Mnemonic public comment, Oversight Board Policy Advisory Opinion 2021-02

Summary of the comment:

Stopgap solutions to bad content moderation enforcement aren't good enough. Meta should not invest significant financial and people resources in bandaids like Meta's cross-check system (Xcheck) when wholescale improvement of Meta's content moderation policies and systems is needed. This is particularly true for at-risk countries and non-English moderation, especially of Arabic and other non-Latin languages. Fixing Meta's "language problem" would improve moderation for all users, not just those lucky enough to make it on the Xcheck list. That being said, Xcheck can be improved with more transparency as well as significant outreach with civil society to ensure ranking factors and the data used to train any automated systems are high quality.

***

We'll start this comment by doing something unusual: agreeing with Meta. Or rather, the Meta employee in a leaked internal document explaining why "Xcheck" will no longer support whitelisting: "A common misuse of the Xcheck product has been to "whitelist/"exempt" content/entities when we are uncomfortable in the quality of our detection….Xcheck was never intended to be a loophole for shipping non-trustworthy classifiers. Please continue to focus on ensuring your classifier quality is high enough to ship."

We agree that stopgap solutions to bad content moderation enforcement aren't good enough. Meta should not invest significant financial and people resources in bandaids like the cross-check system (Xcheck) when it is clear that wholescale, if surgical, improvement of Meta's content moderation policies and systems are needed.

With Xcheck, Meta attempts to address the problems caused by trying to moderate content at scale in a cost-efficient, PR-friendly way. Creating an extra layer of review is not a particularly surprising system, so why is it that Meta has repeatedly misled the public and the Board about Xcheck? The reason is obvious—Meta claims it has the same rules for everyone, no matter where in the world they are. In reality, Meta has created a tiered system for content that favors celebrities, governments, and ultimately Meta's business interests.

This comment focuses on Xcheck in the context of conflict zones, countries at risk of mass killing, and political protests, i.e. countries that fall under Meta's "at-risk" rubric. Ultimately, we don't feel qualified to make a blanket statement that the system should be entirely scrapped. It's possible Xcheck is protecting human rights defenders and their content. It's clear from Meta's internal documents that Xcheck interacts with numerous other related systems, and we do not have enough information about them all. However, we do recommend that the Board push Meta to invest in other, more equitable content moderation solutions and improve Xcheck by working closely with civil society. We also recommend that Meta provide more transparency about Xcheck and the related systems it interacts with.

Should Xcheck exist at all?

The Board asks about how Xcheck can and should be improved. We will start with the question of whether it should be improved at all. Xcheck's additional layer of review is meant for "content that presents a greater risk of false positives." Entities on Facebook and Instagram are added to Xcheck on the request of Meta staff. They look at a number of factors, including potential PR threats to the company.

As a preliminary matter, we argue that the resources that go into Xcheck could instead be used to improve content moderation across the Board- including the problem of false positives (content/accounts incorrectly removed) AND false negatives (content/accounts that should be removed based on Meta's policies but is instead left up). False positives are a very real problem, but so are false negatives—they can cost lives. Meta itself told the board "that despite investing significant resources to improve cross-check it still has difficulties striking a balance between removing content that violates Meta's policies 'while ensuring that it continues to foster open communication and free expression.'" The Board should focus on real-world harm, removing content that incites offline violence or other documentable harm, and focusing its efforts to protect free expression on the most vulnerable users.

Fix Meta's language problem

As noted, surgical improvement of specific and acknowledged trouble areas is a better place to start than "fixing Xcheck," The Board rightly asks about "how the cross-check system should and can be improved for users and entities who do not post in English." We argue that moderation for non-English languages should be improved across the board. An AP report noted that "in some of the world's most volatile regions, terrorist content and hate speech proliferate because the company remains short on moderators who speak local languages and understand cultural contexts." This is compounded by the fact that language problems are happening not only at the moderation stage, but in the creation of training data sets for automation ("labeling"). And once automation starts incorrectly removing content and teaching itself, problems mushroom. One of Meta's internal documents that focused on problems in labeling Arabic content noted that "Portuguese, Spanish, Hindi, French, Mandarin Chinese, and Indonesian are all concerns"....and these are only the large markets. Users in smaller markets are even more sidelined. It's clear that Meta has to improve its non-English moderation globally.

Mnemonic works mainly in Arabic-speaking countries. As the 3rd or 5th most spoken language on the platform (according to internal documents and depending on how one counts), Arabic is a disturbing but helpful example of how poorly Meta divides its resources. We and others from our region have long argued that Meta's moderation of Arabic-language content is substandard and Meta's internal documents confirmed this. One internal document notes that "Facebook is either failing to moderate a large amount of content in certain Arabic countries/dialects or not documenting the decisions made on those closed jobs." As reported by Politico, an internal presentation indicated that " algorithms to detect terrorist content incorrectly deleted non-violent Arabic content 77 percent of the time." The documents also revealed astoundingly poor quality control for labeling of hate speech qualifiers. For example, data "suggests that labellers are deciding content from dialect groups they are unlikely to have expertise in" and "dialect and/or cultural expertise (e.g. country of origin) of the reviewers does not exist in records and even language information is inadequately documented." .

Meta should not just be "investing significant resources" into a tool that their own internal documents acknowledge exists partly to avoid PR nightmares. Instead, they should do their best to improve the platform for all users. For the content we are concerned with, that means addressing the failures in moderation of Arabic-language content in consultation with civil society and investing in the myriad fixes already suggested by their own employees. This process should be undertaken globally, with a focus on conflict zones and countries at risk of mass killing. Meta should also engage in an across-the-board review of its use of automation on non-Latin languages. As we've noted in other comments to the board, co-design and transparency into training data for automation would also help address this issue, from the lists of classifiers to their translations. That means, for example, working directly with civil society to create reliable lists of classifiers.

Finally, Meta is ignoring a valuable resource and a crucial step in content moderation: engaging with its users and listening to their feedback. This is not only important for better transparency and fairness for the users (ie in the appeal process), but also can provide a valuable resource and feedback to improve the corporation's understanding of local contexts and languages. Meta has the capacity to engage in user research on this important topic, and it should do so

Xcheck doesn't seem to help human rights defenders

Meta claims that in addition to "managing Meta's relationships with many of our business partners" and protecting government and celebrity accounts, Xcheck protects journalists and community leaders. Unfortunately, our experience has shown that Xcheck is failing at that. Conversely, as noted in Meta's own documents, it has allowed content that violates Meta's community standards, including threats against vulnerable individuals, to be left up indefinitely. As an organization, Mnemonic has had to repeatedly push Meta to reinstate groups. It's an open secret that civil society groups like ourselves spend thousands of unpaid hours every year providing case management services for Meta and other social media platforms. Journalists also spend a significant amount of resources serving as social media platform watchdogs. Media attention and civil society labor regularly gets content reinstated that was improperly removed. But, we also regularly see that content taken down or accounts suspended again, sometimes only hours or days later. This often seems to be the result of reporting campaigns by bad actors, including state-sponsored "cyber armies," for example the Syrian Electronic Army, a known pro-Assad troll army.

Although we can't publicly comment on all the cases we've provided support for, we can say that we reported suspensions of groups that were used by human rights defenders coordinating documentation and archiving of human rights violations. We've seen these groups reinstated and removed several times, with little communication about why or how these decisions are reached. On a more public-facing note, high profile Palestinian community organiser Mohamed el Kurd, someone who has received broad media attention and recognition, had his account features limited and his account suspended multiple times in May of last year, while he was reporting on increased violence against Palestinians and garnerning international support for Palestinians' human rights.

This process doesn't only waste civil society's time. Meta also fails to meaningfully benefit from the work civil society organizations are doing for the platforms. The fact that accounts are suspended, reinstated, then suspended again, only to be reinstated once more, over and over again, shows little to no interest on Meta's side to use these processes to grow and improve their system.

Harm reduction for Xcheck

Xcheck doesn't seem to be going away. We do think that there are times when Meta should be providing an extra level of review, but that this extra review should be based on the potential impact of content, not simply on the identity of the user who posted the content and certainly not on potential effects on Meta's business interests. In particular extra review is a tool that should be used in contexts of war, violent conflict, and risks of genocide. The Board has reviewed myriad cases from contexts where this is important, including India, Myanmar, and Ethiopia, and its decisions reflect the reasons that content such as incitement to violence and misinformation must be taken seriously.

We are particularly suspicious of the idea of an automated or mostly automated "cross-check ranker.” This would rely on quality training data in myriad languages. Since this will likely happen anyway, we urge the Board to call on Meta to co-design its cross-check ranker with civil society. Meta should especially partner with civil society from conflict zones and at-risk countries, including Syria, Yemen, and Sudan (three countries where we conduct work), as well as the entire MENA region. We also believe that the ranker should be considering the risk of false negatives in addition to the risk of false positives. As noted above, we believe the Board has enough experience with the problems of content that is left up in places like Ethiopia at this point to understand why.

Meta has made it clear that it will also continue to provide secondary review "using a list-based approach." We argue that Groups, profiles, and content that Meta restores based on reports from civil society should be included in that list, not subject to whatever algorithmic ranking Meta will do. The list should not serve as shielding from ALL review, even for groups added at the behest of civil society, but it should ensure that human rights defenders are reviewed at the same level as the politicians that oppose them.

Finally, the transparency questions raised by Xcheck reflect a lot of issues the Board has faced in pushing Meta for more transparency. Some of these issues are legitimate. For example, we understand the security risks associated with publicly explaining when human rights defenders are enrolled in any kind of program with Meta. That being said, more transparency is clearly possible. There don’t seem to be problems with sharing the lists of business partners, celebrities and politicians on the list. We also want Meta to explain what reason could exist for not notifying groups when they are enrolled in Xcheck. We also urge Meta to clarify to civil society how our labor feeds into Xcheck. Finally, the indicators for the cross-check ranker should be clearly explained.

We are pleased that the board is considering the Xcheck program, but we don’t want Xcheck to serve as a distraction from the dozens of improvements the Board and civil society have already suggested to Meta. We hope the Board’s decision will reflect this.