Meta

Ms Marwa Fatafta
MENA Policy and Advocacy Director
Access Now
by email: [redacted]

2 April 2024

Dear Ms Fatafta,

Thank you for your letter of 14 March 2024 on behalf of the Stop Silencing Palestine coalition; for meeting with our executives on 22 February; and for our ongoing dialogue. We know that people in the region and around the world have felt deeply impacted by our response to the ongoing conflict situation in Israel and Palestine.

In the immediate aftermath of the 7 October terrorist attacks against Israel, Meta implemented immediate crisis response measures, including a dedicated 24x7 cross-functional crisis response team. As we did 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, as enshrined in our Corporate Human Rights Policy, to prioritize and mitigate the most salient human rights. We also used international humanitarian law as an important reference.

We initially shared details of our response in a blog post in English, Arabic, and Hebrew on 13 October 2023. We provided further updates to the post on 18 October, 5 December, and 8 December 2023.

As you know, our regional and global teams have also engaged on an ongoing basis with you and other human rights stakeholders, as well as experts in the law of armed conflict. We've continued to refine our approach to reflect changing dynamics, including the ongoing humanitarian crisis in Gaza.

Obviously, in exceptional and fast-moving situations like this one, no response can be perfect, lines are difficult to draw, and people and systems can and will make mistakes.

We're attaching further detail in response to your questions below. We remain ready to review specific cases through our escalation mechanisms and look forward to our engagement in the future.

Yours, very sincerely,

Miranda Sissons
Director, Human Rights Policy

Our overall approach

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, dangerous organizations and individuals, and violent and graphic content. While our platforms are designed to support voice, we must also seek to mitigate risks that may impact the safety and well-being of our community.

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 recently shared an update on our implementation of 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 routine annual human rights reporting.

How we address harmful content

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.

Temporary measures to help keep people safe

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. 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 did adopt 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. That's why we seek to take steps that 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, 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. 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.
- **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.

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 certain 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 political issues, including conflict, across Facebook and Instagram. This change is **not** specific to content related to the conflict in Israel and Palestine nor is it a part of our crisis response actions, but may impact pieces of content related to the conflict.

We acknowledge that some of the temporary measures we implemented may be disruptive, even though we sought to make them specific, proportionate, and time limited. But our goal with these temporary measures—most of which have now been lifted—has been to seek to effectively mitigate salient human rights risks in a violent and dynamic situation. (For more information on Meta's most salient human rights risks, please see our most recent Annual Human Rights Report, pages 14-26.)

On AI and automation

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 harmful 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.

In your letter, you cite an example of problematic machine translations and refer to a WhatsApp issue involving AI-generated stickers. We agree: the outputs you're referring to are unacceptable.

As soon as we became aware of these issues, our engineering teams immediately began an investigation to determine root causes and implement appropriate fixes. Our teams identified that these issues appeared to be related to model hallucinations and the training data used. Unfortunately, both issues are well-documented challenges for AI-powered products.

We delivered an emergency fix to our machine translation product within 90 minutes of our engineering team beginning their investigation. We also worked to rapidly mitigate potentially problematic associations in generated image and sticker outputs.

In addition to these rapid fixes, we have also worked to fine-tune our foundation models for image generation to better address a wide range of potentially problematic associations, and these broader mitigations have now begun being deployed across our products.

We continue to heavily invest in improving our AI products, including in the areas of inaccurate output, model hallucinations, and potential biases in training data and outputs. We have acknowledged these challenges and we've also shared details about this work on our Responsible AI and AI Research websites.

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. We expect to publish data covering the second half of 2023 later this year.

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.

Evidence Retention

We support justice for all international crimes. We've worked since 2019 to explore rights-respecting initiatives for evidence retention and disclosure, consulting extensively with civil society, academia, and international prosecutorial experts and bodies.

As we publicly stated in response to recommendations from the Oversight Board, we worked for several years to develop an approach to allow international courts and accountability mechanisms to make requests to us for extended retention of data that is relevant to their ongoing investigations. This work, which is distinct from our longstanding policies for responding to preservation requests from law enforcement, has now been largely completed.

Last month, we briefed a range of UN-authorized mechanisms and special rapporteurs—including a representative of the UN-authorized Independent International Commission of Inquiry on the Occupied Palestinian Territory, including East Jerusalem, and Israel—on our approach, and also outlined the process for making extended retention requests to Meta. We will carefully review all requests received for consistency with our policies and applicable law. While we will require that requests be as specific and narrow as possible and clearly in scope of the requestor's internationally-authorized mandate or authority, we will not require individual content URLs when they are unavailable or irrelevant to the scope of the request.

This is a novel area without established or tested best practices, and there remain significant legal, privacy, and policy considerations inherent to this work. We expect to share further updates on our work in this area in our Quarterly Updates on the Oversight Board and our annual human rights reporting.

END OF RESPONSE
REMAINDER OF PAGE INTENTIONALLY LEFT BLANK