2022-012-IG-MR

India sexual harassment video

DECEMBER 14, 2022

The Board has upheld Meta's decision to restore a post to Instagram containing a video of a woman being sexually assaulted by a group of men.

Policies and topics: Freedom of expression, Marginalized communities, News events; Sexual exploitation of adults

Region and countries: Central and South Asia; India

Platform: Instagram

Attachments

- India sexual harassment video - public comments

Case summary

The Board has upheld Meta's decision to restore a post to Instagram containing a video of a woman being sexually assaulted by a group of men. The Board has found that Meta's "newsworthiness allowance" is inadequate in resolving cases such as this at scale and that the company should introduce an exception to its Adult Sexual Exploitation policy.

About the case

In March 2022, an Instagram account describing itself as a platform for Dalit perspectives posted a video from India showing a woman being assaulted by a group of men. "Dalit" people have previously been referred to as "untouchables," and have faced oppression under the caste system. The woman's face is not visible in the video and there is no nudity. The text accompanying the video states that a "tribal woman" was

sexually assaulted in public, and that the video went viral. “Tribal” refers to indigenous people in India, also referred to as Adivasi.

After a user reported the post, Meta removed it for violating the Adult Sexual Exploitation policy, which prohibits content that “depicts, threatens or promotes sexual violence, sexual assault or sexual exploitation."

A Meta employee flagged the content removal via an internal reporting channel upon learning about it on Instagram. Meta's internal teams then reviewed the content and applied a “newsworthiness allowance.” This allows otherwise violating content to remain on Meta’s platforms if it is newsworthy and in the public interest. Meta restored the content, placing the video behind a warning screen which prevents anyone under the age of 18 from viewing it, and later referred the case to the Board.

Key findings

The Board finds that restoring the content to the platform, with the warning screen, is consistent with Meta’s values and human rights responsibilities.

The Board recognizes that content depicting non-consensual sexual touching can lead to a significant risk of harm, both to individual victims and more widely, for example by emboldening perpetrators and increasing acceptance of violence.

In India, Dalit and Adivasi people, especially women, suffer severe discrimination, and crime against them has been rising. Social media is an important means of documenting such violence and discrimination and the content in this case appears to have been posted to raise awareness. The post therefore has significant public interest value and enjoys a high degree of protection under international human rights standards.

Given that the video does not include explicit content or nudity, and the majority of the Board finds the victim is not identifiable, a majority finds that the benefits of allowing the video to remain on the platform, behind a warning screen, outweigh the risk of harm. Where a victim is not identifiable, their risk of harm is reduced significantly. The warning screen, which prevents people under-18 from viewing the video, helps to protect the dignity of the victim, and protects children and victims of sexual harassment from exposure to disturbing or traumatizing content.

The Board agrees the content violates Meta's Adult Sexual Exploitation policy and that the newsworthiness allowance could apply. However, echoing concerns raised in the Board's "Sudan graphic video" case, the Board finds that the newsworthiness allowance is inadequate for dealing with cases such as this at scale.

The newsworthiness allowance is rarely used. In the year ending June 1, 2022, Meta only applied it 68 times globally, a figure that was made public following a recommendation by the Board. Only a small portion of those were issued in relation to the Adult Sexual Exploitation Community Standard. The newsworthiness allowance can only be applied by Meta's internal teams. However, this case shows that the process for escalating relevant content to those teams is not reliable. A Meta employee flagged the content removal via an internal reporting channel upon learning about it on Instagram.

The newsworthiness allowance is vague, leaves considerable discretion to whoever applies it, and cannot ensure consistent application at scale. Nor does it include clear criteria to assess the potential harm caused by content that violates the Adult Sexual Exploitation policy. The Board finds that Meta's human rights responsibilities require it to provide clearer standards and more effective enforcement processes for cases such as this one. A policy exception is needed which can be applied at scale, that is tailored to the Adult Sexual Exploitation policy. This should provide clearer guidance to distinguish posts shared to raise awareness from those intended to perpetuate violence or discrimination, and help Meta to balance competing rights at scale.

The Oversight Board's decision

The Oversight Board upholds Meta's decision to restore the post with a warning screen.

The Board also recommends that Meta:

- Include an exception to the Adult Sexual Exploitation Community Standard for depictions of non-consensual sexual touching. This would only be applied by Meta’s internal teams and would permit content where the victim is not identifiable, and that Meta judges is shared to raise awareness, is not shared in a sensationalized context, and does not involve nudity.
- Update its internal guidance to at-scale reviewers on when to escalate content reviewed under the Adult Sexual Exploitation Community Standard that may be eligible for the above policy exception.

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

Full case decision

1.Decision summary

A majority of the Board upholds Meta’s decision to restore the content to the platform and to apply a warning screen over it. The Board finds, however, that while the newsworthiness allowance could be applied in this case, it does not provide a clear standard or effective process to resolve cases such as this one at scale. In line with Meta’s values and human rights responsibilities, Meta should add a clearly defined exception in the Adult Sexual Exploitation policy which is applied in a more consistent and effective way than the newsworthiness allowance. The exception should be designed to protect content that raises awareness of public issues and to help Meta balance the specific risks of harms presented by content that violates the Adult Sexual Exploitation policy. The majority of the Board finds that, rather than relying on the

newsworthiness allowance, it is preferable to apply the recommended exception to the Adult Sexual Exploitation policy, which would permit the content in this case. A minority finds that this exception does not apply to the content in this case.

2. Case description and background

In March 2022, an Instagram account describing itself as a news platform for Dalit perspectives posted a video from India showing a woman being assaulted by a group of men. Dalits, previously referred to as “untouchables,” are socially segregated and economically marginalized in India due to the country’s caste system – a hierarchical system of social stratification. In the video, the woman's face is not visible and she is fully clothed. The text accompanying the video states in English that a "tribal woman" was sexually assaulted and harassed by a group of men in public, and that the video previously went viral. The term “tribal” refers to indigenous people in India, who are also referred to as Adivasi. The account that posted the video has around 30,000 followers, mostly located in India. Dalit and Adivasi women are frequently the target of assaults in the country (see section 8.3.).

The content was reported by another Instagram user for sexual solicitation and sent for human review. Human reviewers determined that the content violated Meta's Adult Sexual Exploitation policy. Under this policy, Meta removes content "that depicts, threatens or promotes sexual violence, sexual assault or sexual exploitation." Following the removal of the content, Meta applied one standard strike (a strike that applies to all violation types), one severe strike (a strike that applies to the most egregious violations, including violations of the Adult Sexual Exploitation policy), and a 30-day feature limit to the content creator’s account. The feature limit prevented the user from starting any live video.

On the day the original content was removed, a member of Meta’s Global Operations team saw a post on their personal Instagram account discussing the content’s removal,

and escalated the original post. When content is escalated, it is reviewed by policy and safety experts within Meta. Upon escalation, Meta issued a newsworthiness allowance, reversed the strikes, restored the content, and placed a warning screen on the video alerting users that it may contain violent or graphic content. The warning screen prevents users under the age of 18 from viewing the content and requires all other users to click through the screen to view the video. A newsworthiness allowance permits content on Meta's platforms that might otherwise violate its policies if the content is newsworthy and keeping it visible is in the public interest. It can only be applied by specialist teams within Meta, and not by human reviewers who review content at scale.

Meta referred this case to the Board, stating that it demonstrates the challenge in striking "the appropriate balance between allowing content that condemns sexual exploitation and the harm in allowing visual depictions of sexual harassment to remain on [its] platforms."

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).

4.Sources of authority

The Oversight Board considered the following authorities and standards:

I. Oversight Board decisions:

The most relevant previous decisions of the Oversight Board include:

- "Sudan graphic video" decision (2022-002-FB-MR): In this case, which dealt with graphic content posted to raise awareness of human rights abuses, the Board concluded that "the newsworthiness allowance is not an effective means of allowing this kind of content on Facebook at scale." It recommended that Meta includes an exception in the Community Standard to assure the application of this exception at scale. The Board also argued that "the lack of clarity surrounding when, and how, the newsworthiness allowance is applied is likely to invite arbitrary application of this policy." The Board explained that warning screens can be a proportionate solution, as they do "not place an undue burden on those who wish to see the content while informing others about the nature of the content and allowing them to decide whether to see it or not. The warning screen also adequately protects the dignity of the individual depicted and their family."
- "Colombia protests" decision: (2021-010-FB-UA): In this case, the Board recommended that Meta "develop and publicize clear criteria for content reviewers to escalate for additional review public interest content." In its policy advisory opinion on sharing private residential information, the Board repeated this recommendation.
- "Ayahuasca brew" decision (2021-013-IG-UA): In this case, the Board decided that content should be left up although the relevant policy prohibited the content. Allowing the content was in line with Meta's values and human rights responsibilities, and the Board recommended Meta change its policies to align with its values and human rights responsibilities.

- "Knin cartoon" decision (2022-001-FB-UA): In this case, the Board argued that, to help users understand how their content will be treated, Meta must provide more detailed information on the escalation process.

II. Meta's content policies:

Instagram Community Guidelines

The Instagram Community Guidelines provide, under the heading "follow the law," that Instagram has "zero tolerance when it comes to sharing sexual content involving minors or threatening to post intimate images of others." The words "intimate images" include a link to the Facebook Community Standards on Adult Sexual Exploitation in Meta's Transparency Center. The Community Guidelines do not expressly address depictions of non-consensual sexual images.

Facebook Community Standards

In the policy rationale for the Adult Sexual Exploitation policy, Meta recognizes "the importance of Facebook as a place to discuss and draw attention to sexual violence and exploitation." Therefore, it "allow[s] victims to share their experiences, but remove[s] content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation." To protect victims and survivors, Meta removes "images that depict incidents of sexual violence and intimate images shared without the consent of the person[s] pictured."

The "do not post" section of this Community Standard states that content consisting of "any form of non-consensual sexual touching," such as "depictions (including real photos/videos except in a real-world art context)" are removed from the platform. This policy also states that Meta "may restrict visibility to people over the age of 18 and

include a warning label on certain fictional videos […] that depict non-consensual sexual touching."

In the Transparency Center, Meta explains that whether a strike is applied "depends on the severity of the content, the context in which it was shared and when it was posted." It aims for its strike system to be "fair and proportionate."

Newsworthiness allowance

Defining the newsworthiness allowance in its Transparency Center, Meta explains that it allows "content that may violate Facebook's Community Standards or Instagram Community Guidelines, if it is newsworthy and keeping it visible is in the public interest." Meta only does this "after conducting a thorough review that weighs the public interest against the risk of harm" and looks to "international human rights standards, as reflected in our Corporate Human Rights Policy, to help make these judgments." The policy states that "content from all sources, including news outlets, politicians, or other people, is eligible for a newsworthy allowance" and "[w]hile the speaker may factor into the balancing test, we do not presume that any person's speech is inherently newsworthy." When the newsworthiness allowance is applied and content is restored, but may be sensitive or disturbing, restoration may include a warning screen.

When weighing public interest against the risk of harm, Meta takes the following factors into consideration: whether the content poses imminent threats to public health or safety; whether the content gives voice to perspectives currently being debated as part of a political process; country-specific circumstances (for example, whether there is an election underway, or the country is at war); the nature of the speech, including whether it relates to governance or politics; and the political structure of the country, including whether it has a free press.

III. Meta's values:

Meta's values are outlined in the introduction to Facebook's Community Standards. The value of "Voice" is described as "paramount":

The goal of our Community Standards has always been to create a place for expression and give people a voice. […] We want 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 service of four values, the relevant ones in this case being "Safety," "Privacy" and "Dignity":

"Safety": We're committed to making Facebook a safe place. 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).
- The right to life: Article 6, ICCPR.
- The right to privacy: Article 17, ICCPR.
- The right to non-discrimination: Article 2, para. 1, ICCPR; Article 1, Convention on the Elimination of All Forms of Discrimination against Women (CEDAW); General recommendation No. 35 on gender-based violence against women, UN Committee on the Elimination of Discrimination against Women; General Recommendation No. 35 on combating racist hate speech, Committee on the Elimination of Racial Discrimination (CERD).
- The right to physical and mental health: Article 12, International Covenant on Economic, Social and Cultural Rights (ICESCR).
- Rights of the child: Art. 3, Convention on the Rights of the Child (CRC) on the best interest of the child; Articles 17 and 19, CRC, on the rights of children to be protected from all forms of physical or mental violence; General Comment No. 25, Committee on the Rights of the Child, 2021.

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

In the rationale provided for this case, Meta explained that the caption and background of the user posting the content indicated intent to condemn and raise awareness of

violence against marginalized communities. However, there is no relevant exception for this in the Adult Sexual Exploitation policy.

With regard to its decision to apply the newsworthiness allowance, Meta explained that the public interest value of the content was high because the content was shared by a news organization that highlights the stories of underrepresented and marginalized populations. Meta said the content appears to have been shared with the intent to condemn the behavior in the video and raise awareness of gender-based violence against tribal women. Adivasi and marginalized voices, Meta argued, have been historically repressed in India and would benefit from greater reach and visibility. Meta also argued that the risk of harm was limited as the depiction did not involve overt nudity or explicit sexual activity, and does not sensationalize. It argued that the "case was exceptional in that the victim's face is not visible and her identity is not readily identifiable."

In response to questions asked by the Board, Meta further explained that "a user's self-description as a news organization is a factor that is considered, but is not determinative, in deciding whether it is treated as a news organization." Subject matter experts and regional market specialists decide which users qualify as news organizations, based on a variety of factors, including their market knowledge and previous classifications of the organizations. Meta argued that its decision and policy is in line with its values and human rights responsibilities.

The Board asked Meta 15 questions in this case. Meta answered 14 fully and did not answer one. The Board asked Meta to share its Human Rights Impact Assessment Report for India with the Board, which Meta declined, citing security risks. Meta failed to provide a satisfactory explanation for why sharing the Human Rights Impact Assessment Report with the Board would entail security risks.

7. Public comments

The Oversight Board considered 11 public comments related to this case. One of the comments was submitted from Asia Pacific and Oceania, four were submitted from Central and South Asia, three from Europe, and three from the United States and Canada.

The submissions covered the following themes: marginalization of Adivasi in India; power relations in the caste system; the potential of depictions of violence to embolden perpetrators and contribute to violence; the difference between non-sexual violence and sexual violence and the importance of contextual assessments with regard to the latter; the risk of victims of sexual harassment being ostracized by their own communities; intersectionality; risks of trauma of sexual assault survivors; harmful effects of hate speech on social media in India; the importance of social media as a tool for raising awareness of violence against marginalized groups; and the importance of Meta keeping a highly secured cache of removed content that is accessible to law enforcement officials.

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 believes the newsworthiness allowance could be applied in this case, but does not provide a clear standard or effective process to assess this kind of content at scale. The Board therefore recommends that, in addition to the newsworthiness allowance, as a general exception to any policy, Meta include an exception to the Adult

Sexual Exploitation policy, which would provide a clear and effective process for moderating content at scale.

A majority of the Board believes the content in this case should be allowed under such an exception, while a minority believes no exception should apply to this specific content and that it should be removed from the platform.

I.Content rules and enforcement

The Board agrees with Meta's assessment that the content in this case violates the prohibition in the Adult Sexual Exploitation Standard on depictions of non-consensual sexual touching.

A majority of the Board agrees with the substance of Meta's reasoning for reinstating the content and believes that the newsworthiness allowance could be applied in this case due to the content's strong public interest value. However, the Board believes that the newsworthiness allowance does not provide an adequate standard or process to assess content such as the post in this case at scale, as it does not assure an effective and consistent application.

The Board agrees with Meta's assessment that there is strong public interest value in keeping this content on the platform, as it raises awareness of violence against a marginalized community. The Board also agrees that leaving content on the platform which depicts non-consensual sexual touching, including assault, can entail significant risks of harm (see section 8.3, below). The Board further agrees with Meta that in cases in which a victim of non-consensual sexual touching is identifiable, potential harm is too great and content generally should be removed, certainly if it is posted without the consent of the victim.

In this case, however, the Board disagrees on whether the victim is identifiable. A majority of the Board believes that the victim is not identifiable. The victim's face is not visible in the video, and the video is shot from a distance and generally of poor quality. The caption does not provide any information on the victim's identity. A minority believes that the content should be removed from the platform on the basis that there is some possibility that the victim could be identified. Viewers of the video who have local knowledge of the area or the incident might be able to identify the victim even if their face is not visible. The likelihood of this, a minority believes, is especially high as the incident was widely reported by local news outlets. The majority acknowledges the minority's concerns but does not believe that local awareness of an incident should, by itself, mean that a victim is "identifiable."

The Board recommends that Meta review its policies and processes based on its values and human rights responsibilities, as analyzed in sections 8.2 and 8.3 below, and introduce a clearly defined exception in the Adult Sexual Exploitation Standard which can be applied in a more consistent and effective way than the newsworthiness allowance.

II.Transparency

Following the Board's recommendations in the "Colombia protests," and "Sudan graphic video" decisions, Meta has provided more information in its Transparency Center on the factors it considers in determining whether its newsworthiness allowance should be applied to a piece of content. It has not, however, developed and publicized "clear criteria for content reviewers to escalate for additional review public interest content that potentially violates the Community Standards but may be eligible for the newsworthiness allowance," as recommended by the Board in its "Colombia protests" decision, and its policy advisory opinion on sharing private residential information. The Board reiterates its concern that Meta should provide more information on the escalation process in the context of the newsworthiness allowance.

8.2 Compliance with Meta’s values

In this case, as in many, Meta's values of “Voice,” “Privacy,” “Safety,” and “Dignity” may point in different directions.

Raising awareness about abuses against Adivasi serves the value of “Voice" and may also help to protect the safety and the dignity of Adivasi. On the other hand, publicity around sexual assault may be unwelcome for the victim or may normalize the conduct, creating negative impacts on the privacy, dignity, and safety of the victim or others in their community.

Because the video in this case was not explicit and the majority of the Board considered that the victim was not identifiable, the majority believes that leaving the video on the platform is consistent with Meta's values, taken as a whole. A minority of the Board maintains, however, that even if the likelihood of the victim being identified were low, a real risk of identification persists. In their view, concerns for the privacy, dignity, and safety of the victim must prevail, and the video should be removed.

8.3 Compliance with Meta’s human rights responsibilities

A majority of the Board finds that keeping the content on the platform is consistent with Meta’s human rights responsibilities. Meta has committed itself to respect human rights under the UN Guiding Principles on Business and Human Rights (UNGPs).
Its Corporate Human Rights Policystates that this commitment includes respecting the International Covenant on Civil and Political Rights (ICCPR).

Freedom of expression (Article 19 ICCPR)

The scope of the right to freedom of expression is broad. Article 19, para. 2, of the ICCPR gives heightened protection to expression on political issues (General Comment

No. 34, paras. 20 and 49). The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) also provides protection from discrimination in the exercise of the right to freedom of expression (Article 5). The Committee on the Elimination of Racial Discrimination has emphasized the importance of the right with respect to assisting "vulnerable groups in redressing the balance of power among the components of society" and to offer "alternative views and counterpoints" in discussions (CERD Committee, General Recommendation 35, para. 29). The content in this case appears to have been shared to raise awareness of violence against Adivasi women in India, and, in line with the standards provided in General Comment No. 34, enjoys a high level of protection.

Under Article 19, para. 3, ICCPR, restrictions on expression must (i) be provided for by law, (ii) pursue a legitimate aim, and (iii) be necessary and proportionate. The ICCPR does not create binding obligations for Meta as it does for states, but this three-part test has been proposed by the UN Special Rapporteur on Freedom of Expression as a framework to guide platforms' content moderation practices (A/HRC/38/35).

I. Legality (clarity and accessibility of the rules)

Rules restricting expression must be clear and accessible so that those affected know the rules and may follow them (General Comment No. 34, paras. 24-25). Applied to Meta, users of its platforms and reviewers enforcing the rules should be able to understand what is allowed and what is prohibited. In this case, the Board concludes that Meta falls short of meeting that responsibility.

The Board finds that the wording of the newsworthiness policy is vague and leaves significant discretion to whoever applies it. As the Board noted in the "Sudan graphic video" case (2022-002-FB-MR) , vague standards invite arbitrary application and fail to assure the adequate balancing of affected rights when moderating content.

The Board has also repeatedly drawn attention to the lack of clarity for Instagram users about which policies apply to their content, particularly, if and when Facebook policies apply (see, for example, the Board's decisions in the "Breast cancer symptoms and nudity" case (2020-004-IG-UA), and the "Ayahuasca brew" case (2021-013-IG-UA)). The Board reiterates that concern here.

The Board further reiterates the critical need for more information around the standards, internal guidance and processes that determine when content is escalated (see, for example, the Board's decision in the "Colombia protests" case, the "Sudan graphic video" case, and the "Knin cartoon" case). For users to understand whether and how the newsworthiness allowance will apply to the content, Meta must provide more detailed information on the escalation process.

II.Legitimate aim

Under Article 19, para. 3, ICCPR, freedom of expression may be limited for the purpose of protecting "the rights of others." The Adult Sexual Exploitation Policy aims to prevent abuse, re-victimization, social stigmatization, doxing and other forms of harassment. It serves the protection of the right to life (Article 6, ICCPR), the right to privacy (Article 17, ICCPR) and the right to physical and mental health (Article 12, ICESCR). It also pursues the goal of preventing discrimination and gender-based violence (Article 2, para. 1, ICCPR, Article 1, CEDAW).

In applying the warning screen, Meta pursues the legitimate aim of mitigating the harms described above, and of protecting other users. The warning that appears on the screen aims to protect victims of sexual harassment from being exposed to potentially retraumatizing, disturbing and graphic content (Article 12, ICESCR). The age restriction also pursues the legitimate aim of protecting children from harmful content (Art. 3, 17 and 19, CRC).

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 34, para. 34).

In this case, a majority finds that removal of the content would not be necessary and proportionate but that applying a warning screen and age restriction satisfies this test. A minority believes that removal would be necessary and proportionate. The Board finds that Meta's human rights responsibilities require it to provide clearer standards and more effective enforcement processes to allow content on the platform in cases such as this one. It therefore recommends that Meta include a clearly defined exception in the Adult Sexual Exploitation Standard, which would be a better way of balancing competing rights at scale. As a general exception which is rarely applied, the newsworthiness allowance would still be available, and allow Meta to assess whether public interest outweighs the risk of harm including in cases where the victim is identifiable.

Decision to leave the content up

The Board recognizes that content such as this may lead to considerable harms. This includes harms to the victim, who, if identified, could suffer re-victimization, social stigmatization, doxing and other forms of abuse or harassment (see PC-10802 (Digital Rights Foundation)). The severity of potential harms is great when the victim is identified. Where the victim cannot be identified, their risk of harm is reduced significantly. In this case, the majority finds that the probability of the victim being identified is low. Even if the victim is not publicly identified, the victim may be harmed by interacting with the content on the platform, and by comments and reshares of that

content. The majority believes that the application of the warning screen over the content addresses this concern.

The Board also considered broader risks. Social media in India has been criticised for spreading caste-based hate-speech (see the report on Caste-hate speech by the International Dalit Solidarity Network, March 2021). Online content can reflect and strengthen existing power structures, embolden perpetrators and motivate violence against vulnerable populations. Depictions of violence against women can lead to attitudes that are more accepting of such violence. Public comments highlighted that sexual harassment is an especially cruel form of harassment (see, for example, PC-10808 (SAFEnet), PC-10806 (IT for change), PC-10802 (Digital Rights Foundation), PC-10805 (Media Matters for Democracy)).

The majority balanced those risks with the fact that it is important for news organizations and activists to be able to rely on social media to raise awareness of violence against marginalized communities (see the public comments PC-10806 (IT for change), PC-10802 (Digital Rights Foundation), PC-10808 (SAFEnet)), especially in a context where media freedom is under threat (see the report by Human Rights Watch). In India, Dalit and Adivasi people, especially women who fall at the intersection of caste and gender (see PC-10806, (IT for Change)), suffer severe discrimination, and crime against them has been on the rise. Civil society organizations report rising levels of ethnic-religious discrimination against non-Hindu and caste minorities which undermines equal protection of the law (see Human Rights Watch report). With the government targeting independent news organizations, and public records underrepresenting crime against Adivasi and Dalit individuals and communities, social media has become an important means of documenting discrimination and violence (see the report by Human Rights Watch and the public comments cited above).

The Board ultimately disagrees on the question whether, in this particular case, there is any reasonable risk of identifying the victim. The majority believes that this risk is

minimal and therefore the interest of keeping the content on the platform outweighs potential harms if a warning screen is applied. The minority believes that the remaining risk requires removing the content from the platform.

Warning screen and age restriction

The Board believes that applying a warning screen is a “lesser restriction” compared to removal. In this case, the majority believes that applying a warning screen would be the least intrusive way to mitigate potential harms inflicted by the content while protecting freedom of expression. As the Board found in the “Sudan graphic video” case, the warning screen “does not place an undue burden on those who wish to see the content while informing others about the nature of the content and allowing them to decide whether to see it or not.” In addition, it "adequately protects the dignity of the individual depicted and their family.” The minority believes that a warning screen does not sufficiently mitigate potential harms and that the severity of those harms requires the removal of the content.

The warning screen also triggers an age restriction, which seeks to protect minors. General Comment No. 25 on Children’s Rights in Relation to the Digital Environment states that “parties should take all appropriate measures to protect children from risks to their right to life, survival and development. Risks relating to content ... encompass, among other things, violent and sexual content...” (para 14). It further states that children should be “protected from all forms of violence in the digital environment” (paras. 54, 103). The majority of the Board agrees with Meta’s reasoning that an age restriction reconciles the objective of protecting minors with the objective of allowing content which is in the public interest to be seen.

Design of policy and enforcement processes

While the majority agrees with Meta's ultimate decision to allow the content on the platform, the Board believes that the newsworthiness allowance is an ineffective mechanism to be applied at scale. The Board unanimously finds that allowing depictions of sexual violence against marginalized groups on the platform should be based on clear policies and accompanied by adequately nuanced enforcement. This should distinguish posts such as this one, which are being shared to raise awareness, from posts being shared to perpetuate violence or discrimination against these individuals and communities. It should include clear criteria to assess the risks of harm presented by such content, to help Meta balance competing rights at scale.

The newsworthiness allowance is an ineffective mechanism for moderating content at scale (see the Board's decision in the "Sudan graphic video" case). This is indicated by the fact it is used so rarely. According to Meta the newsworthiness allowance was applied just 68 times across all policies globally between June 1, 2021 to June 1, 2022. Only a small portion of those were issued in relation to the Adult Sexual Exploitation Community Standard. This case demonstrates that the internal escalation process for application of the newsworthiness allowance is not reliable: the content was not escalated by any of the at-scale human reviewers who initially reviewed the content, but by a member of the Global Operations team. Upon learning about the content removal on Instagram, they flagged the issue via an internal reporting channel. In Meta's content moderation process, most content is reviewed by external at-scale reviewers rather than Meta's internal specialized teams. Content can be escalated for additional review by these internal specialized teams where at scale reviewers consider that the newsworthiness allowance may apply – however, the escalation process is only effective when at-scale human reviewers have clear guidance on when to escalate content. The newsworthiness allowance is a general exception, that can be applied to content violating any of Meta's policies. It does therefore not include criteria to assess or balance the harms presented by content violating the Adult Sexual Exploitation policy in particular.

A more effective means of protecting freedom of expression and allowing people to raise awareness of the sexual harassment of marginalized groups while protecting the rights of the victim and marginalized communities would be to include an exception to the Adult Sexual Exploitation Standard, to be applied "at escalation." In addition, at-scale reviewers should be instructed to escalate content when the exception potentially applies, rather than relying on the rarely applied newsworthiness allowance (for similar reasoning, see the Board's decision in the "Sudan graphic video" case). The Board therefore recommends that an exception from the Adult Sexual Exploitation policy should be introduced for depictions of non-consensual sexual touching. This would allow content violating the policy to remain on Meta's platforms where, based on a contextual analysis, Meta judges that the content is shared to raise awareness, the victim is not identifiable, the content does not involve nudity, is not shared in a sensationalized context and thus entails minimal risks of harm for the victim. This exception should be applied on escalation-level only, that is by Meta's specialist internal teams. Meta should also provide clear guidance to at-scale reviewers on when to escalate content which potentially falls under this exception. This exception does not preclude the application of the newsworthiness allowance.

Including an exception to the Adult Sexual Exploitation policy and updating guidance to at-scale reviewers would ensure that an assessment on escalation becomes part of a standard procedure which can be triggered by at-scale moderators in every relevant case. At-scale reviewers would still remove content depicting non-consensual sexual touching but would escalate content in cases where the exception potentially applies. Building on regional expertise, policy and safety experts can then decide whether the exception applies. If it does not, they can decide whether strikes should be imposed, and, if so, which strikes, in line with Meta's goal to apply strikes in a proportionate and fair manner. Limiting the application of the exception to specialized teams encourages consistency as well as adequate consideration of potential harms.

Non-discrimination

Meta has a responsibility to respect equality and non-discrimination on its platforms (Articles 2 and 26 ICCPR). In its General Recommendation No. 35, the Committee on the Elimination of Racial Discrimination highlighted the “contribution of speech to creating a climate of racial hatred and discrimination” (para. 5) and the potential of hate speech “leading to mass violations of human rights” (para. 3).

The Board recognizes that there is a difficult tension between allowing content on the platform which raises awareness of violence against marginalized groups and removing content which might potentially harm the privacy and security of an individual who is part of those groups. The Board believes that a significant potential for individual harm could outweigh the benefits of raising awareness of harassment on Instagram. However, in this case, the majority believes that, as the victim is not identifiable and the risk for individual harm is low, the content should remain on the platform with a warning screen. A minority believes that the risk is not low enough and therefore the content should be removed.

9. Oversight Board decision

The Board upholds Meta’s decision to leave the content on the platform with a warning screen.

10. Policy advisory statement

Policy

1. Meta should include an exception to the Adult Sexual Exploitation Community Standard for depictions of non-consensual sexual touching, where, based on a contextual analysis, Meta judges that the content is shared to raise awareness, the victim is not identifiable, the content does not involve nudity and is not shared in a sensationalized context, thus entailing minimal risks of harm for the victim. This

exception should be applied at escalation only. The Board will consider this recommendation implemented when the text of the Adult Sexual Exploitation Community Standard has been changed.

Enforcement

2. Meta should update its internal guidance to at-scale reviewers on when to escalate content reviewed under the Adult Sexual Exploitation Community Standard, including guidance to escalate content depicting non-consensual sexual touching, with the above policy exception. The Board will consider this recommendation implemented when Meta shares with the Board the updated guidance to at-scale reviewers.

***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.