UPHELD

2023-004-FB-MR

Armenian prisoners of war video

The Oversight Board has upheld Meta's decision to leave up a Facebook post that included a video depicting identifiable prisoners of war and add a "mark as disturbing" warning screen to the video.

Policies and topics

War and conflict, Freedom of expression, Safety

Coordinating harm and publicizing crime

Region and countries

Europe

Armenia, Azerbaijan

Platform

Facebook

Attachments

Public comments appendix

Armenian translation

Azerbaijani translation

This decision is also available in Armenian and Azerbaijani.

Այս որոշումը հայերեն կարդալու համար սեղմեք այստեղ:

Bu qərarı azərbaycan dilində oxumaq üçün buraya kliklәyin.

Case summary

The Oversight Board has upheld Meta's decision to leave up a Facebook post that included a video depicting identifiable prisoners of war and add a "mark as disturbing" warning screen to the video. The Board found that Meta correctly applied a newsworthiness allowance to the post, which would have otherwise been removed for violating its Coordinating Harm and Promoting Crime Community Standard. However, the Board recommends that Meta strengthen internal guidance around reviewing this type of content and develop a protocol for preserving and sharing evidence of human rights violations with the appropriate authorities.

About the case

In October 2022, a Facebook user posted a video on a page that identifies itself as documenting alleged war crimes committed by Azerbaijan against Armenians in the context of the Nagorno-Karabakh conflict. This conflict reignited in September 2020 and escalated into fighting in Armenia in September 2022, leaving thousands dead, and hundreds of people missing.

The video begins with a user-inserted age warning that it is only suitable for people over the age of 18, and an English text, which reads "Stop Azerbaijani terror. The world must stop the aggressors." The video appears to depict a scene where prisoners of war are being captured.

It shows several people who appear to be Azerbaijani soldiers searching through rubble, with their faces digitally obscured with black squares. They find people in the rubble who are described in the caption as Armenian soldiers, whose faces are left unobscured and identifiable. Some appear to be injured, others appear to be dead. The video ends with an unseen person, potentially the person filming, continuously shouting curse words and using abusive language in Russian and Turkish at an injured soldier sitting on the ground.

In the caption, which is in English and Turkish, the user states that the video depicts Azerbaijani soldiers torturing Armenian prisoners of war. The caption also highlights the July 2022 gas deal between the European Union and Azerbaijan to double gas imports from Azerbaijan by 2027.

Key findings

The Board finds that although the content in this case violates the Coordinating Harm and Promoting Crime Community Standard, Meta correctly applied the newsworthiness allowance to allow the content to remain on Facebook, and the contents of the video required a "mark as disturbing" warning screen under the Violent and Graphic Content Community Standard. These decisions were consistent with Meta's values and human rights responsibilities.

The case raises important questions about Meta's approach to content moderation in conflict situations, where revealing identities and locations of prisoners of war could undermine their dignity or expose them to immediate harm. Concerns regarding human dignity are acute in situations where prisoners are shown in degrading or inhumane circumstances. At the same time, such exposure can inform public debate and raise awareness of potential mistreatment, including violations of international human rights and international humanitarian law. It can also build momentum for action that protects rights and ensures accountability. Meta is in a unique position to assist in the preservation of evidence that may be of relevance in prosecuting international crimes and supporting human rights litigation.

The scale and speed at which imagery of prisoners of war can be shared via social media complicates the task of resolving these competing interests. Given the acute harms and risks facing prisoners of war, the Board finds that Meta's default rule prohibiting the posting of information that could reveal the identities or locations of prisoners of war is consistent with the company's human rights responsibilities under the UN Guiding Principles of Business and Human Rights (UNGPs, commentary to Principle 12). These responsibilities are heightened during armed conflict and must be informed by the rules of international humanitarian law. The Board agrees with Meta that the public interest value in keeping the content on the platform with a warning screen outweighed the risk to the safety and dignity of the prisoners of war.

The Oversight Board's decision

The Oversight Board upholds Meta's decision to leave the post on Facebook with a "mark as disturbing" warning screen.

The Board also recommends that Meta:

- 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.
- Provide further guidance to reviewers and escalation teams to better inform the newsworthiness of escalation and assessment of content revealing the identity or locations of prisoners of war.
- Add an example of content that revealed the identity or location of prisoners of war but was left up due to the public interest, to its public explanation of the newsworthiness allowance in the Transparency Center, in order to provide greater clarity to users.
- Publicly share the protocol on evidence preservation related to atrocity crimes and grave human rights violations.

* Case summaries provide an overview of the case and do not have precedential value.

Full case decision

1. Decision summary

The Oversight Board upholds Meta's decision to leave up a Facebook post that included a video depicting identifiable prisoners of war, and adding a "mark as disturbing" warning screen to the video. The Board found that Meta correctly applied the newsworthiness allowance to the content, which it would have otherwise removed for violating the Community Standard on Coordinating Harm and Promoting Crime for revealing the identity of prisoners of war in the context of an armed conflict. The "mark as disturbing" warning screen was required under the Community Standard on Graphic and Violent Content. These decisions were consistent with Meta's values and human rights responsibilities.

The case, which Meta referred to the Board, raises important questions about the company's approach to content moderation in conflict situations, where the disclosure of the identities or locations of prisoners of war could expose them to immediate harm or affect their dignity. Concerns regarding human dignity may be acute in situations where prisoners are shown as defenseless or in humiliating circumstances, engaging their right to life, security, and privacy, and right to be free from torture, inhuman and degrading treatment as well as their families' rights to privacy and security. At the same time, such exposure can also inform public debate and raise awareness of potential mistreatment, including violations of international human rights and humanitarian law. It can also build momentum for action that protects rights and ensures accountability. Meta is also in a unique position to assist in the preservation of evidence that may be of use in the prosecution of international crimes and in support of human rights litigation, whether the content is removed or left up.

The scale and speed at which imagery of prisoners of war can be shared via social media complicates the task of resolving these competing interests. Given the acute harms and risks facing prisoners of war, the Board finds that Meta's default rule prohibiting the posting of information that could reveal the identities or locations of prisoners of war is consistent with the company's human rights responsibilities under the UN Guiding Principles of Business and Human Rights (UNGPs, commentary to principle 12), which are heightened during armed conflict and must be informed by the rules of international humanitarian law. The Board agrees with Meta's assessment that the public interest value in keeping the content on the platform with a warning screen outweighed the risk to the safety and dignity of the prisoners of war. Keeping the content on the platform was necessary to ensure the public's right to know about severe wrongdoing, and in this specific context, to potentially prevent, mitigate and remedy severe human rights harms through public disclosure of wrongdoing.

The Board recommends that Meta provides further guidance to reviewers and escalation teams to ensure that content revealing the identity or locations of prisoners of war can be reviewed on a case-by-case basis by those with the necessary expertise. Meta should develop more granular criteria to guide assessments of newsworthiness in these cases, which should be shared transparently. The Board calls on Meta to preserve and, where appropriate, share with competent authorities, information to assist in investigations and legal processes to remedy or prosecute grave violations of international criminal, human rights and humanitarian law.

2. Case description and background

In October 2022, a Facebook user posted a video on a page that identifies itself as documenting alleged war crimes committed by Azerbaijan against Armenians in the context of the Nagorno-Karabakh conflict. This conflict reignited in September 2020 during the 44-day Second Nagorno-Karabakh war, and escalated into fighting in Armenia in September 2022, leaving thousands dead, and hundreds of people missing.

In the caption, which is in English and Turkish, the user states that the video depicts Azerbaijani soldiers torturing Armenian prisoners of war. The caption also calls attention to the July 2022 gas deal between the European Union and Azerbaijan to double gas imports from Azerbaijan by 2027 to reduce European reliance on Russian gas.

The video begins with a user-inserted age warning that it is only suitable for people over the age of 18, and an English text, which reads "Stop Azerbaijani terror. The world must stop the aggressors." The video shows soldiers in the process of being detained as prisoners of war. It shows several people who appear to be Azerbaijani soldiers searching through rubble; their faces have been digitally obscured with black squares. They find people in the rubble who are described in the caption as Armenian soldiers; their faces have been left unobscured and are identifiable. Some appear to be injured, others appear to be dead. They pull one soldier from the rubble, who cries out in pain. His face is visible, and he appears injured. The video ends with an unseen person, potentially the person filming, continuously shouting curse words and using abusive language in Russian and Turkish at an injured soldier sitting on the ground, telling him to stand up. The individual attempts to do so. The page the content was posted to has fewer than 1,000 followers. This content has been viewed fewer than 100 times, and received fewer than 10 reactions. It has not been shared, or been reported as violating, by any user.

Meta informs the Board it was monitoring the situation as the conflict was ongoing. Meta's Global Operations team coordinated with Meta's security team to conduct risk monitoring that involved monitoring of external signals (such as news and social media trends) related to the issue. During the monitoring, the security team found Twitter posts that showed a video of Azerbaijani soldiers torturing Armenian prisoners of war circulating online, and then identified the same video on Facebook in this case. The security team sent the post on Facebook for additional review, a process Meta describes as an "escalation." When content is escalated, it is

sent to additional teams within Meta for policy and safety review. In this case, Meta's Global Operations team decided to escalate the content further to Meta's policy teams for newsworthiness review.

Upon review, within two days of the content being posted, Meta issued a newsworthiness allowance, which permits content on Meta's platforms that might otherwise violate its policies if the public interest in the content outweighs the risk of harm. The newsworthiness allowance can only be applied by specialist teams within Meta after content has been escalated for additional layers of review.

As part of escalated review by Meta's policy teams, a "marked as disturbing" warning screen was applied under the Violent and Graphic Content policy, and the content was added to a Graphic Violence Media Matching Service (MMS) Bank that automatically places a warning screen over the video and identical videos identified on the platform. However, due to a combination of technical and human errors, this failed, and had to be completed manually about one month later.

Meta referred this case to the Board, stating that it demonstrates the challenge required "to balance the value of raising awareness of these issues against the potential harm caused by revealing the identity of prisoners of war." Meta asked the Board to consider whether Meta's decision to allow the content represents an appropriate balancing of its values of "Safety," "Dignity," and "Voice," and is consistent with international human rights principles.

3. Oversight Board authority and scope

The Board has authority to review decisions that Meta submits for review (Charter Article 2, Section 1; Bylaws Article 2, Section 2.1.1).

The Board may uphold or overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

4. Sources of authority and guidance

The following standards and precedents informed the Board's analysis in this case:

I. Oversight Board decisions:

The most relevant previous decisions of the Oversight Board include:

- "India sexual harassment video" (2022-012-IG-MR)
- "Video after Nigeria church attack" (2022-011-IG-UA)
- "Russian poem" (2022-008-FB-UA)
- "Sudan graphic video" (2022-002-FB-MR)
- "Colombia protests" (2021-010-FB-UA)
- "Former President Trump's suspension" (2021-001-FB-FBR)

II. Meta's content policies:

The Coordinating Harm and Promoting Crime Community Standard states under the heading "policy rationale" that "[i]n an effort to prevent and disrupt offline harm and copycat behaviour, we prohibit people from facilitating, organizing, promoting or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals." It further states "we allow people to [...] draw attention to harmful or criminal activity that they may witness or experience as long as they do not advocate for or coordinate harm." Under a rule added on May 4, 2022, beneath the heading "additional information and/or context to enforce," Meta specifically prohibits "content that reveals the identity or location of a prisoner of war in the context of an armed conflict by sharing their name, identification number and/or imagery" and does not enumerate any specific exceptions to this rule.

The Violent and Graphic Content Community Standard states under the heading "policy rationale" that it exists to "protect users from disturbing imagery." The policy further specifies that "imagery that shows the violent death of a person or people by accident or murder" and "imagery that shows acts of torture committed against a person or people" is placed behind a warning screen so that "people are aware that the content may be disturbing," and only adults aged 18 and over are able to view the content. The "do not post" section of the rules explains

that users cannot post sadistic remarks towards imagery that requires a warning screen under the policy.

The Board's analysis was informed by the Meta's commitment to " Voice," which the company describes as "paramount," and its values of "Safety," "Privacy" and "Dignity."

The newsworthiness allowance is a general policy exception that can potentially be applied across all policy areas within the Community Standards, including to the rule on prisoners of war. The newsworthiness allowance is explained under Meta's "commitment to voice." It allows otherwise violating content to be kept on the platform if the public interest value in doing so outweighs the risk of harm. According to Meta's approach to newsworthy content, which is linked from the introduction to the Community Standards, such assessments are made only in "rare cases," following escalation to the Content Policy Team. This team assesses whether the content in question surfaces an imminent threat to public health or safety, or gives voice to perspectives currently being debated as part of a political process. This assessment considers country-specific circumstances, including if the country is at war. While the identity of the speaker is a relevant consideration, the allowance is not limited to content that is posted by news outlets.

III. Meta's human rights responsibilities

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. In 2021, Meta announced its Corporate Human Rights Policy, where it reaffirmed its commitment to respecting human rights in accordance with the UNGPs. Significantly, the UNGPs 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's analysis of Meta's human rights responsibilities in this case was informed by the following international standards, including in the field of international humanitarian law (also known as 'the law of armed conflict'):

- The right to freedom of expression: Article 19, International Covenant on Civil and Political Rights (ICCPR), General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression reports A/68/362 (2013), A/HRC/38/35

freedom of opinion and expression, reports A/68/362 (2013); A/HRC/38/35 (2018); A/74/486 (2019); A/77/288 (2022).

- The right to life: Article 6, ICCPR.
- The right to be free from torture, inhuman or degrading treatment: Article 7, ICCPR.
- The right to privacy: Article 17, ICCPR.
- Protection of prisoners of war from insults and public curiosity: Article 13, para. 2 Convention III relative to the Treatment of Prisoners of War (Geneva Convention III), Geneva Convention (III) Commentary, International Red Cross Committee (ICRC), 2020 (Geneva Convention III Commentary).

5. User submissions

Following Meta's referral and the Board's decision to accept the case, the user was sent a message notifying them of the Board's review and providing them with an opportunity to submit a statement to the Board. The user did not submit a statement.

6. Meta's submissions

Meta referred this case to the Board "because it highlights the challenges [Meta] face[s] when determining whether the value of content's newsworthiness outweighs the risk of harm in the context of war and violence." While the content violated Meta's Coordinating Harm and Promoting Crime policy, the newsworthiness allowance was applied "in order to raise awareness of the violence against prisoners of war" in the conflict between Azerbaijan and Armenia. Meta said the case was "significant" due to its relationship to an ongoing military conflict, and "difficult" because it requires "balancing the value of raising awareness of these issues against the potential harm caused by revealing the identity of prisoners of war."

Meta explains that it prohibits "outing" prisoners of war, but its escalation teams require "additional context" to enforce that rule. Meta added this rule to reflect the fact that they take the safety and dignity of prisoners of war seriously, given the risks of their platforms being used to expose them to public curiosity, noting both freedom of expression principles from Article 19 of the ICCPR, and the guidance of the Geneva Conventions. The Board understands this to mean that for such content to be removed, in-house investigation teams would need to flag it for review to Meta's internal teams, or at-scale reviewers would need to escalate the content to

those teams. These teams are able to consider contextual factors outside of the content to decide an enforcement action. Content would only be removed through automation if it is identical or near identical to content Meta's internal teams have already assessed as violating and added to media matching banks.

In response to the Board's questions, Meta confirmed that the additional context it considered to identify the content as violating in this case included (i) the uniforms confirmed the identifiable prisoners were Armenian soldiers, and (ii) knowledge of the ongoing conflict between Azerbaijan and Armenia. "Prisoner of war" is defined in internal guidance for reviewers called the Known Questions as "a member of the armed forces who has been captured or fallen into the hands of an opposing power during or immediately after an armed conflict." In the Internal Implementation Standards accompanying this rule, Meta explains that content which exposes a prisoner of war's identity or location by sharing either the first name, last name, their identification number, and/or imagery identifying their face, even if it is shared with "a condemning or raising awareness context," should be removed.

In response to the Board's questions, Meta noted that the Crisis Policy Protocol, was not used during the second Nagorno-Karabakh war or the ongoing border clashes between Azerbaijan and Armenia. The Crisis Policy Protocol provides Meta review teams with additional opportunities to apply escalation-only policies, and was created in response to the Board's recommendation in the "former President Trump's suspension" case. Meta explained that as there is a permanent policy to remove content outing prisoners of war, the standard process of the internal escalations team removing this content would not have changed had the protocol been activated.

On applying the newsworthiness allowance, Meta says it conducted a balancing test that weighs the public interest against the risk of harm, considering several factors: (i) whether the content surfaces imminent threats to public health or safety; (ii) whether the content gives voice to perspectives currently being debated as part of a political process; (iii) the nature of the speech, including whether it relates to governance or politics; (iv) the political structure of the country, including whether it has a free press; and (v) other country-specific circumstances (for example, whether there is an election underway, or the country is at war).

Meta recognized the graphic nature of this video and the risks that prisoners of war may face

when they are identified on social media. In response to the Board's questions, Meta acknowledged that family members and friends may be the target of ostracism, suspicion, or even violence when prisoners of war are identified or exposed. Meta also noted that this kind of imagery can have a variety of effects on civilians and people in the military, including reinforcing antagonism towards the other side and intensifying prejudice. Meta noted that a prisoner of war who is recorded criticizing their own nation or their military's conduct may be at higher risk of ostracism and reprisal upon their return home than prisoners who are shown being mistreated by the enemy. In the context of this conflict, Meta did not have evidence that videos of this kind were producing these negative impacts but did see evidence that international organizations were using such videos to increase pressure on Azerbaijan to end mistreatment of prisoners of war. Given the potential public interest value of this content in raising awareness and serving as evidence of possible war crimes, Meta concluded that, given its overall newsworthiness, removing this content would not be a proportionate action.

Meta also added that content that identifies witnesses, informants, hostages, or other detained people may be removed under the Coordinating Harm and Promoting Crime Community Standard, if public knowledge of the detention may increase risks to their safety. Content identifying or outing individuals may also be removed under the Privacy Violations policy when personally identifiable information is shared on Meta platforms. Finally, the Violent and Graphic Content policy would have applied even if the victims were not prisoners of war, as under that policy, Meta considers both the dignity and safety of the victims of violence and the fact that people may not want to see this content.

In response to the Board's questions, Meta also provided examples of how it applies the newsworthiness allowance to content identifying prisoners of war more broadly. For example, Meta informed the Board that it generally removes content that reveals the identity of prisoners of war in Ethiopia but makes a case-by-case newsworthy determination for some content. Factors considered in applying the newsworthiness allowance in previous cases include whether the content (i) reports on the capture of senior combatants, such as high-ranking officers or leaders of armed groups; (ii) reveals the identity of a prisoner when it is potentially in their interest to do so (e.g. when they have been reported missing); or (iii) raises awareness about potential human rights abuses. Meta also stated it had granted newsworthiness allowances to "leave some content up that shows Russian [prisoners of war] in Ukraine."

Meta also noted that it assesses content at its face value, unless authenticity is in question, or where there are indicators of manipulated media or where they have context that the information is false. In this case, Meta saw no indications that its misinformation policies were engaged.

The Board asked Meta 16 questions. Questions related to application of the newsworthiness allowance; factors in assessing context in Meta's decision; and the application of the Crisis Policy Protocol. Meta answered the 16 questions.

7. Public comments

The Oversight Board received 39 public comments relevant to this case. One comment was submitted from Asia Pacific and Oceania, three from Central and South Asia, 23 from Europe, four from the Middle East and North Africa and eight from the United States and Canada.

The submissions covered the following themes: background on the Nagorno-Karabakh conflict and recent escalations; application of international humanitarian law to the moderation of content revealing the identity or location of prisoners of war; concern about content on social media showing the faces of the prisoners of war; potential adverse and positive impacts that can result from leaving up or removing content depicting prisoners of war; independent mechanisms preserving potential evidence of international crimes; cooperation between social media companies, civil society organizations and international justice mechanisms; concern over verification of video content; operational/technical suggestions on how to keep the content on social media platforms and protect the safety and dignity of prisoners of war; potential of such content to assist in preventing further atrocities, and the public's right to know about mistreatment of prisoners of war.

To read public comments submitted for this case, please click here.

In April 2023, as part of ongoing stakeholder engagement, the Board consulted representatives of advocacy organizations, academics, inter-governmental organizations and other experts on issues relating to the moderation of content depicting prisoners of war. A roundtable was held under the Chatham House Rule. This focused on motivations, potential risk factors and advantages to posting content depicting identifiable prisoners of war and ways to balance the benefits of raising awareness of violence against prisoners of war against the potential harm

benefits of raising awareness of violence against prisoners of war against the potential harm caused by revealing their identity. The insights provided at this meeting were valuable, and the Board extends its appreciation to all participants.

8. Oversight Board analysis

The Board analyzed Meta's content policies, human rights responsibilities and values to determine whether this content should be kept up with a warning screen. The Board also assessed the implications of this case for Meta's broader approach to content governance, particularly in conflict and crisis situations.

The Board selected this case as an opportunity to assess Meta's policies and practices in moderating content that depicts identifiable prisoners of war. Additionally, the case allows the Board to examine Meta's compliance with its human rights responsibilities in crisis and conflict situations generally.

8.1 Compliance with Meta's content policies

The Board finds that while the content violates the Coordinating Harm and Promoting Crime Community Standard, Meta correctly applied the newsworthiness allowance to allow the content to remain on Facebook, and the contents of the video required a "mark as disturbing" warning screen under the Violent and Graphic Content Community Standard.

I. Content rules

Coordinating Harm and Promoting Crime

The Board finds that the content in this case exposed the identity of prisoners of war, through imagery in the video that showed the faces of detained Armenian soldiers. It therefore clearly violated the rule prohibiting such content in the Coordinating Harm and Promoting Crime Community Standard.

Acknowledging that this rule requires additional information and/or context to enforce, the Board agrees with Meta that the soldiers' uniforms indicated that the individuals with their faces visible were members of the Armenian armed forces. The context of the war indicated that these

soldiers were being detained by the opposing Azerbaijani armed forces, meeting the definition of "prisoner of war" contained in Meta's internal guidance to reviewers. This information was sufficient to find that the content was contrary to the rule prohibiting content revealing the identity of prisoners of war through the sharing of imagery.

Newsworthiness allowance

The Board finds that the public interest in the video outweighed the potential risks of harm, and that it was appropriate for escalation teams with access to expertise and additional contextual information, including cross-platform trends, to apply the newsworthiness allowance to keep the content on the platform. While it is not presumed that anyone's speech is inherently newsworthy, the assessment of the newsworthiness allowance accounts for various factors, including the country-specific circumstances and the nature of the speech and the speaker. In such cases, Meta should conduct a thorough review that weighs the public interest, including the public's right to know about serious wrongdoing and the potential to prevent, mitigate and remedy severe human rights harms through public disclosure of wrongdoing, against the risks of harm to privacy, dignity, security and voice, pursuant to the international human rights standards, as reflected in Meta's Corporate Human Rights Policy.

The application of the newsworthiness allowance in a situation as complex and fast-moving as an armed conflict requires a case-by-case contextual assessment to mitigate risks and secure the public's access to important information. The factors Meta identified in its assessment, detailed in Section 6, were all pertinent to assessing the potential for serious harm resulting from the display of the video against adverse impacts that could result from suppressing this kind of content. The absence of evidence of videos like this being used in this particular conflict to further mistreatment of detainees, taken together with clear trends of similar content being primarily available through social media and highly relevant to campaigns and legal proceedings for accountability of serious crimes, militated in favour of keeping the content on the platform.

The Board emphasizes that it is important that Meta has systems in place to gain the kind of highly context specific insights required to enable a rapid case-by-case assessment of potential harms, taking into account Meta's human rights responsibilities.

Violent and Graphic Content

Following its decision that the content should be left up under the newsworthiness allowance, the Board finds that the violent and graphic nature of the video justified the imposition of a "mark as disturbing" warning screen, which serves a dual function of warning users of the graphic nature of the content and limiting the ability to view the content to adults over the age of 18. Although Meta did not specify the policy line it relied upon to impose this screen, the Board finds two rules were engaged.

First, the video shows what appear to be dead bodies of Armenian soldiers lying in the rubble. While the internal guidelines to moderators exclude violence committed by one or more uniformed personnel preforming a police function, in which case a "mark as sensitive" warning screen would be applied, the internal guidelines further define "police function" as "maintaining public order by performing crowd control and/or detaining people" and clarifies that "war does not qualify as a police function." As the content concerns an armed conflict situation, the Board finds it was consistent with Meta's policies to add a "mark as disturbing" warning screen.

The Board notes the video further engaged a second policy line as it showed acts meeting Meta's definition of torture against people. For the purpose of Meta's policy enforcement, the internal guidelines to moderators define such "torture" imagery as (i) imagery of a person in a dominated or forcibly restrained position and any of the following: (a) there is an armament pointed at the person; (b) there is evidence of injury on the person; or (c) person is being subjected to violence; or (ii) imagery of a person subjected to humiliating acts. Meta further defines "dominated position" as "any position including where the victim is kneeling, cornered, or unable to defend themselves" and "forcibly restrained" as "being physically tied, bound, buried alive or otherwise held against one's will." Noting that Meta's definition of "torture" is much broader than the term as understood under international law, the Board finds that it was consistent with Meta's rules to apply the "mark as disturbing" screen and accompanying age-gating restrictions. In line with the internal guidance, there are sufficient indicators in the content that individuals are being held against their will as detainees, and are unable to defend themselves. Moreover, several detainees seemed to be injured, while others appeared to be deceased.

8.2 Compliance with Meta's human rights responsibilities

The Board finds that Meta's decision to leave up the content is consistent with Meta's human rights responsibilities, which are heightened in a situation of armed conflict.

Freedom of expression (Article 19 ICCPR)

Article 19, para. 2 of the ICCPR provides for broad protection of expression, including the right to access information. These protections 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, including to protect prisoners of war (General Comment 31, Human Rights Committee, 2004, para. 11; Commentary to UNGPs, Principle 12; see also UN Special Rapporteur's report on Disinformation and freedom of opinion and expression during armed conflicts, Report A/77/288, paras. 33-35 (2022); and OHCHR report on International legal protection of human rights in armed conflict (2011) at p. 59).

International humanitarian law provides specific guarantees for the treatment of prisoners of war, particularly prohibiting acts of violence or intimidation against prisoners of war as well as exposing them to insults and public curiosity (Article 13, para. 2 of the Geneva Convention (III)). In a situation of armed conflict, the Board's freedom of expression analysis is informed by the more precise rules in international humanitarian law. The ICRC commentary to Article 13 explains that "being exposed to 'public curiosity' as a prisoner of war, even when such exposure is not accompanied by insulting remarks or actions, is humiliating in itself and therefore specifically prohibited [...] irrespective of which public communication channel is used, including the internet" and provides narrow exceptions to this prohibition that are discussed below (ICRC Commentary, at 1624).

The UN Special Rapporteur has stated that "[d]uring armed conflict, people are at their most vulnerable and in the greatest need of accurate, trustworthy information to ensure their own safety and well-being. Yet, it is precisely in those situations that their freedom of opinion and expression, which includes 'the freedom to seek, receive and impart information and ideas of all kinds, is most constrained by the circumstances of war and the actions of the parties to the conflict and other actors to manipulate and restrict information for political, military and strategic objectives" (Report A/77/288, para. 1).

The connection between the right of access to information, including for victims of human rights violations, has also been emphasized by the mandate holder (Report A/68/362, para. 92 (2013)). Some of the most important journalism in conflict situations has included sharing information and imagery of prisoners of war. Eyewitness accounts of detainees following the liberation of Nazi concentration camps in 1945, as well as in Omarska camp in Bosnia in 1992, were crucial in galvanizing global opinion regarding the horrors of these wars and the atrocities committed. Similarly, widely circulated images of detainee abuse at Abu Ghraib prison in Iraq in 2004 led to public condemnation and several prosecutions for these abuses.

Where restrictions on expression are imposed by a state, 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 human rights responsibilities, which include the responsibility to respect freedom of expression. As the UN Special Rapporteur on freedom of expression has stated, although "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 users' right to freedom of expression" (report A/74/486, para. 41).

I. Legality (clarity and accessibility of the rules)

The principle of legality under international human rights law requires rules that limit expression to be clear and publicly accessible (General Comment No.34, para. 25). Rules restricting expression "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (Ibid). Applied to rules that govern online speech, the UN Special Rapporteur on freedom of expression has said they should be clear and specific (A/HRC/38/35, para. 46). People using Meta's platforms should be able to access and understand the rules, and content reviewers should have clear guidance on their enforcement.

The Board finds that Meta's rule prohibiting content which reveals the identity of prisoners of war is sufficiently clear to govern the content in this case, as is the potential for Meta's Content Policy Team to exceptionally leave up content that would otherwise violate this rule where the public interest requires it. The policy lines for applying a "mark as disturbing" screen to graphic

and violent content are sufficiently clear to govern the content in this case.

At the same time, Meta's public explanations of the newsworthiness allowance could provide further detail on how it may apply to content revealing the identity of prisoners of war. Of the three examples of content where a newsworthiness allowance was applied, illustrating Meta's approach to newsworthy content, none concern the Coordinating Harm and Promoting Crime Community Standard. While one example relates to a conflict situation, the criteria or factors particular to conflict could be more comprehensively made public as part of the explanations of the underlying policy rules. In response to Board's prior recommendations, Meta has already provided more clarity around the newsworthiness allowance, including through adding information to its public explanation of the newsworthiness allowance as to when it will apply a warning screen ("Sudan graphic video," recommendation no. 2), and linking the public explanation to the landing page of the Community Standards and adding examples to the newsworthiness page, including about protests ("Colombia protests," recommendation no. 2). The Board stresses the importance of enhanced transparency and guidance to users, especially in crisis and conflict situations.

II. Legitimate aim

a. The Community Standard prohibiting depictions of identifiable prisoners of war

Respecting the rights of others, including the right to life, privacy, and protection from torture or cruel, inhuman, or degrading treatment, is a legitimate aim for restrictions on the right to freedom of expression (Article 19, para. 3, ICCPR). In this case, the assessment of the legitimacy of the aim underlying the prohibition on depicting identifiable prisoners of war is informed by the situation of armed conflict and the more specific rules of international humanitarian law which call for the protection of the life, privacy and dignity of prisoners of war, when the content exposes prisoners of war to "insult" and "public curiosity" (Article 13, para. 2 of the Geneva Convention (III)).

In the context of an armed conflict, Article 13 of the Geneva Convention III provides protection for the humane treatment of prisoners of war, and Meta's general rule coupled with the availability of the newsworthiness allowance supports that function. In addition to potential offline violence, the sharing of the images themselves can be humiliating and violate the detainees' right to

privacy, especially as detained individuals cannot meaningfully consent to such images being taken or shared. Experiencing those images being shared can revictimize and shows how social media can be abused to directly violate the laws of war. This applies not only to the depicted prisoners of war but serves a protective function to prisoners of war more broadly, as well as family members and others who could be targeted. The protection of these rights relates closely to Meta's values of privacy, safety and dignity. The Board finds that Meta's Community Standard prohibiting depictions of identifiable prisoners of war is legitimate.

b. Meta's rules on warning screens

The Board affirmed that Meta's rules on violent and graphic content pursue legitimate aims in the "Sudan graphic video" case, and in several cases since. In the context of this case and for other content like it, the rules providing for a "mark as disturbing" warning screen seek to empower users with more choices over what they see online.

III. Necessity and proportionality

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's analysis of necessity and proportionality is informed by the more specific rules in international humanitarian law. According to the ICRC, Geneva Convention III, Article 13 para. 2 requires a "reasonable balance" to be struck between the benefits of public disclosure of materials depicting prisoners of war, given the high value of such materials when used as evidence to prosecute war crimes, promote accountability, and raise public awareness of abuse, and the potential humiliation and even physical harm that may be caused to the persons in the shared materials. Further, the ICRC notes in its guidance towards media that such materials may be exceptionally disclosed, if there is a "compelling public interest" in revealing the identity of the prisoner or if it is in the prisoner's "vital interest" to do so (ICRC Commentary on Article 13 at (1627)).

Meta's default rule is consistent with goals embodied in international humanitarian law.

Determining whether a person depicted is an identifiable prisoner of war in the context of an armed conflict requires expert consideration. Therefore, the rule requiring "additional context to enforce," and thus requiring escalation to internal teams before it can be enforced, is necessary. Where content reveals the identity or location of prisoners of war, removal will generally be proportionate considering the severity of harms that can result from such content. Many public comments shared examples of such harms. Concerns were raised about the use of content depicting prisoners of war for propaganda purposes (see e.g., PC-11137 from Digital Rights Foundation, and PC-11144 from Igor Mirzakhanyan), especially when those images are disseminated by the detaining power.

Prisoners of war can face many potential harms when their identities are revealed (see e.g., PC-11096 from Article 19). These can include the humiliation of the prisoner, and the ostracization of them or their family on release. Severe harms may still result even where the user shares the content with well-meaning intent to raise awareness or condemn mistreatment. In this case, the faces of alleged prisoners of war are visible, and they are depicted as they are being captured. This process is accompanied with continuous shouting of abusive language and curse words directed at the prisoners, some of whom seem to be injured, while others appear to be deceased.

However, in this case the potential for newsworthiness was correctly identified, leading to escalation to the Content Policy Team. It is important to ensure that escalations of this kind reach teams with the expertise necessary for assessing complex human rights implications, where the potential harm is imminent. The seriousness of these risks in the context of an armed conflict distinguishes this case from prior decisions where the Board has raised concerns about the scalability of the newsworthiness allowance (see e.g., "Sudan graphic video," or "India sexual harassment video").

In this case, the video documents alleged violations of international humanitarian law. While the video may have been made by the detaining power, it appears that the user's post was aimed at raising awareness of potential violations. This is important to the public's right to information around the fact of the detainees' capture, proof of them being alive and physical conditions of detention as well as shedding light on potential wrongdoing.

It is correct that Meta's newsworthiness allowance can apply to content that is not shared by professional media. Nevertheless, guidance available to journalists on responsible reporting in

conflict situations indicates a presumption against disclosure of images identifying prisoners of war, and that even where there is a compelling public interest, efforts should still be taken to safeguard detainees' dignity.

Social media companies preserving content depicting grave human rights violations or atrocity crimes, such as those crimes specified under the Rome Statute of the International Criminal Court, including against prisoners of war, is important. Public comments highlighted the need for greater clarity from Meta on its practices in this area, especially for cooperation with international mechanisms (see: PC-11128 from Trial International, PC-11136 from Institute for International Law of Peace and Armed Conflict, Ruhr University, Bochum, and PC-11140 from Syria Justice and Accountability Centre). They underlined that keeping such content up is important to identify not only the perpetrators, but also the victims (see e.g., PC-11133 from Center for International and Comparative Law, PC-11139 from Digital Security Lab Ukraine, and PC-11145 from Protection of Rights Without Borders NGO). In the Board's view, this content, properly assessed in its particular context, not only informed the public but contributed to the pressure on the detaining power in real time to protect the rights of the detainees. In accordance with the Geneva Convention III, the ICRC facilitates the exchange of correspondence between the prisoners of war and their family members to "prevent missing cases and maintain family links without compromising the dignity or safety of the prisoners of war."

The decision to apply a warning screen to the content was necessary and proportionate, showing respect for the rights of the prisoners and their families, who could experience mental anguish as a result of being involuntarily exposed to such content. Similar to the Board's "video after Nigeria church attack" decision, the content in this case included a video that showed deceased bodies and injured people at close range, with their faces visible, and audio of prisoners of war experiencing severe discomfort while being verbally abused by their captors. The case content contrasts with the content in the "Russian poem" case, which also concerned a conflict situation, where the Board decided the content should not have been placed behind a "marked as disturbing" screen. That case content was a still image of a body lying on the ground at long range, where the face of the victim was not visible. The Board concluded that "the photographic image lacked clear visual indicators of violence, as described in Meta's internal guidelines to content moderators, which would justify the use of the warning screen." In this case, while the warning screen would likely have reduced the reach of the content and therefore its impact on public discourse, providing users with the choice of whether to see disturbing

its impact on public discourse, providing users with the choice of whether to see disturbing content is a proportionate measure. Many public comments, including from people in regions experiencing conflict, favoured the application of warning screens given the graphic nature of the video and the high public interest in keeping the content (see e.g., PC-11139 from Digital Security Lab Ukraine, PC-11144 from Igor Mirzakhanyan and PC-11145 from Protection of Rights without Borders NGO).

9. Oversight Board Decision

The Oversight Board upholds Meta's decision to leave up the content with a "mark as disturbing" warning screen.

10. Recommendations

A. Content Policy

1. In line with recommendation no. 14 in the "former President Trump's suspension" case, Meta should commit to preserving, and where appropriate, sharing with competent authorities evidence of atrocity crimes or grave human rights violations, such as those specified in the Rome Statute of the International Criminal Court, by updating its internal policies to make clear the protocols it has in place in this regard. The protocol should be attentive to conflict situations. It should explain the criteria, process and safeguards for (1) initiating and terminating preservation including data retention periods, (2) accepting requests for preservation, (3) and for sharing data with competent authorities including international accountability mechanisms and courts. There must be safeguards for users' rights to due process and privacy in line with international standards and applicable data protection laws. Civil society, academia, and other experts in the field should be part of developing this protocol. The Board will consider this recommendation implemented when Meta shares its updated internal documents with the Board.

B. Enforcement

2. To ensure consistent enforcement, Meta should update the Internal Implementation Standards to provide more specific guidance on applying the newsworthiness allowance to content that identifies or reveals the location of prisoners of war, consistent with the factors

outlined in Section 8 of this decision, to guide both the escalation and assessment of this content for newsworthiness. The Board will consider this recommendation implemented when Meta incorporates this revision and shares the updated guidance with the Board.

C. Transparency

3. To provide greater clarity to users, Meta should add to its explanation of the newsworthiness allowance in the Transparency Center an example of content that revealed the identity or location of prisoners of war but was left up due to the public interest. The Board will consider this recommendation implemented when Meta updates its newsworthiness page with an example addressing prisoners of war.

4. Following the development of the protocol on evidence preservation related to atrocity crimes and grave human rights violations, Meta should publicly share this protocol in the Transparency Center. This should include the criteria for initiating and terminating preservation, data retention periods, as well as the process and safeguards for accepting requests for preservation and for sharing data with competent authorities, including international accountability mechanisms and courts. There must be safeguards for users' rights to due process and privacy in line with international standards and applicable data protection laws. The Board will consider this recommendation implemented when Meta publicly shares this protocol.

***Procedural note:**

The Oversight Board's decisions are prepared by panels of five Members and approved by a majority of the Board. Board decisions do not necessarily represent the personal views of all Members.

For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg which draws on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, an organization that engages in open-source research on social media trends, also provided analysis.