OVERTURNED

2023-049-IG-UA

Al-Shifa Hospital

The Board overturns Meta's original decision to remove the content from Instagram. It finds that restoring the content to the platform, with a "mark as disturbing" warning screen, is consistent with Meta's content policies, values and human-rights responsibilities.

Policies and topics

- Safety, Violence, War and conflict
- Violent and graphic content

Region and countries

- Middle East and North Africa

Platform

- Instagram

Attachments

Hebrew translation.pdf

In the weeks following the publication of this decision, we will upload a translation in Hebrew here and an Arabic translation will become available through the 'language' tab accessed in the menu at the top of this screen.

לקריאת החלטה זו בעברית יש ללחוץ כאן.

1. Summary

This case involves an emotionally powerful video of the aftermath of a strike on or near Al-Shifa hospital in Gaza during Israel's ground offensive, with a caption condemning the attack. Meta's automated systems removed the post for violating its Violent and Graphic Content Community Standard. After unsuccessfully contesting this decision with Meta, the user appealed to the Oversight Board. After the Board identified the case for review, Meta reversed its decision and restored the content with a warning screen. The Board holds that the original decision to remove the content did not comply with Meta's content policies or the company's human-rights responsibilities. The Board approves the decision to restore the content with a warning screen but disapproves of the associated demotion of the content barring it from recommendations. This case, together with Hostages Kidnapped From Israel (2023-050-FB-UA), are the Board's first cases decided under its expedited review procedures.

2. Context and Meta's Response

On October 7, 2023, Hamas, a designated Tier 1 organization under Meta's Dangerous Organizations and Individuals Community Standard, led unprecedented terrorist attacks on Israel from Gaza that killed an estimated 1,200 people and resulted in roughly 240 people being taken hostage (Ministry of Foreign Affairs, Government of Israel). Israel immediately undertook a military campaign in Gaza in response to the attacks. Israel's military action has killed more than 18,000 people in Gaza as of mid-December 2023 (UN Office for the Coordination of Humanitarian Affairs, drawing on data from the Ministry of Health in Gaza), in a conflict where both sides have been accused of violating international law. Both the terrorist attacks and Israel's subsequent military actions have been the subjects of intense worldwide publicity, debate, scrutiny, and controversy, much of which has taken place on social media platforms, including Instagram and Facebook.

Meta immediately designated the events of October 7 a terrorist attack under its Dangerous Organizations and Individuals policy. Under its Community Standards, this means that Meta would remove any content on its platforms that "praises, substantively supports or represents" the October 7 attacks or the perpetrators of them.

In reaction to an exceptional surge in violent and graphic content being posted to its platforms following the terrorist attacks and military response, Meta put in place several temporary measures, including a reduction of the confidence thresholds for its Graphic and Violent Content automatic classification system (classifier) to identify and remove content. Meta informed the Board that these measures applied to content originating in Israel and Gaza across all languages. The changes to these classifiers increased the automatic removal of content where there was a lower confidence score for the content violating Meta's policies. In other words, Meta used its automated tools more aggressively to remove content that might violate its policies. Meta did this to prioritize its value of safety, with more content removed than would have occurred under the higher confidence threshold in place prior to October 7. While this reduced the likelihood that Meta would fail to remove violating content that might otherwise evade detection or where capacity for human review was limited, it also increased the likelihood of Meta mistakenly removing non-violating content related to the conflict.

When escalation teams assessed videos as violating its Violent and Graphic Content, Violence and Incitement and Dangerous Organizations and Individuals policies, Meta relied on Media Matching Service banks to automatically remove matching videos. This approach raised the concern of over-enforcement, including people facing restrictions on or suspensions of their accounts following multiple violations of Meta's content policies (sometimes referred to as "Facebook jail"). To mitigate this concern, Meta withheld " strikes" that would ordinarily accompany content post removals that occur automatically based on Media Matching Service banks (as Meta announced in its newsroom post).

Meta's changes in the classifier confidence threshold and its strike policy are limited to the Israel-Gaza conflict and intended to be temporary. As of December 11, 2023, Meta had not restored confidence thresholds to pre-October 7 levels.

3. Case Description

The content in this case involves a video posted on Instagram in the second week of November, showing what appears to be the aftermath of a strike on or near Al-Shifa Hospital in Gaza City during Israel's ground offensive in the north of the Gaza Strip. The Instagram post in this case shows people, including children, lying on the ground lifeless or injured and crying. One child appears to be dead, with a severe head injury. A caption in Arabic and

English below the video states that the hospital has been targeted by the "usurping occupation," a reference to the Israeli army, and tags human rights and news organizations.

Meta's Violent and Graphic Content Community Standard, which applies to content on Facebook and Instagram, prohibits "[v]ideos of people or dead bodies in non-medical settings if they depict … [v]isible internal organs." At the time of posting, the policy allowed "[i]magery that shows the violent death of a person or people by accident or murder," provided that such content was placed behind a "mark as disturbing" warning screen and was only visible to people over the age of 18. This rule was updated on November 29, after the content in this case was restored, to clarify that the rule applies to the "moment of death or the aftermath" as well as imagery of "a person experiencing a life-threatening event."

Meta's automated systems removed the content in this case for violating the Violent and Graphic Content Community Standard. The user's appeal against that decision was automatically rejected because Meta's classifiers indicated "a high confidence level" that the content was violating. The user then appealed Meta's decision to the Oversight Board.

Following the Board's selection of this case, Meta said it could not conclusively determine that the video showed visible internal organs. Meta therefore concluded that it should not have removed this content, though it was on the "borderline" of violating. Meta further explained that even if internal organs had been visible, the post should have been kept up with a "mark as disturbing" warning screen as it was shared to raise awareness. The company reiterated that, in line with the Graphic and Violent Content policy rationale, such content is permitted when shared to raise awareness "about important issues such as human-rights abuses, armed conflicts or acts of terrorism."

Meta therefore reversed its original decision and restored the content with a warning screen. The warning screen tells users that the content may be disturbing. Adult users can click through to see the post, but Meta removes these posts from the feeds of Instagram users under 18 and also removes them from recommendations to adult Instagram users. Meta also added a separate instance of the same video to a Media Matching Service bank, so other videos identical to this one would be automatically kept up with a warning screen and would only be visible to people over the age of 18.

4. Justification for Expedited Review

The Oversight Board's Bylaws provide for expedited review in "exceptional circumstances, including when content could result in urgent real-world consequences," and decisions are binding on Meta (Charter, Art. 3, section 7.2; Bylaws, Art. 2, section 2.1.2). The expedited process precludes the level of extensive research, external consultation or public comments that would be undertaken in cases reviewed on ordinary timelines. The case is decided on the information available to the Board at the time of deliberation and is decided by a five-member panel without a full vote of the Board.

The Oversight Board selected this case and one other case, Hostages Kidnapped From Israel (2023-050-FB-UA), because of the importance of freedom of expression in conflict situations, which has been imperiled in the context of the Israel-Hamas conflict. Both cases are representative of the types of appeals users in the region have been submitting to the Board since the October 7 attacks and Israel's subsequent military action. Both cases fall within the Oversight Board's crisis and conflict situations priority. Meta's decisions in both cases meet the standard of "urgent real-world consequences" to justify expedited review, and accordingly the Board and Meta agreed to proceed under the Board's expedited procedures.

In its submissions to the Board, Meta recognized that "the decision on how to treat this content is difficult and involves competing values and trade-offs," welcoming the Board's input on this issue.

5. User Submissions

The author of the post stated in their appeal to the Board that they did not incite any violence, but shared content showing the suffering of Palestinians, including children. The user added that the removal was biased against the suffering of the Palestinians. The user was notified of the Board's review of their appeal.

6. Decision

While members of the Board have disagreements about Israel's military response and its justification, they unanimously agree on the importance of Meta respecting the right to freedom of expression and other human rights of all those impacted by these events, and their ability to communicate in this crisis.

The Board overturns Meta's original decision to remove the content from Instagram. It finds that restoring the content to the platform, with a "mark as disturbing" warning screen, is consistent with Meta's content policies, values and human-rights responsibilities. However, the Board also concludes that Meta's demotion of the restored content, in the form of its exclusion from the possibility of being recommended, does not accord with the company's responsibilities to respect freedom of expression.

6.1 Compliance With Meta's Content Policies

The Board agrees with Meta that it is difficult to determine whether the video in this case shows "[v]isible internal organs." Given the context of this case, where there is exceptionally high public interest value in protecting access to information and providing avenues for raising awareness of the impact of the conflict, content that is "on the borderline" of violating the Violent and Graphic Content policy should not be removed. As the content includes imagery that shows a person's violent death, depicting a bloody head injury, Meta should have applied a warning screen and made it available only to people over the age of 18 in line with its policies.

The Board also agrees with Meta's subsequent determination that even if this video had included visible internal organs, the post's language condemning or raising awareness of the violence also means that it should have been left up with a "mark as disturbing" warning screen, and not be available to users under 18. The Community Standard does not provide for warning screens in relation to the applicable policy line ("[v]ideos of people or dead bodies in a medical setting if they depict […] [v]isible internal organs"). In the Sudan Graphic Video case, the Board explained that Meta instructs reviewers to follow the letter of its "do not post" policies. The rationale states that "[i]n the context of discussions about important issues such as human-rights abuses, armed conflicts or acts of terrorism, we allow graphic content (with some limitations) to help people to condemn and raise awareness about these situations." The Community Standard rule, however, prohibits all videos depicting "visible internal organs" in a non-medical context, without providing reviewers the option of adding a warning screen where the policy rationale exception is engaged. Meta's automated systems do not appear to be configured to apply warning screens to videos depicting graphic content where there is context condemning or raising awareness of the violence. It is also not clear that where this context is present, the applicable classifiers would be able to send the content to human reviewers for further assessment.

6.2 Compliance With Meta's Human-Rights Responsibilities

In line with its human-rights responsibilities, Meta's moderation of violent and graphic content must respect the right to freedom of expression, which includes freedom to seek, receive and impart information (Art. 19, para. 2, International Covenant on Civil and Political Rights (ICCPR)). As the Board stated in the Armenian Prisoners of War Video case, the protections for freedom of expression under Article 19 of the International Covenant on Civil and Political Rights (ICCPR) "remain engaged during armed conflicts, and should continue to inform Meta's human rights responsibilities, alongside the mutually reinforcing and complementary rules of international humanitarian law that apply during such conflicts." The UN Guiding Principles on Business and Human Rights impose a heightened responsibility on businesses operating in a conflict setting ("Business, human rights and conflict-affected regions: towards heightened action," A/75/212).

The Board has emphasized in previous cases that social media platforms like Facebook and Instagram are an important vehicle for transmitting in real-time information about violent events, including news reporting (see e.g. Mention of the Taliban in News Reporting). They play an especially important role in contexts of armed conflicts, especially where there is limited access for journalists. Furthermore, content depicting violent attacks and human-right abuses is of great public interest (See Sudan Graphic Video).

When restrictions on expression are imposed by a state, under international human rights law they must meet the requirements of legality, legitimate aim and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's voluntary human-rights commitments, both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. In doing so, the Board attempts to be sensitive to how those rights may be different as applied to a private social media company than as applied to a government. Nonetheless, as the UN Special Rapporteur on freedom of expression has stated that while companies do not have the obligations of governments "their impact is of a sort that requires them to assess the same kind of questions about protecting their right to freedom of expression" (report A/74/486, para. 41.).

Legality requires that any restriction on freedom of expression should be accessible and clear enough to provide guidance as to what is permitted and what is not. The Board has previously expressed concern that the rules of the Violence and Graphic Content Community

Standard do not align fully with the rationale of the policy, which sets out the aims of the policy (See Sudan Graphic Video and Video After Nigeria Church Attack). The Board reiterates the importance of recommendations no. 1 and no. 2 in the Sudan Graphic Video case, which called on Meta to amend its Violent and Graphic Content Community Standard to allow videos of people or dead bodies when shared for the purpose of raising awareness of or documenting human-rights abuses (that case concerned visible dismemberment.) Meta has conducted a policy development process in response to these recommendations and intends to report on its progress in its next quarterly update to the Board. In the Board's view, this recommendation should apply to the rules for videos showing visible internal organs, and specifically provide for warning screens as an enforcement measure where the raising awareness (including factual reporting) and condemnation exception is engaged.

Under Article 19, para. 3 of the ICCPR, expression may be restricted for a defined and limited list of reasons. The Board has previously found that the Violent and Graphic Content policy legitimately aims to protect the rights of others, including the privacy of the depicted individual (See Sudan Graphic Video and Video After Nigeria Church Attack). The present case demonstrates, additionally, that restricting access to the content for people under 18 served the legitimate aim of protecting the right to health of minors (Convention on the Rights of the Child, Article 24).

The principle of necessity and proportionality provides that any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; [and] they must be proportionate to the interest to be protected" (General Comment No. 34, para. 34).

The Board has previously found in relation to violent and graphic content that a warning screen "does not place an undue burden on those who wish to see the content while informing others about the nature of the content and allowing them to decide whether to see it or not" (See Sudan Graphic Video). Warning screens prevent users from unwillingly seeing potentially disturbing content. Victims' rights are further protected by Meta's policy to remove videos and photos that show the violent death of someone (or its immediate aftermath) when a family member requests this. The content in this case can be distinguished from that in the Russian Poem case, which showed a still image of a body lying on the ground at long range, where the face of the victim was not visible, and where there were no clear visual indicators of violence. Applying a warning screen in that case was inconsistent with Meta's guidance to reviewers and not a necessary or proportionate

restriction on expression. The content in this case is more similar to the content in the Video After Nigeria Church Attack decision, showing dead and injured people at close range, with very clear visual indicators of violence.

In this case, the depiction of injured and lifeless children makes the video especially distressing. In circumstances like these, providing users with the choice of whether to see disturbing content is a necessary and proportionate measure (see also Armenian Prisoners of War Video).

The Board finds that excluding content raising awareness of potential human-rights abuses and violations of the laws of war, conflicts or acts of terrorism from recommendations reaching adults is not a necessary or proportionate restriction on freedom of expression, in view of the very high public interest in such content. Warning screens and removal from recommendations serve separate functions, and should in some instances be decoupled, in particular in crisis situations. Recommendations on Instagram are generated by automated systems that suggest content to users based on users' predicted interests. Removing content from recommendation systems means reducing the reach that this content would otherwise get. The Board finds this practice interferes with freedom of expression in disproportionate ways in so far as it applies to content that is already limited to adult users and that is posted to raise awareness, condemn, or report on matters of public interest such as the development of a violent conflict.

The Board recognizes that immediate responses to a crisis can require exceptional temporary measures, and that in some contexts it is legitimate to prioritize safety concerns and to temporarily and proportionally place greater restrictions on freedom of expression. Some of these are outlined, for example, in the commitments to counter "terrorist and violent extremist content" established in the Christchurch Call. However, the Board notes that the Christchurch Call emphasizes the need to respond to such content in a manner consistent with human rights and fundamental freedoms. The Board believes that safety concerns do not justify erring on the side of removing graphic content that has the purpose of raising awareness about or condemning potential war crimes, crimes against humanity, or grave violations of human rights. Such restrictions can even obstruct information necessary for the safety of people on the ground in those conflicts.

Measures such as not imposing strikes do help to mitigate the potentially disproportionate adverse effects of enforcement errors due to emergency measures like reducing confidence

thresholds for removal of content during conflict situations. They are, however, not sufficient to protect the ability of users to share content that raises awareness about potential human-rights abuses and violations of humanitarian law, and other critical information in conflict situations.

The Board has repeatedly highlighted the need to develop a principled and transparent framework for content moderation during crises and in conflict zones (See Haitian Police Station Video and Tigray Communication Affairs Bureau). It is precisely at times of rapidly changing conflict that large social media companies must devote the resources necessary to ensure that freedom of expression is not needlessly curtailed. At such times, journalistic sources are often subject to physical and other attacks, making news reporting by ordinary citizens on social media especially essential.

The Board has also previously observed that in contexts of war or political unrest, there will be more graphic and violent content captured by users and shared on the platform for the purpose of raising awareness of or documenting abuses (See Sudan Graphic Video). In contexts such as the Israel-Gaza conflict, where there is an alarming number of civilians killed or injured, a high proportion of children among them, amid a worsening humanitarian crisis, these kinds of allowances are especially important. While acknowledging Meta's ongoing policy development process on its Violent and Graphic Content policy, the Board would expect Meta to be ready to rapidly deploy temporary measures to allow this kind of content with warning screens, and not remove it from recommendations.

The Board notes that the situation in Gaza at the time this content was posted did not engage the same set of challenges for Meta as the October 7 attacks. In Gaza, there have been difficulties in attaining information from people on the ground, while journalist access to the territory is limited and Internet connectivity has been disrupted. Moreover, unlike the early aftermath of the October 7 attacks, the Gaza situation presented in this case did not involve terrorists using social media to broadcast their atrocities. In the context of armed conflict, by contrast, Meta should be ensuring that its actions are not making it more difficult for people to share content that provides information that raises awareness about harms against civilians, and may be relevant to determining whether violations of international humanitarian law and international human rights law have occurred. The question of whether content was shared to raise awareness of or condemn events on the ground should be the starting point for any reviewer assessing such content, and Meta's automated systems should be designed to avoid incorrectly removing content that should benefit from applicable exceptions.

This case further illustrates that insufficient human oversight of automated moderation in the context of a crisis response can lead to erroneous removal of speech that may be of significant public interest. Both the initial decision to remove this content as well as the rejection of the user's appeal were taken automatically based on a classifier score, without any human review. This, in turn, may have been exacerbated by Meta's crisis response of lowering the removal threshold of content under the Violent and Graphic Content policy following the October 7 attacks. This means that even if the classifier gives a relatively lower score to the likelihood of violation than would usually be required, Meta removes that content.

For Meta to employ its automated systems in a manner compatible with its human-rights commitments, the Board reminds Meta of recommendation no. 1 in the Colombia Police Cartoon case. In that case, the Board called on Meta to ensure that content with high rates of appeal and high rates of successful appeal be reassessed for possible removal from its Media Matching Service banks. In response to this recommendation, Meta established a designated working group committed to governance improvements across its Media Matching Service banks (See Meta's most recent updates on this here). The Board notes that it is important for this group to pay particular attention to the use of Media Matching Services in the context of armed conflicts. In the Breast Cancer Symptoms and Nudity case (recommendation no. 3 and no. 6), the Board recommended that Meta inform users when automation is used to take enforcement action against their content, and to disclose data on the number of automated removal decisions per Community Standard and the proportion of those decisions subsequently reversed following human review. This is particularly important when the confidence thresholds for content that is likely violating have reportedly been significantly lowered. The Board urges Meta to make further progress in the implementation of recommendation no. 6 and share evidence of implementation with the Board for recommendation no. 3.

Restrictions on freedom of expression must be non-discriminatory, including on the basis of nationality, ethnicity, religion or belief, or political or other opinion (Article 2, para. 1, and Article 26, ICCPR). Discriminatory enforcement of the Community Standards undermines this fundamental aspect of freedom of expression. In the Shared Al Jazeera Post case, the Board raised serious concerns that errors in Meta's content moderation in Israel and the Occupied Palestinian Territories may be unequally distributed, and called for an independent investigation (Shared Al Jazeera Post decision, recommendation no. 3). The Business for Social Responsibility (BSR) Human Rights Impact Assessment, which Meta commissioned in response to that recommendation, identified "various instances of unintentional bias where

Meta policy and practice, combined with broader external dynamics, does lead to different human-rights impacts on Palestinian and Arabic speaking users." The Board encourages Meta to deliver on commitments it made in response to the BSR report.

Finally, Meta has a responsibility to preserve evidence of potential human-rights violations and violations of international humanitarian law, as also recommended in the BSR report (recommendation 21) and advocated by civil society groups. Even when content is removed from Meta's platforms, it is vital to preserve such evidence in the interest of future accountability (See Sudan Graphic Video and Armenian Prisoners of War Video). While Meta explained that it retains all content that violates its Community Standards for a period of one year, the Board urges that content specifically related to potential war crimes, crimes against humanity, and grave violations of human rights be identified and preserved in a more enduring and accessible way for purposes of longer-term accountability. The Board notes that Meta has agreed to implement recommendation no. 1 in the Armenian Prisoners of War Video case. This called on Meta to develop a protocol to preserve and, where appropriate, share with competent authorities, information to assist in investigations and legal processes to remedy or prosecute atrocity crimes or grave human-rights violations. Meta has informed the Board that it is in the final stages of developing a "consistent approach to retaining potential evidence of atrocity crimes and serious violations of international human rights law" and expects to provide the Board with a briefing about its approach soon. The Board expects Meta to fully implement the above recommendation.

***Procedural Note:**

The Oversight Board's expedited decisions are prepared by panels of five members and are not subject to majority approval of the full Board. Board decisions do not necessarily represent the personal views of all members.