Meta

August 29, 2024

Prof Robert McCorquodale and
other expert mandate holders
Chair-Rapporteur of the United Nations Working Group
on the issue of human rights and
transnational corporations and other business enterprises
Palais des Nations
1211 Geneva 10
SWITZERLAND

by email

Mr Chairman, Special Rapporteurs;

Thank you for your joint letter (Ref.: AL OTH 20/2024) of April 18th in which you raise a number of questions about Meta's response to events in Israel and Palestine since October 7th.

We apologize for the delay in responding. As we've discussed, a UN communications error meant we became aware of your letter in late July. And while the initial deadline has expired, we believe it's important to share information on our human rights risk management in response to the issues you have raised.

In the immediate aftermath of the 7 October terrorist attacks Meta implemented immediate crisis response measures, including a dedicated 24x7 cross-functional crisis response team.

In doing so, we were guided by core human rights principles, including respect for the right to life and security of person; protection of the dignity of victims; non-discrimination; and freedom of expression. We looked to the UN Guiding Principles on Business and Human Rights, to prioritize and mitigate for the most salient human rights. We also used international humanitarian law as an important reference. We publicly shared response details in our newsroom on October 13 (and updated on October 18 and December 5) in English, Arabic, and Hebrew.

During the ongoing conflict in Israel and Palestine, there has been a surge in related content on our platforms: this includes large volumes of non-violating content discussing and raising awareness of events, but also of content that violates our policies on hate speech, violence and incitement, bullying and harassment, dangerous organizations and individuals, and violent and graphic content. While our platforms are designed to support voice, we also seek to mitigate risks that may impact the safety and well-being of our community, in line with Meta's responsibility to prevent or mitigate its most salient human rights risks. (For more information on Meta's salient human rights risks, please see our most recent Annual Human Rights Report, pages 13-26, available in multiple languages).

Taking both safety and voice into consideration is difficult in peaceful contexts and even more so in conflict situations—especially those involving sanctioned entities such as Hamas.

Our response is guided by our prior crisis experience, as well as from recommendations made in the independent human rights due diligence on Israel and Palestine by Business for Social Responsibility (BSR) that we commissioned and disclosed in 2022 (and shared in an update on our implementation in September 2023).

We also have used our Crisis Policy Protocol, first launched in 2022 after extensive consultation, to guide our actions.

Our Human Rights Team has been closely involved in Meta's response and has conducted ongoing, integrated human rights due diligence throughout, in line with our Corporate Human Rights Policy and the UN Guiding Principles on Business and Human Rights. We plan to include information on this work, as well as on our continuing efforts to address the recommendations made by BSR, in our forthcoming annual human rights report.

A major priority for us is to work to ensure that we are not amplifying harmful and inflammatory content, which is present in all relevant markets.

During critical moments with elevated risk of violence or other severe human rights risks, we may adapt our standard approach to keeping people safe while still enabling them to express themselves and initiate temporary measures to help keep people safe.

We closely monitor offline events and track platform trends: for example, how much violating content people are seeing on Facebook or Instagram and whether we're starting to see new forms of abusive behavior that warrant changes in our response.

As we've detailed in our blog post describing our response to the conflict, we have adopted a number of temporary product and policy measures to help keep people safe and mitigate salient human rights risks.

We don't implement such temporary measures lightly: we know that they can have unintended consequences, like inadvertently limiting harmless or even helpful content. And we have heard complaints from many organizations about the unintended consequences of these measures.

That's why we seek to ensure that the steps that are taken are time limited and proportionate to the risks as we are aware of them. That's also why our Human Rights Team is embedded within our crisis response process to carry out integrated human rights due diligence that informs our approach.

Some examples of the safety measures we implemented include:

- Changes to how we recommend unconnected content: We temporarily reduced the threshold at which borderline or potentially violating content—like images or videos depicting graphic violence—may be made ineligible for recommendation. This measure applied to unconnected content—that is, content from people that someone hasn't already chosen to follow that may appear on surfaces like Feed, Search, Explore, and Reels.
- Adjustments to confidence thresholds for automatically actioning content: We use machine learning classifiers to identify potentially harmful content and automatically action it when we have a high confidence that it violates our policies. In crisis situations, we may lower the confidence level at which we automatically take action—as we did in this conflict—to address a persistent spike in violations. In doing so, we seek to reduce the level of violating content (such as hate speech, violence and incitement, bullying and harassment, or graphic violence) to a prevalence that is equitable across languages and markets. This means that confidence thresholds for classifiers in each relevant language may be adjusted individually to reflect differences in content trends in specific languages or markets.
- Blocking certain hashtags from search: We temporarily made a number of hashtags unsearchable on Instagram after we determined that they were frequently being used in association with content that violated our policies. Content that used these hashtags but that did not violate our policies was not removed.
- Product changes to address unwanted and problematic comments: Following a spike in unwanted and problematic comments, we changed the default settings for who can comment on public posts made by people in the region impacted by the conflict to include only friends or established followers; a poster could change this setting back to allow comments from anyone if desired.
- Launching the Lock Your Profile tool in the region: To address safety and harassment concerns, we gave people in the region impacted by the conflict the ability to lock their Facebook profiles if they wished to do so. When someone's profile is locked, people who aren't their friends can't download, enlarge or share their profile photo, nor can they see posts or other photos on someone's profile, regardless of when they may have posted it.

We also provided tools that made it easier for people to bulk delete comments on their posts, and we stopped showing the first one or two comments on a post automatically in Feed.

At the same time, we also made other changes specifically aimed at ensuring we were protecting voice. For example, in response to a large spike in usage of our products, we temporarily adjusted some automated rate limits designed to prevent spam to make them more permissive, reducing the risk of restrictions on legitimate users.

For some policy areas, like the most graphic types of violent and graphic content, we're removing violating content without applying strikes—the penalties for violations that result in escalating account restrictions as they accumulate—to ensure we're not overly penalizing or restricting users who are trying to raise awareness of the conflict's impacts.

Separately, we globally limit recommendations of unconnected content related to politics and social issues, including conflict, across Facebook and Instagram.

Our Community Standards prohibit a wide range of potentially harmful content, including violence and incitement, hate speech, dangerous organizations and individuals, and violent and graphic content. These policies apply to all content shared on Facebook and Instagram—regardless of who posts it—and are informed by human rights standards, international humanitarian law, extensive stakeholder input, and two decades of practical experience with content moderation.

Our policies are designed to address content that may amount to incitement to violence, hatred, or genocide. We have adapted the principles of the Rabat Plan of Action into actionable content policy tools, including escalation-based frameworks to evaluate speech attacking concepts (as opposed to people) and content involving state threats to use force.

In rare cases, we allow content that may violate our policies if it's newsworthy and if keeping it visible is in the public interest. We only do this after conducting a thorough review that weighs the public interest against the risk of harm. We look to international human rights standards, as reflected in our Corporate Human Rights Policy, to help make these judgments. For content we allow that may be sensitive or disturbing, we include a warning screen. In these cases, we can also limit the ability to view the content to adults, aged 18 and older.

We removed multiple pieces of content from the Israeli group "Tzav9" for organizing efforts to blockade humanitarian aid trucks, which violated our Coordinating Harm policy, and led to the disabling of their Facebook and Instagram accounts. These actions were in line with international humanitarian law and the need to allow and facilitate the rapid and unimpeded passage of humanitarian relief for civilians in need.

Overall, our response has benefited significantly from our prior crisis experience, as well as from the human rights due diligence we conducted and disclosed in 2022, and are continuing to implement (see our September 2023 update here).

We leverage a combination of technology and human review teams to detect and enforce on content that violates our policies. The technologies we use include a wide range of both language-specific classifiers and language-agnostic classifiers; these include classifiers to address policy violating content in both Arabic and Hebrew languages.

Based on recommendations emerging from the independent Israel/Palestine human rights due diligence we conducted in 2022, we have taken a number of specific steps to improve our Arabic and Hebrew language classifiers. These include developing and launching a hostile speech classifier for Hebrew and expanding language identification for Arabic to recognize content in different Arabic dialects. We shared details on this work in our September 2023 Israel/Palestine Human Rights Due Diligence update.

Throughout the crisis, we have been in touch with experts in human rights and international humanitarian law to ensure that we are taking account of their expertise .

You also ask about responses to government takedown requests. We want to be clear: we do not remove content simply because a government entity (or anyone else) requests it. When we receive a content takedown request from a government entity, we review it following a consistent global process.

First, we evaluate it in the same way we would a report from any other source, reviewing it against our Community Standards. If we ourselves determine that the reported content violates one of our policies, we take action and notify the person who posted it that we did so.

If we determine that the content does not go against our policies but a government has alleged that it violates local law, we may restrict access to the content in the country where it's alleged to be illegal after a careful legal and human rights assessment conducted in line with our commitments as a member of the Global Network Initiative.

When we do restrict content in specific jurisdictions on the basis of local law, we're transparent about our actions: we directly notify the person who posted the content as well as anyone who tries to view it but is blocked from doing so, and we also publish data on the restriction in our biannual Content Restrictions Report. Our most recent report for the second half of 2023 is available here.

As we shared in our September 2023 Israel/Palestine Human Rights Due Diligence update and most recent Quarterly Update on the Oversight Board, we are still in the process of developing consistent and reliable systems for gathering metrics on the number of pieces of content removed under the Community Standards as a result of government requests. We continue to evaluate approaches to building the necessary internal data logging infrastructure to enable us to publicly report this information across the diversity of request formats we receive, but we expect this to be a complex, long-term project.

If it would be of interest to meet to discuss our responses in more depth, please do not hesitate to let us know.

Yours sincerely,

Miranda Sissons
Director, Human Rights Policy