderation decisions to users in all cases**: when under review, when the decision has been made to leave the content up, or take it down. We recommend that Meta keeps a publicly accessible running tally of the number of moderation decisions made on a piece of content till date with their explanations. This would allow third parties (like journalists, activists, researchers and the OSB) to keep Facebook accountable when it does not follow its own policies, as has previously been the case.

In the same case decision, the Board has also previously recommended that Meta "*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-checking compared with ordinary enforcement procedures.*" Thus, Meta should **publicly explain the cross check system** in detail with examples, and make public the list of attributes that qualify a piece of content for secondary review.

The Operational Principles further provide actionable steps that Meta can take to improve the transparency of their content moderation systems. Drawing from Principle 2: Notice and Principle 3: Appeals, Meta should make a satisfactory **appeals process available** to users - whether they be decisions to leave up or takedown content. The appeals process should be handled by context aware teams. Meta should then **publish the results** of the cross check system and the appeals processes as part of their transparency reports including data like total content actioned, rate of success in appeals and cross check process, decisions overturned and preserved etc, which would also satisfy the first Operational Principle: Numbers.

Resources needed to improve the system for users and entities who do not post in English.

Recommendations for the Board: The Board must urge Meta to urgently invest in resources to expand Meta's content moderation services into the local contexts in which the company operates and invest in training data for local languages.

Explanation: The cross-check system is not a fundamentally different problem than content moderation. It has been shown time and time again that Meta's handling of content from non-Western, non-English language contexts is severely lacking. It has been shown how content hosted on the platform has been used to inflame existing tensions in developing countries, promote religious hatred in India, genocide in Mynmar, and continue to support human traffickers and drug cartels on the platform even when these issues have been identified.

There is an urgent need to invest resources to expand Meta's content moderation services into the local contexts in which the company operates. The company should make all policies and rule documents available in the languages of its users; invest in creating automated tools that are capable of flagging content that is not posted in English; and add people familiar with the local contexts to provide context aware second level reviews. The Facebook Files show that even according to company engineering, automated content moderation is still not very effective in identifying hate speech and other harmful content. Meta should focus on hiring, training and retaining human moderators who have knowledge of local contexts. Bias training of all content moderators, but especially those who will participate in the second level reviews in the cross check system is also extremely important to ensure acceptable decisions.

Additionally, in keeping with Meta's human rights commitments, the company should develop and publish a policy for responding to human rights violations when they are pointed out by activists, researchers, journalists and employees as a matter of due process. It should not wait for a negative news cycle to stir them into action as it seems to have done in previous cases.

Benefits and limitations of automated technologies.

Meta recently changed its moderation practice wherein it uses technology to prioritize content for human reviewers based on their severity index. Facebook has not specified the technology it uses to prioritize high-severity content but its research record shows that it uses a host of automated frameworks and tools to detect violating content, including image recognition tools, object detection tools, natural language processing models, speech models and reasoning models. One such model is the Whole Post Integrity Embeddings ("WPIE") which can judge various elements in a given post (caption, comments, OCR, image etc.) to work out the context and the content of the post. Facebook also uses image matching models (SimSearchNet++) that are trained to match variations of an image with a high degree of precision and improved recall; multi-lingual masked language models on cross-lingual understanding such as XLM-R that can accurately identify hate-speech and other policy-violating content across a wide range of languages. More recently, Facebook introduced its machine translation model called the M2M-100 whose goal is to perform bidirectional translation between 7000 languages.

Despite the advances in this field, there are inherent limitations of such automated tools. Experts have repeatedly maintained that AI will get better at understanding context but it will not replace human moderators for the foreseeable future. One such instance where these limitations were exposed was during the COVID-19 pandemic, when Facebook sent its human moderators home - the number of removals flagged as hate speech on its platform more than doubled to 22.5 million in the second quarter of 2020 but the number of successful content appeals was dropped to 12,600 from the 2.3 million figure for the first three months of 2020.

The Facebook Files show that Meta's AI cannot consistently identify first-person shooting videos, racist rants and even the difference between cockfighting and car crashes. Its automated systems are only capable of removing posts that generate just 3% to 5% of the views of hate speech on the platform and 0.6% of all content that violates Meta's policies against violence and incitement. As such, it is difficult to accept the company's claim that nearly all of the hate speech it takes down was discovered by AI before it was reported by users.

However, the benefits of such technology cannot be discounted, especially when one considers automated technology as a way of reducing trauma for human moderators. Using AI for prioritizing content for review can turn out to be effective for human moderators as it can increase their efficiency and reduce harmful effects of content moderation on them. Additionally, it can also limit the exposure of harmful content to internet users. Moreover, AI can also reduce the impact of harmful content on human moderators by allocating content to moderators on the basis of their exposure history. Theoretically, if the company's claims are to be believed, using automated technology for prioritizing content for review can help to improve the mental health of Facebook's human moderators.