2022-011-IG-UA

Video after Nigeria church attack

DECEMBER 14, 2022

The Board has overturned Meta's decision to remove a video from Instagram showing the aftermath of a terrorist attack in Nigeria.

Policies and topics: Mistreatment, Safety, War and conflict; Violent and graphic content

Region and countries: Subsaharan Africa; Nigeria

Platform: Instagram

Attachments

- Video after Nigeria church attack - public comments

Case summary

The Board has overturned Meta's decision to remove a video from Instagram showing the aftermath of a terrorist attack in Nigeria. The Board found that restoring the post with a warning screen protects victims' privacy while allowing for discussion of events that some states may seek to suppress.

About the case

On June 5, 2022, an Instagram user in Nigeria posted a video showing motionless, bloodied bodies on the floor. It appears to be the aftermath of a terrorist attack on a church in southwest Nigeria, in which at least 40 people were killed and many more injured. The content was posted on the same day as the attack. Comments on the post included prayers and statements about safety in Nigeria.

Meta's automated systems reviewed the content and applied a warning screen. However, the user was not alerted as Instagram users do not receive notifications when warning screens are applied.

The user later added a caption to the video. This described the incident as “sad,” and used multiple hashtags, including references to firearms collectors, allusions to the sound of gunfire, and the live-action game “airsoft” (where teams compete with mock weapons). The user had included similar hashtags on many other posts.

Shortly after, one of Meta’s Media Matching Service banks, an “escalations bank,” identified the video and removed it. Media Matching Service banks can automatically match users’ posts to content that has previously been found violating. Content in an “escalations bank” has been found violating by Meta's specialist internal teams. Any matching content is identified and immediately removed.

The user appealed the decision to Meta and a human reviewer upheld the removal. The user then appealed to the Board.

When the Board accepted the case, Meta reviewed the content in the “escalations bank,” found it was non-violating, and removed it. However, it upheld its decision to remove the post in this case, saying the hashtags could be read as “glorifying violence and minimizing the suffering of the victims.” Meta found this violates multiple policies, including the Violent and Graphic Content policy, which prohibits sadistic remarks.

Key findings

A majority of the Board finds that restoring this content to Instagram is consistent with Meta’s Community Standards, values and human rights responsibilities.

Nigeria is experiencing an ongoing series of terrorist attacks and the Nigerian government has suppressed coverage of some of them, though it does not appear to have done so in relation to the June 5 attack. The Board agrees that in such contexts freedom of expression is particularly important.

When the hashtags are not considered, the Board is unanimous that a warning screen should be applied to the video. This would protect the privacy of the victims, some of whose faces are visible, while respecting freedom of expression. The Board distinguishes this video from the image in the “Russian poem” case, which was significantly less graphic, where the Board found a warning screen was not required. It also distinguishes it from the footage in the “Sudan graphic video" case, which was significantly more graphic, where the Board agreed with Meta’s decision to restore the content with a warning screen, applying a “newsworthiness allowance,” which permits otherwise violating content.

A majority of the Board finds that the balance still weighs in favor of restoring the content when the hashtags are considered, as they are raising awareness and are not sadistic. Hashtags are commonly used to promote a post within a community. This is encouraged by Meta’s algorithms, so the company should be cautious in attributing ill-intent to their use. The majority notes that Meta did not find that these hashtags are used as coded mockery. Users commenting on the post appeared to understand that it was intended to raise awareness, and responses from the post’s author were sympathetic to the victims.

A minority of the Board finds that adding shooting-related hashtags to the footage appears sadistic, and could traumatize survivors or victims’ families. A warning screen would not reduce this effect. Given the context of terrorist violence in Nigeria, Meta is justified in exercising caution, particularly when victims are identifiable. The minority therefore finds this post should not be restored.

The Board finds the Violence and Graphic Content policy should be clarified. The policy prohibits “sadistic remarks,” yet the definition of that term included in the internal guidance for moderators is broader than its common usage.

The Board notes that the content was originally removed because it matched a video that had wrongly been added to the escalations bank. In the immediate aftermath of a crisis, Meta was likely attempting to ensure that violating content did not spread on its platforms. However, the company must now ensure content mistakenly removed is restored, and resulting strikes, reversed.

The Oversight Board's decision

The Oversight Board overturns Meta's decision to remove the post and finds it should be restored to the platform with a “disturbing content” warning screen.

The Board recommends that Meta:

- Review the language in the public Violent and Graphic Content policy to ensure it aligns with the internal guidance for moderators.
- Notify Instagram users when a warning screen is applied to their content and provide the specific policy rationale for doing so.

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

Full case decision

1. Decision summary

Meta removed an Instagram post containing a captioned video depicting the aftermath of an attack at a church in Nigeria, for violating its policies on Violent and Graphic Content, Bullying and Harassment, and Dangerous Individuals and Organizations. A majority of the Board finds that the content should be restored to the platform with a “disturbing content” warning screen, requiring users to click through to see the video. A minority of the Board disagrees and would uphold Meta’s decision to remove the content.

2. Case description and background

On June 5, 2022, terrorists attacked a Catholic church in Owo, southwestern Nigeria, killing at least 40 people and injuring approximately 90 others. Within hours of the attack, an Instagram user in Nigeria posted a video on their public account that appears to be of the aftermath, showing motionless and bloodied bodies on the church floor, some with their faces visible. Chaotic sounds, including people wailing and screaming, can be heard in the background. The video was initially posted without a caption. There were fewer than 50 comments. Those seen by the Board included prayers for victims, crying emojis, and statements about safety in Nigeria. The author of the post had responded to several showing solidarity with those sentiments.

After the user posted the content, it was identified by one of Meta's Violent and Graphic Content Media Matching Service banks, which contained a substantially similar video. This bank automatically flags content which has previously been identified by human reviewers as violating the company's rules. In this case, the bank referred the user's video to an automated content moderation tool called a classifier, which can assess how likely content is to violate a Meta policy. The classifier determined that the video should be allowed on Instagram. It also determined that the content likely contained imagery of violent deaths and, as a result, automatically applied a "disturbing content" warning screen as required by the Violent and Graphic Content policy. Meta did not notify the user that the warning screen had been applied. Over the following 48 hours, three users reported the content, including for depicting death and severe injury.

At the same time, Meta's staff were working to identify and deal with content arising from the attack. Meta's policy team was alerted about the attack by regional staff and added videos of the incident to a different Media Matching Service bank, an "escalations bank." Content in this escalations bank has been found violating by Meta's specialist internal teams, and any matching content is immediately removed. Videos added to the escalations bank after the incident included footage that showed visible human innards

(which the video at issue in this case did not). Other teams at Meta were invited to refer potentially similar videos to the policy team. The policy team would then determine if they should also be added to the escalations bank.

Three days after the attack, Meta added a video almost identical to the content in this case to the escalations bank. As a result, Meta's systems compared that video to content already on the platform to check for matches. While this retroactive review was taking place, the user edited their original post, adding an English-language caption to the video. It states that the church was attacked by gunmen, that multiple people were killed, and describes the incident as "sad." The caption included a large number of hashtags. The majority of these were about the live-action game "airsoft" (where teams compete to tag each other out of play using plastic projectiles shot with mock weapons). Another, according to Meta, alluded to the sound of gun-fire and is also used to market firearms. Other hashtags referenced people who collect firearms and firearm paraphernalia, as well as military simulations.

Shortly after the caption was added, the escalations bank's retroactive review matched the user's post to the recently added near-identical video, and removed it from the platform. The user appealed. A human moderator reviewed the content and maintained the removal decision. The user then appealed to the Board.

At this point, the three reports users had made on the content had still not been reviewed and were closed. Meta told the Board that the reports had mistakenly been assigned to a low priority queue.

In response to the Board selecting this case, Meta reviewed the near-identical video that had been placed in the escalations bank. Meta determined that it did not violate any policies because there were no "visible innards" and no sadistic caption, and removed it from the bank. However, Meta maintained its decision to remove the content in this case, as it stated that, while the narrative about the event and user's expression of

sadness were not violating, the hashtags in the caption added by the user violated multiple policies. In response to questions from the Board, Meta analyzed the user's posting history and found that the user had included similar hashtags on many of their recent Instagram posts.

The Board notes as relevant context the recent history of violence and terrorist incidents in Nigeria. Experts consulted by the Board stated that the Nigerian government has at times suppressed domestic reporting of terror attacks but does not appear to have done so to a significant degree with regards to the June 5 attack, which was widely covered by traditional media. Graphic imagery of the attack and its victims was widely circulated on social media platforms, including Instagram and Facebook, but was not shown to the same extent by traditional media. In response to questions from the Board, Meta confirmed that the Nigerian government did not contact Meta regarding the attack or request that the content be taken down.

3. Oversight Board authority and scope

The Board has authority to review Meta's decision following an appeal from the user whose content was removed (Charter Article 2, Section 1; Bylaws Article 3, Section 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).

4. Source of authority

The Oversight Board considered the following authorities and standards:

I. Oversight Board decisions:

- “Russian poem” (case decision 2022-008-FB-UA): the Board discussed challenges arising from content moderation in conflict situations and noted that the Violent and Graphic Content policy was unclear.
- “Mention of the Taliban in news reporting” (case decision 2022-005-FB-UA): the Board discussed how users can comment on the activities of terrorist entities.
- “Colombian police cartoon” (case decision 2022-004-FB-UA): the Board recommended Meta improve procedures to remove non-violating content incorrectly added to Media Matching Service banks.
- “Sudan graphic video” (case decision 2022-002-FB-MR): the Board discussed the need to amend and clarify the Violent and Graphic Content policy.
- "Pro-Navalny protests in Russia” (case decision 2021-004-FB-UA): the Board discussed legitimate aim of the Bullying and Harassment Community Standard.
- “Nazi quote” (case decision 2020-005-FB-UA): the Board discussed comments left on content by its authors friends and followers, as an indicator of a poster’s likely intent.

II. Meta’s content policies:

This case involves Instagram's Community Guidelines and Facebook's Community Standards. Meta’s third quarter transparency reportstates that "Facebook and Instagram share Content Policies. This means that if content is considered violating on Facebook, it is also considered violating on Instagram."

The Instagram Community Guidelines say that Meta “may remove videos of intense, graphic violence to make sure Instagram stays appropriate for everyone.” This links to

the Facebook Violent and Graphic Content Community Standard, where the policy rationale states:

To protect users from disturbing imagery, we remove content that is particularly violent or graphic, such as videos depicting dismemberment, visible innards or charred bodies. We also remove content that contains sadistic remarks towards imagery depicting the suffering of humans and animals. In 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 Violent and Graphic Content policy states that "imagery that shows the violent death of a person or people by accident or murder" will be placed behind a disturbing content warning screen. 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. It also states that content will be removed if there are "visible innards."

Meta's Bullying and Harassment policy rationale explains that it removes a variety of content "because it prevents people from feeling safe and respected on Facebook." Under Tier 4 of the specific rules, the company prohibits content that "praises, celebrates or mocks the death or serious injury" of private individuals.

Meta's Dangerous Individuals and Organizations policy rationale explains that Meta prohibits "content that praises, substantively supports or represents events that Facebook designates as violating violent events – including terrorist attacks, hate events, multiple-victim violence or attempted multiple-victim violence." Under Tier 1 of the specific rules Meta removes any praise of such events.

III. Meta's values:

Meta's values are outlined in the introduction to Facebook's Community Standards and the company has confirmed that these values apply to Instagram. The value of "Voice" is described as "paramount":

The goal of our Community Standards is to create a place for expression and give people a voice. Meta wants people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Meta limits "Voice" in the service of four other values, three of which are relevant here:

Safety: *We're committed to making Facebook a safe place. We remove content that could contribute to a risk of harm to the physical security of persons. Content that threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook.*

Privacy:*We're committed to protecting personal privacy and information. Privacy gives people the freedom to be themselves, choose how and when to share on Facebook and connect more easily.*

Dignity: *We believe that all people are equal in dignity and rights. We expect that people will respect the dignity of others and not harass or degrade others.*

IV. International human rights standards:

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. The Board's analysis of Meta's human rights responsibilities in this case was informed by the following human rights standards:

- 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 freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019).
- The right to privacy: Article 17, ICCPR.

5. User submissions

In their statement to the Board, the user explained that they shared the video to raise awareness of the attack and to let the world know what was happening in Nigeria.

6. Meta's submissions

Meta explained in its rationale that, under the Violent and Graphic Content policy, imagery, including video, that shows the violent death of people is usually placed behind a warning screen that indicates it may be disturbing. Adult users may click through to view the content, whereas minors do not have that option. However, Meta also explained that when such content is accompanied by sadistic remarks, it is removed. According to Meta, this is to stop people using the platforms to glorify violence or celebrate the suffering of others. Meta confirmed that, without a caption, the video would be permitted on Instagram behind a disturbing content warning screen. If the video had included "visible innards," as other videos of the same incident had, it would be removed under the Violent and Graphic Content policy without the need for sadistic remarks.

Initially, Meta told the Board that in this case the user was not notified of the warning screen, or the policy used to apply it, because of a technical error. However, after further questioning from the Board, Meta disclosed that while Facebook users generally receive notification of the addition of a warning screen and the reason, Instagram users receive no notification.

Meta explained that its internal guidance for moderators, the Known Questions, define sadistic remarks as those that "are enjoying or deriving pleasure from the suffering/humiliation of a human or animal." The Known Questions provide examples of remarks that qualify as sadistic, divided into those that show an "enjoyment of suffering" and "humorous responses." Meta also confirmed that sadistic remarks can be expressed through hashtags as well as emojis.

In its analysis of the hashtags used in this case, Meta explained that the reference to the sound of gunfire was a "humorous response" to violence that made light of the June 5 terror attack. Meta explained the same hashtag is also used to market weapons. Meta also stated that the gunfire hashtag, as well as the hashtag referring to individuals who collect firearms and firearm paraphernalia, "could be read as glorifying violence and minimizing the suffering of the victims by invoking humor and speaking positively about the weapons and gear used to perpetrate their death." Meta also explained that the hashtag referring to military simulations compared the attack to a simulation, "minimizing the actual tragedy and real-world harm experienced by the victims and their community." Meta also stated that the hashtags referring to "airsoft" compared the attack to a game in a way that glorifies violence as something done for pleasure.

Meta explained that the user's caption asserting that they do not support violence and that the attack was a sad day "do not clearly indicate that they are sharing the video to raise awareness of the attack." The company also clarified that, even if the user showed intent to raise awareness, the use of "sadistic hashtags" would still result in removal. To support this position, Meta explained that some users attempt to evade moderation by including deceptive or contradictory language in their posts. Meta distinguished this from the Board's decision in the "Sudan graphic video" case (2022-002-FB-FBR) where the user made clear their intent to raise awareness while sharing disturbing content. In response to the Board's questions, Meta informed the Board that the user included the same hashtags in most of their recent posts. Meta could not determine why this user was repeatedly using the same hashtags.

Meta also stated that the user's post violated the Bullying and Harassment policy which prohibits content that mocks the death of private individuals. In this case, the hashtag referencing the sound of gunfire was deemed to be a humorous response to the violence shown in the video.

In response to questions from the Board, Meta also determined that the content violated the Dangerous Individuals and Organizations policy. Meta had designated the June 5 attack as a "multiple-victim violence event" and, as a result, any content deemed to praise, substantively support, or represent that event is prohibited under the Dangerous Individuals and Organizations policy. Meta explained that this was in line with its commitments under the Christchurch Call for Action, and that it brought the June 5 attack to the attention of industry partners in the Global Internet Forum to Counter Terrorism. Meta explained that, while it was a "close call," the content in this case appears to mock the victims of the attack and speak positively about the weapons used, and therefore qualifies as praise of a designated event under its policy.

Meta stated that removing the content in this case strikes the appropriate balance between its values. The user's caption demonstrated a lack of respect for the dignity of the victims, their families, and the community impacted by the attack – all of which outweigh the value of the user's own voice. In response to the Board's questions, Meta confirmed that it did not issue any newsworthiness allowances in relation to content containing violating imagery related to the June 5 attack.

Finally, Meta explained its actions were consistent with international human rights law, stating that its policy on sadistic remarks is clear and accessible, the policy aims to protect the rights of others, as well as public order and national security, and all actions short of removal would not adequately address the risk of harm. Meta pointed to the European Court of Human Rights decision in *Hachette Filipacchi Associes v. France* (2007), which held that journalists who published photos of someone's violent death in a widely distributed magazine "intensified the trauma suffered by the relatives."

Meta also pointed to a 2010 article by Sam Gregory, “Cameras Everywhere: Ubiquitous Video Documentation of Human Rights, New Forms of Video Advocacy, and Considerations of Safety, Security, Dignity and Consent” in the Journal of Human Rights Practice, which explains that “the most graphic violations” such as violent attacks “most easily translate into a loss of dignity, privacy, and agency, and which carries with it the potential for real re-victimization.” Meta noted that its policy is not to remove graphic content, but to place it behind a warning screen, which limits its access to minors. It said that its policy to remove sadistic remarks “goes a step further” because “the value of dignity outweighs the value of voice.”

The Board asked Meta 29 questions, 28 of which were answered fully. Meta was unable to answer a question on the percentage of user reports that are closed without review in the Sub-Saharan Africa market.

7. Public comments

The Oversight Board considered nine public comments related to this case. One of the comments was submitted from Asia Pacific and Oceania, one from Central and South Asia, one from the Middle East and North Africa, one from Sub-Saharan Africa, and five from the United States and Canada.

The submissions covered themes including the need to clarify the Violent and Graphic Content policy, and Nigeria-specific issues that the Board should be aware of while deciding this case.

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

8.Oversight Board analysis

The Board looked at the question of whether this content should be restored through three lenses: Meta's content policies, the company's values, and its human rights responsibilities.

8.1 Compliance with Meta's content policies

The Board analyzed three of Meta's content policies: Violent and Graphic Content; Bullying and Harassment, and Dangerous Individuals and Organizations. The majority of the Board finds that no policy was violated.

I. Content rules

Violent and Graphic Content

The policy rationale states that Meta "removes content that contains sadistic remarks towards imagery depicting the suffering of humans and animals." However, it also states that it allows graphic content with some limitations to help people condemn and raise awareness about "important issues such as human rights abuses, armed conflicts or acts of terrorism." The policy also provides for warning screens to alert people that content may be disturbing, including where imagery shows violent deaths. In the rules immediately under the policy rationale, Meta explains that users cannot post "sadistic remarks towards imagery that is deleted or put behind a warning screen under this policy." Meta does not provide any further public explanation or examples of sadistic remarks.

The Board agrees with Meta that the video at issue in this case shows violent deaths and that, without a caption, it should have a warning screen. Distinct from the content in the Board's "Sudan graphic video" decision, the video in this case, though depicting bloodied dead bodies, does not show visible innards, which would require removal under the policy. In the "Sudan graphic video" case, the hashtags indicated a clear

intent to document human rights abuses, and the Board relied in part on the clear intent of those hashtags and on Meta's newsworthiness allowance to restore the content. In this case, the Board's assessment of the content against Meta's content policies is based in part on the absence of any visible innards, dismemberment or charring in the video footage, as well as the hashtags used. The difference between the majority and minority positions turns on the purpose or meaning that should be attributed to the hashtags in this case.

A majority of the Board premises its position on the common use of hashtags by users to promote a post within a certain community, and to associate with others who share common interests and signify relationships. When used in this way, they are not necessarily implying commentary on an image or issue. A majority of the Board finds that the hashtags in the caption are not sadistic, as they are not used in a way that shows the user is "enjoying or deriving pleasure" from the suffering of others. Interpreting the long list of idiosyncratic hashtags as commentary on the video is, in this case, misguided. This distinguishes this case from the Board's decision in the "Sudan graphic video" case, in which hashtags clearly indicated the user's intent in sharing a graphic video. The user's inclusion of hashtags about the game airsoft, as well as those related to firearms and military simulations, should not have been read as "glorifying violence" (under the Dangerous Individuals and Organizations policy) and still less as "mocking" (under the Bullying and Harassment policy) or showing that the user was "enjoying or deriving pleasure from the suffering of others" (under the Graphic Violence policy). Many users of social media have and share interests in airsoft, firearms, or military simulation and may use hashtags to connect with others without in any way expressing support for terrorism or violence against individuals. The airsoft hashtags are more directly associated with enthusiasm for the game, and are, as a whole, incongruous with the content of the video and commentary the user shared immediately below it. This should have indicated to Meta that the user was trying to raise awareness amongst the people they normally communicate with on Instagram, and to reach others. As Instagram's design incentivizes the liberal use of hashtags as a means to promote

content and connect with new audiences, it is important that Meta is cautious before attributing ill intent to their use. Independent research commissioned by the Board confirmed that the hashtags used in this post are used widely among airsoft and firearm enthusiasts, and Meta did not find that these hashtags had been used as coded mockery to evade detection on its platforms.

For the majority, it was also clear that the commentary the user added to the video, after the warning screen was applied, did not indicate they were enjoying or deriving pleasure from the attack. The user stated that the attack represented a "sad day," and that they do not support violence. Comments on the post further indicated that the user's followers understood the intent to be awareness raising, like the situation in the "Nazi quote" case. The user's responses to those comments also showed further sympathy with victims. The Board accepts Meta's argument that explicit user statements in content that they do not support violence should not always be accepted at face value, as some users may attempt to evade moderation by including them, contrary to the actual purpose of their posts. The Board recalls its finding that it should not be necessary for a user to expressly state condemnation when commenting on the activities of terrorist entities, and that expecting them to do so could severely limit expression in regions where such groups are active (see the Board's decision in the "Mention of the Taliban in news reporting" case).

A minority of the Board concludes that when assessed together, the juxtaposition of shooting-related hashtags against the footage appears sadistic, comparing the murder of those depicted to games and appearing to promote weapons imitating those used in the attack. This would appear sadistic to survivors of the attack and relatives of those deceased, and the potential for re-traumatization would not be reduced by placing the content behind a warning screen. Given the context of terrorist violence in Nigeria, the minority finds that Meta is justified to err on the side of caution where commentary on graphic violence appears sadistic, even if there is a degree of ambiguity. This is especially relevant for content like this video, where specific victims are identifiable as

their faces are visible, and where escalating violence or further retaliation against survivors from attackers cannot be ruled out. A minority of the Board also finds that the statements in the caption in this case do not negate the sadistic effect of juxtaposing hashtags associated with gun enthusiasts with a video depicting the horrific aftermath of violence inflicted with guns. While hashtags may serve an associative purpose for members of a community, a minority of the Board believes that it is appropriate, in situations such as this, for Meta to apply its policies in a manner that considers the content from the perspective of survivors and the victims' families.

It is also important to consider how Meta can swiftly and consistently enforce its content policies in crisis situations, such as in the aftermath of terrorist acts where imagery quickly spreads across social media. The minority considered it pertinent that a moderator or casual reader would not know that the user had routinely included these hashtags on most of their recent posts. In a fast moving situation, the minority find that Meta was correct to interpret the use of firearms-related hashtags as indicating that the user is deriving enjoyment from the suffering depicted. The majority acknowledges that the removal of the content was a reasonable mistake and agrees with Meta that it was "a difficult call." Nevertheless, the Board's independent analysis (assisted by experts providing contextual information on the shooting, on violence in Nigeria more generally and its relationship to social media, and on the meaning and use of the hashtags) leads the majority to conclude that it is an error to characterize these hashtags as sadistic merely because they are associated with users of firearms.

Bullying and Harassment

Tier 4 of the Bullying and Harassment policy prohibits content that mocks the death or serious injury of private individuals.

For the same reasons set forth in the previous section, a majority of the Board finds that the content is not mockery, as the purpose of the hashtags is not an attempt at humor

but an attempt to associate with others, this is confirmed by the responses to the post and the user's engagement with them. Meta erred by presuming a string of hashtags are commentary on the shared video. As noted above, that the user was not asking firearms enthusiasts to mock the victims appears to be confirmed by responses to the post expressing shock and sympathy, which Meta confirms were mostly from users in Nigeria, and the user's engagement with those responses (see the Board decision in the "Nazi quote" case). While the majority agrees it is important to consider the perspectives of survivors and victims' families, the responses to this content indicate that those perspectives do not necessarily weigh against keeping content on the platform, particularly given the frequency of attacks on Christians in Nigeria.

A minority of the Board disagrees. By adding hashtags involving imitation firearms it also appears that the user was intentionally directing a video depicting the victims of a shooting to firearms enthusiasts. Meta was correct to find that this appears mocking, and it is appropriate for the company to prioritize the perspective of survivors and the victims' families in making this assessment.

Dangerous Individuals and Organizations

Tier 1 of the Dangerous Individuals and Organizations policy prohibits content that praises, substantively supports, or represents "multiple-victim violence."

The Board agrees that according to Meta's definition of "multiple-victim violence," the June 5 attack qualifies. However, the majority finds that the use of hashtags in the caption is not "praise" of the attack, for the same reasons it was not sadistic. The minority disagrees and finds that, while it is a close call, for the same reasons articulated in the previous sections, the juxtaposition between the hashtags and the content could be viewed as praise of the attack itself.

II. Enforcement action

Meta initially informed the Board that the user was not sent a message when their content was put behind a warning screen due to a technical issue. However, in response to questions from the Board, Meta investigated and determined that Instagram users are not notified when their content is placed behind a warning screen. In this case, the addition of a caption in which the user explicitly states that they do not support violence may have been an attempt to respond to the imposition of the warning screen. Meta should ensure that all users are notified when their content is placed behind warning screens, and told why this action has been taken.

The Board notes that the content in this case was removed because the video matched with a near identical video that was mistakenly added to an escalations bank that automatically removes matched content. In the “Colombian police cartoon” case, the Board said Meta must ensure that it has robust systems and screening processes to assess content before it is added to any Media Matching Service banks that delete matches without further review. The Board understands that in the immediate aftermath of a crisis, Meta was likely attempting to ensure that violating content did not spread on its platforms. However, given the multiplying impacts of Media Matching Service banks, controls remain critical. Meta should ensure that all content mistakenly removed due to this wrongful banking is restored and any related strikes are reversed.

The Board is concerned that the three user reports of the content were not reviewed in the five days before the content was removed. In response to questions from the Board, Meta explained that this was due to an unknown technical issue which it is investigating. In response to further questions, Meta stated that it is unable to ascertain what percentage of Instagram user reports in Sub-Saharan Africa are closed without review.

8.2 Compliance with Meta’s values

The Board concludes that removing the content in this case was inconsistent with Meta’s value of “Voice.”

The Board recognizes the competing interests in situations such as the one in this case. The content in this case implicates the dignity and privacy of the victims of the June 5 attack, as well as that of their families and communities. A number of the victims in the video have their faces visible and are likely identifiable.

The Board recalls that in its “Sudan graphic video" case, and its “Russian poem” case, it called for improvements to Meta’s policy on Violent and Graphic Content to align it with Meta’s values. The Board found the policy’s treatment of content sharing graphic content to “raise awareness” was insufficiently clear. In a number of cases, the Board has found that warning screens can be appropriate mechanisms to balance Meta’s values of “Voice,” “Privacy,” “Dignity,” and “Safety” (see the Board’s decisions in the “Sudan graphic video” and “Russian poem” cases). The Board is in agreement that, in contexts where civic space and media freedom is illegitimately restricted by the state, as is the case in Nigeria, Meta’s value of "Voice" becomes even more important (see its decision in the "Colombia protests" case). It also agrees that raising awareness of human rights abuses is a particularly important aspect of “Voice,” which can in turn advance “Safety” by ensuring access to information. Warning screens can further this exercise of “Voice,” though they may be inappropriate where the content is not sufficiently graphic as they significantly reduce reach and engagement with content (see the Board’s decision in the "Russian poem” case).

For essentially the same reason laid out in the preceding section, the majority and minority reached different conclusions about what Meta’s values require for interpreting the hashtags added to the video caption. The majority notes that there is a particular need to protect “Voice” where content draws attention to serious human rights violations and atrocities, including attacks on churches in Nigeria. The majority finds these hashtags do not contradict the user’s stated sympathy to the victims, expressed in the caption, and that their use is consistent with the user’s efforts to raise awareness. As the caption is not “sadistic,” it is consistent with Meta’s values to restore the content with an age-gated warning screen. While the majority acknowledges that adding a warning

screen may impact “Voice” by limiting the reach of content raising awareness of human rights abuses, given the identifiability of the victims it is required to properly balance the values of “Dignity” and “Safety”.

The minority find removal of the content justified to protect the “Dignity” and “Safety” of the victims’ families and survivors, who are at a high-risk of re-traumatization from exposure to content that appears to be providing sadistic and mocking commentary on the killings of their loved ones. That several victims’ faces are visible and identifiable in the video and are not blurred is pertinent. In respect of “Voice,” the minority finds it relevant that similar content without hashtags was shared on the platform behind a warning screen, and it remained possible for this user to share similar content without firearm-related hashtags. Meta’s removal of this content did not therefore excessively hinder efforts of the community in Nigeria to raise awareness of, or seek accountability for, these atrocities.

8.3 Compliance with Meta’s human rights responsibilities

A majority of the Board finds that removing the content in this case is inconsistent with Meta’s human rights responsibilities. However, as in the “Sudan graphic video” case, the majority and minority agree that Meta should amend the Violent and Graphic Content policy in order to make clear which policy rules impact content that aims to raise awareness of human rights abuses and violations.

Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides broad protection for freedom of expression, including the right to seek and receive information. However, the right may be restricted under certain specific conditions, as evaluated according to a three-part test of legality, legitimacy, and necessity and proportionality. The Board has adopted this framework to analyze Meta’s content policies and enforcement practices. The UN Special Rapporteur

on freedom of expression has encouraged social media companies to be guided by these principles when moderating online expression, mindful that regulation of expression at scale by private companies may give rise to concerns particular to that context (A/HRC/38/35, paras. 45 and 70).

I.Legality (clarity and accessibility of the rules)

The principle of legality requires any restriction on the right to freedom of expression to be clear and accessible, so that individuals know what they can and cannot do (General Comment No. 34, para. 25 and 26). Lack of specificity can lead to subjective interpretation of rules and their arbitrary enforcement. The Santa Clara Principles on Transparency and Accountability in Content Moderation, which have been endorsed by Meta, are grounded in ensuring companies' respect for human rights in line with international standards including freedom of expression. They provide that companies must have "understandable rules and policies," including "detailed guidance and examples of permissible and impermissible content."

Before addressing each content policy, the Board notes its previous recommendations for Meta to clarify the relationship between the Instagram Community Guidelines and Facebook's Community Standards ("Breast cancer symptoms and nudity" case, 2020-004-IG-UA-2and "Öcalan's isolation" case, 2021-006-IG-UA-10), and urges Meta to complete its implementation of this recommendation as soon as possible.

Violent and Graphic Content

The Board reiterates its concern that the Violent and Graphic Content policy is insufficiently clear with regards to how users may raise awareness of graphic violence under the policy. In this case, there are further concerns about the content captured under Meta's definition of "sadistic."

In the “Sudan graphic video” case decision, the Board stated that:

[T]he Violent and Graphic Content policy does not make clear how Meta permits users to share graphic content to raise awareness of or document abuses. The rationale for the Community Standard, which sets out the aims of the policy, does not align with the rules of the policy. The policy rationale states that Meta allows users to post graphic content "to help people raise awareness about" human rights abuses, but the policy prohibits all videos (whether it is shared to raise awareness or not) "of people or dead bodies in non-medical settings if they depict dismemberment."

The Board recommended that Meta amend the policy to specifically allow imagery of people and dead bodies to be shared to raise awareness or document human rights abuses. It also recommended that Meta to develop criteria to identify videos shared for that purpose. Meta has stated that it is assessing the feasibility of those recommendations and will conduct a policy development process to determine whether they can be implemented. Meta has also updated the policy rationale “to ensure that it reflects the full range of enforcement actions covered in the policy and adds clarification about the deletion of exceptionally graphic content and sadistic remarks.” However, the Board notes that the rules on what can and cannot be posted under this policy still do not provide clarity on how otherwise prohibited content may be posted to “raise awareness.”

The Board also notes that after it publicly announced its selection of this case and sent questions to the company, Meta updated the policy rationale to include a reference to its existing prohibition on “sadistic” remarks. However, the term is still not publicly defined, as the policy simply lists types of content that users cannot make sadistic remarks towards. The Board finds the common usage of the term “sadistic” has connotations of intentional depravity and seriousness, which do not align adequately with Meta’s internal guidance for moderators, the Known Questions. That internal guidance shows that Meta’s definition of “sadistic” is broadly defined to extend to any humorous

response or positive speech about human or animal suffering. This appears to set a lower bar for the removal of content than the public-facing policy communicates.

Bullying and Harassment

Under Meta's Bullying and Harassment policy, the company prohibits content that mocks the death or serious physical injury of private individuals. The Board did not find that the framing of this rule raised legality concerns in this case.

Dangerous Individuals and Organizations

Under Tier 1 of this policy, Meta prohibits praise of designated "violating violent events," a category which includes terrorist attacks, "multiple-victim violence, and multiple murders." The Board notes that Meta does not appear to have a consistent policy regarding when it publicly announces events that it has designated. Without this information, users in many scenarios may not know why their content was removed.

II. Legitimate aim

Restrictions on freedom of expression should pursue a legitimate aim, which includes the protection of the rights of others, such as the right to privacy of the identifiable victims, including those who are deceased, depicted in this content (General Comment 34, para. 28).

The Board has previously assessed the three policies at issue in this case and determined that each pursues the legitimate aim of protecting the rights of others. The Violent and Graphic Content policy was assessed in the "Sudan graphic video" case, the Bullying and Harassment policy was assessed in the "Pro-Navalny protests in Russia" case, and the Dangerous Individuals and Organizations policy was assessed in the "Mention of the Taliban in news reporting" case.

III. Necessity and proportionality

Restrictions on 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 interests to be protected" (General Comment 34, para. 34).

The Board has discussed whether warning screens are a proportionate restriction on expression in the “Sudan graphic video” decision and the “Russian poem” decision. The nature and severity of graphic violence has been determinative in those decisions, and Meta’s human rights responsibilities have at times been in tension with its stated content policies and their application. The “Russian poem” case concerned a picture, taken at a distance, of what appeared to be a dead body. The face was not visible, the person was not identifiable, and there were no visible graphic indicators of violence. In that case, the Board found that a warning screen was not necessary. By contrast, the “Sudan graphic video” case concerned a video showing dismemberment and visible innards, shot at closer range. In that case, the Board found that the content was sufficiently graphic to justify the application of a warning screen. The latter decision relied on the newsworthiness allowance, which is used to permit otherwise violating content. This was used because the policy itself was not clear on how it could be applied to permit content raising awareness of human rights violations.

In the present case, the Board agrees that, absent the hashtags, a warning screen was necessary to protect the privacy rights of victims and their families, primarily because the victims’ faces are visible and the location of the attack was known. This makes victims identifiable, and more directly engages their privacy rights and the rights of their families. The depictions of death are also significantly more graphic than in the “Russian poem” case, with bloodied bodies shown at much closer range. However, there is no dismemberment and there are no "visible innards.” If either of these features had been present the content would have to be removed or given a newsworthiness allowance to

allow it to remain on the platform. While a warning screen will reduce both reach and engagement with the content, it is a proportional measure to both respect expression while also respecting the rights of others.

The majority of the Board finds that removing the content was not a necessary or proportionate restriction on the user's freedom of expression, and that it should be restored with a "disturbing content" warning screen. The majority finds that the addition of the hashtags did not increase the risk of harming the privacy rights and dignity of victims, survivors or their families, as substantially similar footage is already on Instagram behind a warning screen.

By drastically reducing the number of people who would see the content, the application of a warning screen in this case served to respect the victims' privacy (as with other instances of similar videos), while also allowing for discussion of events that some states may seek to suppress. In contexts of ongoing insecurity, it is particularly important that users are able to raise awareness of recent developments, document human rights abuses, and promote accountability for atrocities.

For a majority of the Board, the caption as a whole, including the hashtags, was not sadistic and would need to have more clearly demonstrated sadism, mockery, or glorification of the violence for removal of the content to be considered necessary and proportionate.

A minority of the Board agrees with the majority in terms of their analytical approach and overall view of Meta's policies in this area, but disagrees with their interpretation of the hashtags, and therefore the outcome of their human rights analysis.

For the minority, removal of the post was in line with Meta's human rights responsibilities and the principles of necessity and proportionality. When events like this attack occur, videos of this nature frequently go viral. The user in this case had a large

number of followers. It is crucial that in response to incidents like this, Meta acts quickly and at-scale, including through collaboration with industry partners, to prevent and mitigate harms to the human rights of victims, survivors and their families. This also serves a broader public purpose of countering the widespread terror that perpetrators of such attacks seek to instill, knowing that social media will amplify their psychological impacts. For the minority, it is therefore less important in human rights terms whether the user in this case primarily intended to use the hashtags to connect with their community or increase their reach. The value of those associations, to the individuals concerned and the broader public, while not insignificant, are far outweighed by the importance of respecting the right to privacy and dignity of the survivors and victims. Victims' faces are visible and identifiable at close range in the video, in a place of worship, with their bodies covered in blood. The juxtaposition between this and the militaristic hashtags about weapons in the caption is jarring and appears mocking. Exposing victims and their family members to such content would likely re-traumatize them, even if that is not what the posting user intended.

For the minority, this is distinct from the Board's "Sudan graphic video" case, where the hashtags very clearly indicated intent to document human rights abuses. While explicit statements of intent to raise awareness should not be a policy requirement (see the Board's "Wampum belt," and "Mention of the Taliban in news reporting" decisions), it is consistent with Meta's human rights responsibilities to remove hashtags that non-critically evoke enthusiasm for weapons alongside identifiable imagery of persons killed by gunfire. In these circumstances, the minority believes Meta should err in favor of removal.

9. Oversight Board decision

The Oversight Board overturns Meta's decision to take down the content, requiring the post to be restored with a "mark as disturbing" warning screen.

10. Policy advisory statement

Content policy

1. Meta should review the public facing language in the Violent and Graphic Content policy to ensure that it is better aligned with the company's internal guidance on how the policy is to be enforced. The Board will consider this recommendation implemented when the policy has been updated with a definition and examples, in the same way as Meta explains concepts such as "praise" in the Dangerous Individuals and Organizations policy.

Enforcement

2. Meta should notify Instagram users when a warning screen is applied to their content and provide the specific policy rationale for doing so. The Board will consider this recommendation implemented when Meta confirms notifications are provided to Instagram users in all languages supported by the platform.

***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. An independent research institute headquartered at the University of Gothenburg and drawing 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, and Memetica, a digital investigations group providing risk advisory and threat intelligence services to mitigate online harms.