Annex II: Meta's Response Letter to Human Rights Watch

Meta

December 6, 2023

Tirana Hassan
Executive Director
Human Rights Watch
350 Fifth Avenue, 34th floor
New York, NY 10118-3299
United States

Dear Ms Hassan,

Thank you very much for your letter of 15 November 2023. We deeply appreciate the work of Human Rights Watch to document violations and advance human rights, often in the most difficult circumstances.

In the wake of the 7 October terrorist attacks in Israel, Meta took immediate crisis response measures. As we did so, we were guided by core human rights principles, including respect for the right to life and security of the person, the protection of the dignity of victims, and the right to non-discrimination - as well as balancing those with the right to freedom of expression. We looked to the UN Guiding Principles on Business and Human Rights to prioritize and mitigate the most salient human rights risks: in this case, that people may use Meta platforms to further inflame an already violent conflict. We also looked to international humanitarian law (IHL) as an important source of reference for assessing online conduct.

As conflict-related content exponentially surged on our platforms, we implemented a number of temporary measures across both Arabic and Hebrew markets, seeking equitable outcomes in limiting prevalence of violating content on our platforms. We note we've sought to make these measures time-bound, proportionate, and consistent with established human rights guidance on permissible limitations for freedom of expression.

We publicly shared details in our newsroom on October 13 and updated on October 18 and December 5 in English, Arabic, and Hebrew, and have attached a copy to this letter.

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 most recent update here). For example, we were able to route Arabic

content more appropriately across our systems to enable greater precision, and also to use our Crisis Policy Protocol to guide our actions.

Obviously, in exceptional and fast-moving situations like this, our response can never be perfect, lines are difficult to draw, and people or systems can and will make mistakes. We know that our users in both the Arabic and Hebrew markets have felt deeply impacted by our decisions. In this respect, we note that none of your questions appear to relate to Hebrew-related markets, measures, or user feedback.

As the conflict evolves, we'll no doubt face new challenges and will continue to learn from civil society feedback and insights from this current period. And we have actively sought out and have heard strong viewpoints from civil society and UN actors in the region and beyond.

We're providing further detail below the signature line – including on our continuing efforts to improve our response. Before doing so we'd like to note: a number of your questions ask us for our opinion on content and/or accounts we have not seen. It is important that we see the content and conduct an assessment based upon all of the available context to reach a determination on whether it violates our Community Standards or Community Guidelines.

We are more than happy to review content samples with you, or to escalate them through our systems, as we are doing daily for many individuals and groups across all relevant markets and languages. Indeed, there are significant benefits in conducting such escalations directly and in real time, both to remedy errors and to share context on specific decisions or policies.

We plan to report back in future on our continuing efforts to address the recommendations made by Business for Social Responsibility in the Israel Palestine human rights due diligence, including any learnings from our current efforts.

Yours sincerely,

Miranda Sissons
Director of Human Rights Policy

Attached:
Response Details
Newsroom Post (English. Arabic here and Hebrew here).

Response Details

Our overall approach

During the current conflict, there has been a surge in related content on our platforms, including 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 ensure the safety and well-being of our community – and also respond to adversarial coordinated behaviors.

The balance between voice and safety is often not easy to strike in peaceful contexts. In conflict situations – and especially conflict situations involving sanctioned entities, such as Hamas – it is much more difficult.

Content volumes have risen sharply, as has the amount of content people are reporting to us for review. These significantly higher volumes mean we're aware that content that doesn't violate our policies could be removed in error, and that even a small margin of error at scale may result in thousands of negative impacts. To mitigate these risks, we've introduced an important fairness mitigation: for some policy areas, we are temporarily removing violating content without enforcing user strikes, meaning these content removals won't cause accounts to be disabled.

For example, certain posts that violate our Violent and Graphic Content policies may be removed without a strike to ensure that we are not overly penalizing users who are trying to share graphic content and raise awareness about the casualties of war.

We also continue to provide tools for users to appeal our decisions if they think we made a mistake. We have provided more details regarding these steps in our newsroom post.

Hostage video content

In fast-moving crisis or conflict situations, we may need to adapt our guidelines as a situation evolves. This enables our content reviewers to address emerging issues, and also ensure consistency in the application of our Community Standards and Community Guidelines.

In the context of the Israel-Gaza conflict, we initially expanded our Violence and Incitement policy to remove content that depicted identifiable hostages being kidnapped or held in captivity, even when it was done to condemn or raise awareness of their situation. We did so to protect the dignity of the victims in line with international humanitarian law standards (rule against public curiosity in the Geneva Conventions), and also to ensure that Hamas propaganda was not appearing on our platforms in line with our Dangerous Organizations and Individuals policy.

This approach was based on feedback from international humanitarian law experts and civil

society in Israel that have raised concerns about the use of footage produced by Hamas showing hostages in captivity.

As the war continues, we have seen the reasons for which people may be sharing hostage related content evolve. In particular, people began sharing hostage content to rebut emerging narratives that the October 7 kidnappings did not actually take place.

We adopted a more nuanced approach accordingly: content showing moment-of-kidnapping footage with speech that condemns the act, or which includes information for awareness-raising purposes is now allowed, with the addition of a warning screen to inform users that it may be disturbing (for example, this could include activists raising awareness of instances of violence or journalists reporting from conflict zones). Under our privacy policy, we may also remove hostage release imagery of minors (under 13). We continue to remove any Hamas-produced footage, unless it is shared in a condemning or news reporting context.

As always, teams are considering context around this imagery, and newsworthy allowances are available where appropriate to balance the public interest against the risk of harm.

You asked if we could share our internal guidelines with you. We want to be clear that Meta does not share internal guidance with external parties unless legally bound to do so.

Appeals

Where a decision is eligible for appeal, people are given the option to ask us to take another look after receiving a notification that their content has been removed or covered with a warning. These appeal mechanisms are available on both Facebook and Instagram. Both human review teams and technology play a role in reviewing user reports and appeals, and we aim to prioritize appeals with potentially harmful content first. During busy periods, such as during conflict situations, we may not always be able to review everything based on our review capacity. Appeals for content demotions are currently not available outside of the EU.

In response to your question about how we define our severest policies, our severe policies refer to those that have a heightened risk of imminent physical harm to individual or public safety. These are assessed and frequently updated to reflect external user research, input from industry stakeholders, safety experts, and law enforcement, as well as a reflection of Meta's company values. These policies would typically include posting content involving terrorism, child exploitation, human trafficking, suicide promotion, sexual exploitation, the sale of non-medical drugs or the promotion of dangerous individuals and organizations.

Temporary measures to help people stay safe

During critical moments with elevated risk of violence or other severe human rights risks, we rigorously seek to tailor our approach to keeping people safe while respecting their ability to express themselves. We closely monitor real world events and platform trends and track different metrics: for example, how much violating content is on Facebook or Instagram, or

whether we're starting to see new forms of abuse where we need to quickly adjust our response.

As we mentioned in our newsroom post, we decided to adopt a number of temporary measures in order to keep our users safe and mitigate the risks that our platforms could be used to further exacerbate tensions both online and offline.

We do not implement temporary measures lightly — we know that there could be 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.

We use machine learning classifiers to identify harmful content and automatically action content when we have high confidence it violates our policies. In emergencies, we may lower the confidence level at which we automatically take action, as we did in this conflict, to address a persistent spike in potentially hateful comments across the region. While these measures are disruptive, our goal is to seek to effectively mitigate salient human rights risks, including the right to life, liberty and security of person; the right to non-discrimination; and others. (For more information on Meta's most salient human rights risks, please see our most recent human rights report, pages 14-26).

Once we see spikes in potentially harmful content return to normal levels and determine the risk to safety on or off our platform has subsided, we will turn off the associated temporary measures.

We note that in conflicts, there are also frequent technical and connectivity challenges that impair user experience, but may result from technical limitations; actions of third parties; or impacts on bandwidth and infrastructure. We know you are aware of the Instagram technical problem on October 16 that impacted users' reach equally across the world, and was fixed as quickly as possible.

Government takedown requests

We do not remove content simply because a government entity (or anyone) requests it. When we receive a content takedown request from a government entity, we evaluate the content in the same way we would for any other piece of content. We first review it against the Facebook Community Standards or Instagram Community Guidelines and, if we determine that the content goes against our policies, we take action. When governments report content that does not go against our policies but is alleged to violate local law, we may restrict access to the content in the country where it's alleged to be illegal, following a careful legal review and a human rights assessment.

As we noted in our September 2023 Israel Palestine Human Rights Due Diligence update, 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. The objective is to produce government takedown request metrics in

the most efficient manner given ongoing challenges including confidentiality obligations and data logging and taxonomy gaps from internal systems. 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 in which we receive it, but we expect it to be a complex, long-term project.

You asked whether we would be instituting any firewalls between our Public Policy staff, "including former Israeli and other government officials" to prevent any undue influence over content moderation decisions. The idea that the background of our team members has an inappropriate influence on the company's content decisions and enforcement is misplaced and inaccurate: employees of different backgrounds from around the world, including Israelis and Palestinians, are represented in teams working across the company - a fact that's been confirmed by the independent human rights due diligence exercise we commissioned from Business for Social Responsibility.

All Meta employees have to respect our Code of Conduct, which defines the expectations for how we act and how we make decisions. The Code includes specific provisions on human rights and on interacting with governments and political entities responsibly. Our Public Policy staff help us better understand the realities of the situation on the ground for the countries they cover. They regularly engage with government officials, lawmakers, regulators and civil society stakeholders to foster greater understanding of Meta's products, policies and positions across a wide range of issues. The Content Policy team is responsible for writing our Community Standards and deciding how these should be applied in crisis situations. The team is made up of subject matter experts who consult with a wide range of internal cross-functional partners, external experts where appropriate and utilize experience from previous events to understand the specific context to reach these decisions.

Our continuing efforts to improve our response

Since the onset of the current conflict, Meta has been conducting weekly investigations to provide directional analysis to assess the performance of our content moderation systems. The sampling methodology has been designed to assess whether there are issues of over- and under-enforcement on the conflict-related content based on the internal data sets and the submissions by external partners. These included content in Arabic, Hebrew, and English. A dedicated team identifies core issues causing misalignment with what is an expected performance of Meta systems and conducts root cause analysis to identify the solutions to fix them.

The assessment about Meta's enforcement accuracy on entities related to the Israel-Gaza conflict has been conducted in two parts, with slight changes to the company's sampling methodology to capture perceived instances of under-enforcement as well as over-enforcement. Throughout the process, sampled content was derived from review of viral content in relevant policy areas based on Viewer per Views (VPVs), a review of a subset of actioned viral content in those areas, content escalated through user reporting and captured

by Quality Assurance, and a samples of content escalated to Meta by external parties that were perceived to be over enforcement.

When internal teams come across an enforcement error, we work to either restore (or remove) the content based on additional review. This work is ongoing.

We plan to report back on our continuing efforts to address the recommendations made by Business for Social Responsibility in the Israel Palestine human rights due diligence, including any learnings from our current monitoring of the conflict.

RESPONSE DETAIL ENDS
REMAINDER OF PAGE INTENTIONALLY LEFT BLANK

Meta

Meta's Ongoing Efforts Regarding the Israel-Hamas War

October 13, 2023
Hebrew translation, *Arabic translation*

Update on December 5, 2023 at 9:00PM PT:

At the beginning of the war, we designated the October 7 attack by Hamas as a Terrorist Attack under our Dangerous Organization and Individuals policy. Consistent with that designation, we removed all content showing identifiable victims at the moment of the attack.

Following that, people began sharing this type of footage in order to raise awareness and condemn the attacks. Meta's goal is to allow people to express themselves while still removing harmful content. In turn, we began allowing people to post this type of footage within that context only, with the addition of a warning screen to inform users that it may be disturbing. If the user's intent in sharing the content is unclear, we err on the side of safety and remove it.

Under our Dangerous Organizations and Individuals policy, we continue to remove any imagery that is produced by a Dangerous Organization or Individual, unless it is clear that the user is sharing it in a news reporting or condemnation context, and no minors under thirteen years old are depicted.

Update on October 18, 2023 at 3:00AM PT:

After the terrorist attack by Hamas against Israel last week, and Israel's response in Gaza, our teams introduced a series of measures to address the spike in harmful and potentially harmful content spreading on our platforms. Our policies are designed to keep people safe on our apps while giving everyone a voice. We apply these policies equally around the world and there is no truth to the suggestion that we are deliberately suppressing voice. However, content containing praise for Hamas, which is designated by Meta as a Dangerous Organization, or violent and graphic content, for example, is not allowed on our platforms. We can make errors and that is why we offer an appeals process for people to tell us when they think we have made the wrong decision, so we can look into it.

Some additional updates and steps we are taking as this situation continues to unfold:

Fixing Bugs: We identified and fixed some bugs this past week.

- One impacting all Stories that re-shared Reels and Feed posts on Instagram, meaning they weren't showing up properly in people's Stories, leading to significantly reduced reach. This bug affected accounts equally around the globe – not only people trying to post about what's happening in Israel and Gaza – and it had nothing to do with the subject matter of the content. We fixed this bug as quickly as possible.
- Another prevented people from going Live on Facebook for a short time. This was also a global issue that was fixed within a few hours. We understand people rely on these tools and we're sorry to anyone who felt the impact of these issues.

Comment and Profile Settings:

- As a temporary measure to protect people in the region from potentially unwelcome or unwanted comments, we have:
    - Changed the default setting for who can comment on newly created public Facebook posts of people in the region to Friends and/or established followers only. Users globally can choose to use this setting and opt in or out at any time, and we are notifying people in the region with specific instructions on how to change this setting.
    - We've made it easier for people to bulk delete comments on their posts.
    - Disabled the feature that normally displays the first one or two comments under posts in Feed.
- We recently rolled out the Lock Your Profile tool in the region that allows people to lock their Facebook profile in one step. 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.
- Fundraising on Facebook and Instagram: Since October 7, people have raised more than $11.5 million for nonprofits on Facebook and Instagram to help with relief efforts in Israel and Palestine. This includes over 340,000 donations to 262 charities – providing disaster relief, ambulance and blood services, medical care and more.

Originally published on October 13, 2023 at 1:00AM PT:

Like many, we were shocked and horrified by the brutal terrorist attacks by Hamas, and our thoughts go out to civilians who are suffering in Israel and Gaza as the violence continues to unfold.

Since the terrorist attacks by Hamas on Israel on Saturday, and Israel's response in Gaza, expert teams from across our company have been working around the clock to monitor our platforms, while protecting people's ability to use our apps to shed light on important developments happening on the ground. The following are some of the specific steps we have taken:

Taking Action on Violating Content

- We quickly established a special operations center staffed with experts, including fluent Hebrew and Arabic speakers, to closely monitor and respond to this rapidly evolving situation in real time. This allows us to remove content that violates our Community Standards or Community Guidelines faster, and serves as another line of defense against misinformation.
- We continue to enforce our policies around Dangerous Organizations and Individuals, Violent and Graphic Content, Hate Speech, Violence and Incitement, Bullying and Harassment, and Coordinating Harm.
    - In the three days following October 7, we removed or marked as disturbing more than 795,000 pieces of content for violating these policies in Hebrew and Arabic.
    - As compared to the two months prior, in the three days following October 7, we have removed seven times as many pieces of content on a daily basis for violating our Dangerous Organizations and Individuals policy in Hebrew and Arabic alone.
- Hamas is designated by the US government as both a Foreign Terrorist Organisation and Specially Designated Global Terrorists. It is also designated under Meta's Dangerous Organizations and Individuals policy. This means Hamas is banned from our platforms, and we remove praise and substantive support of them when we become aware of it, while continuing to allow social and political discourse — such as news reporting, human rights related issues, or academic, neutral and condemning discussion.