Oversight Board

ANULADO

VIOLENCE AGAINST WOMEN

The Oversight Board has overturned Meta's decisions to remove two Instagram posts which condemned gender-based violence.

TIPO DE DECISIÓN

Estándar:

POLÍTICAS Y TEMAS

Igualdad de sexo y género, Libertad de expresión

Lenguaje que incita al odio

REGIONES/PAÍSES

Suecia

PLATAFORMA

Instagram

ARCHIVOS ADJUNTOS

Public comments appendix

CASE SUMMARY

The Oversight Board has overturned Meta's decisions to remove two Instagram posts which condemned gender-based violence. The Board recommends that Meta include the exception for allowing content that condemns or raises awareness of gender-based violence in the public language of the Hate Speech policy, as well as update its internal guidance to reviewers to ensure such posts are not mistakenly removed.

About the cases

In this decision, the Board considers two posts from an Instagram user in Sweden together. Meta removed both posts for violating its Hate Speech Community Standard. After the Board identified the cases, Meta decided that the first post had been removed in error but maintained its decision on the second post.

The first post contains a video with an audio recording and its transcription, both in Swedish, of a woman describing her experience in a violent intimate relationship, including how she felt unable to discuss the situation with her family. The caption notes that the woman in the audio recording consented to its publication, and that the voice has been modified. It says that there is a culture of blaming victims of gender-based violence, and little understanding of how difficult it is for women to leave a violent partner. The caption says, "men murder, rape and abuse women mentally and physically – all the time, every day." It also shares information about support organizations for victims of intimate partner violence, mentions the International Day for the Elimination of Violence against Women, and says it hopes women reading the post will realize they are not alone.

After one of Meta's classifiers identified the content as potentially violating Meta's rules on hate speech, two reviewers examined the post and removed it. This decision was then upheld by the same two reviewers on different levels of review. As a result of the Board selecting this case, Meta determined that it had removed the content in error, restoring the post.

As the Board began to assess the first post, it received another appeal from the same user. The second post, also shared on Instagram, contains a video of a woman speaking in Swedish and pointing at words written in Swedish on a notepad. In the video, the speaker says that although she is a man-hater, she does not hate all men. She also states that she is a man-hater for condemning misogyny and that hating men is rooted in fear of violence. Meta removed the content for violating its rules on hate speech. The user appealed the removal to Meta, but the company upheld its original decision after human review. After being informed that the Board had selected this case, Meta did not change its position.

Since at least 2017, digital campaigns have highlighted that Facebook's hate speech policies result in the removal of phrases associated with calling attention to gender-based violence and harassment. For example, women and activists have coordinated posting phrases such as "men are trash" and " **men are scum**" and protested their subsequent removal on the grounds of being anti-men hate speech.

Key findings

The Board finds that neither of the two posts violates Meta's rules on hate speech.

On the first post, the Board finds that the statement "Men murder, rape and abuse women mentally and physically – all the time, every day" is a qualified statement which does not violate Meta's Hate Speech policy. Given that the post refers to international campaigns against violence against women and provides local resources for organizations that work to help women victims, it is clear the language describes men who commit violence against women.

In addition, the Board finds that the second post is not an expression of contempt towards men but condemns violence against women and discusses the roots of gender-based hate. While Meta argues that the user's statement that she does not hate all men does not impact the assessment of other parts of the post, the Board disagrees and assesses the post as a whole. The Board finds that the other aspects of the post that Meta cited as potentially violating are not violating when read within the context of the post. Some Board Members disagreed that the posts in question did not violate Meta's hate speech rules.

The Board is concerned that Meta's approach to enforcing gender-based hate speech may result in the disproportionate removal of content raising awareness of and condemning gender-based violence. Meta states, for example, that the first post should be allowed on its platforms and that the H' Speech policy is "designed to allow room for raising awareness of gender-based violence." ver, neither the public-facing policy nor its internal guideline documents to moderators contain

clear guidance to ensure that posts like these would not be mistakenly removed. The company's confusing guidance makes it virtually impossible for moderators to reach the right conclusion. While Meta relied on contextual cues to determine the first post was not violating once it was identified by the Board, the company's guidance for moderators limits the possibility of contextual analysis significantly.

The Board finds that within this context, it is critical that statements that condemn and raise awareness of gender-based violence not be mistakenly removed. The Board's concern that this may be happening is particularly pronounced given that an allowance for this type of content, while highlighted by Meta, is not communicated clearly to the public and the guidance provided to moderators is confusing. To address this, Meta should clarify its public rules and provide appropriate guidance to moderators that better reflects this allowance.

The Oversight Board's decision

The Oversight Board overturns Meta's decisions to remove both posts.

The Board recommends that Meta:

- Include the exception for allowing content that condemns or raises awareness of gender-based violence in the public language of the Hate Speech policy.
- Update guidance to its at-scale moderators with specific attention to rules around qualification to ensure that content condemning and raising awareness of gender-based violence is not removed in error.
- Update its Transparency Center with information on what penalties are associated with the accumulation of strikes on Instagram. The Board appreciates that Meta has provided additional information about strikes for Facebook users in response to Board recommendations. It believes this should be done for Instagram users as well.
- Assess how its current review routing protocol impacts accuracy. The Board believes Meta would improve content moderation accuracy by adjusting this protocol to prioritize sending secondary review jobs to different reviewers than those who previously assessed the content.

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

FULL CASE DECISION

1. Decision summary

The Oversight Board overturns Meta's decisions in two cases about Instagram posts condemning gender-based violence that Meta removed as anti-men hate speech. Meta has acknowledged that its initial decision in the first case was wrong but maintains the second post violates the Hate Speech policy. In both cases, the Board finds the posts do not violate the Hate Speech policy. It is also recommended that Meta should include a clearer exception to allow content that condemns or raises awareness of gender-based violence in the public language of the Hate Speech policy, as well as update its internal guidance so that moderators can effectively implement this exception. This would help ensure that Meta does not incorrectly remove content condemning or raising awareness about gender-based violence.

2. Case description and background

This decision concerns two content decisions made by Meta, which the Oversight Board is addressing here together. An Instagram user in Sweden created two posts with videos and captions. Meta removed both posts for violating its Hate Speech Community Standard. After the Board identified the cases, Meta reversed its decision on the first post stating that it had been removed in error. However, it maintained its decision on the second post.

In the first post, the user posted a video with an audio recording and its transcription, both in Swedish, of a woman describing her experience in a violent intimate relationship, including feeling unable to discuss her situation with her family. The audio does not contain graphic details of violence. The caption to the post notes the woman in the audio recording consented to it being published, and that the voice has been modified. It says that there is a culture of blaming victims of gender-based violence, and little understanding of how difficult it is for women to leave a violent partner. The caption says, "men murder, rape and abuse women mentally and physically - all the time, every day." It also provides a helpline number, shares information about support organizations for victims of intimate partner violence, mentions the International Day for the Elimination of Violence against Women, and says it hopes women reading the post will realize they are not alone. The post was viewed around 10,000 times.

On the same day, a Meta classifier identified the content as potentially violating the Hate Speech policy. Meta stated that due to a bug, the classifier created two review jobs. It then sent the content twice to two reviewers, who each decided twice that the content violated the Hate Speech policy. Meta removed the post and applied a " **strike**" to the user's Instagram account. When Meta removes content, it sometimes applies "strikes," which correspond to different penalties against an account as they accumulate.

The content creator appealed Meta's decision on the same day, and one of the reviewers who had already examined the content upheld the removal. After this, about an hour after the content had initially been posted, it was automatically sent to a High Impact False Positive Override (HIPO) channel, which aims to identify wrongfully removed content. This resulted in the content being sent to the same two reviewers who had originally examined the content. Both reviewers decided once again that the post violated the Hate Speech policy. In total, the content was examined seven times by the same two human reviewers who, on every occasion, found the content to be violating.

As a result of the Board selecting this appeal, Meta reviewed the relevant post and determined that its decision to remove it was in error, restored it, and reversed the strike.

While the Board began to assess the first post, it received another appeal from the same user. This concerned an Instagram video of a woman speaking in Swedish and pointing at words written in Swedish on a notepad. In the video, the speaker says that although she is a man-hater, she does not hate all men. She further explains that this means she talks about and condemns violence against women, and that these feelings of hate are rooted in fear of violence. Within this discussion of fear, the person in the video draws an analogy between venomous snakes and men who commit violence against women. She notes that although many snakes are not poisonous, the fact that some are impacts how people approach them in general, just as the fear towards men stems from a worldwide social problem of violence against women. In the caption of the post, the user calls on men who are "allies" to help women in their fight. The post was viewed around 150,000 times.

Following user reports, Meta removed the content of the second post for violating the Hate Speech policy and again applied a strike against the account, preventing the user from creating live videos. On the same day, the content creator appealed Meta's removal, but the company upheld its original decision after human review. After being informed that the Board had selected this case, Meta did not change its position.

When assessing cases, the Board notes as relevant context research, reporting, and public comments that highlight similar issues. Since at least 2017, digital campaigns have highlighted that Facebook's hate speech policies result in the removal of phrases associated with calling attention to gender-based violence and harassment. For example, women and activists have coordinated posting phrases such as "men are trash" and " **men are scum**" and protested their subsequent removal on the grounds of being anti-men hate speech.

Meta, itself, has reflected on the complexities of its policy approach to gender-based hate speech. In Mark Zuckerberg **explained his rationale for considering** such posts hate speech, citing the challenges the company perceived in enforcing a policy that acknowledged power differences

among different groups. Meta also held a **policy forum** in which it debated potentially modifying the Hate Speech policy, and ultimately decided to continue with its current approach.

3. Oversight Board authority and scope

The Board has authority to review Meta's decision following an appeal from the person 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 non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

When the Board selects cases like the first post, where Meta subsequently acknowledges that it made an error, the Board reviews the original decision to increase understanding of the content moderation process and to make recommendations to reduce errors and increase fairness for people who use Facebook and Instagram. The Board also aims to make recommendations to lessen the likelihood of future errors and treat users more fairly moving forward.

When the Board identifies cases that raise similar issues, they may be assigned to a panel simultaneously to deliberate together. Binding decisions will be made with respect to each piece of content.

4. Sources of authority and guidance

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

I. Oversight Board decisions

The most relevant previous decisions of the Oversight Board include:

- "Mention of the Taliban in news reporting" case (**2022-005-FB-UA**) recommending that Meta release more information on its strike system.
- "Wampum belt" case (**2021-012-FB-UA)** recommending that Meta study the impacts on reviewer accuracy when content moderators are informed they are engaged in secondary review.
- "'Two buttons' meme" case (**2021-005-FB-UA**) examining Meta's procedures to assess relevant text and recommending that Meta provide content moderators necessary time and resources to review content.

- “Armenians in Azerbaijan” case (**2020-003-FB-UA**) examining the Hate Speech policy (use of slurs) where the Board emphasized that context is key in determining the potential for adverse outcomes.

II. Meta’s content policies

The **Instagram Community Guidelines** note that content containing hate speech may be removed and link to Facebook’s **Hate Speech** policy. The Hate Speech policy rationale defines hate speech as a direct attack against people on the basis of protected characteristics, including sex and gender. Meta does not allow Hate Speech on its platforms because it “creates an environment of intimidation and exclusion, and in some cases may promote offline violence.” The rules prohibit “violent” or “dehumanizing” speech and “expressions of contempt” against people based on these characteristics, including men.

Tier 1 of the Hate Speech policy prohibits “dehumanizing speech” includes “comparisons, generalizations, or unqualified behavioral statements to or about ... violent and sexual criminals.” Meta’s internal policy guidelines define “qualified” and “unqualified” behavioral statements and provide examples. Under these guidelines, ‘qualified statements’ do not violate the policy, while ‘unqualified statements’ are violating and removed. Meta states qualified behavioral statements use statistics, reference individuals, or describe direct experience. According to Meta, unqualified behavioral statements “explicitly attribute a behavior to all or a majority of people defined by a protected characteristic.”

Tier 2 of the Hate Speech policy prohibits direct attacks against people on the basis of protected characteristics with "expressions of contempt,” which includes “self-admission to intolerance on the basis of protected characteristics" and “expressions of hate, including, but not limited to: despise, hate.”

The Board’s analysis of the content policies was informed by Meta’s **commitment to "Voice,"** which the company describes as “paramount,” and its values of “Safety” and “Dignity.”

III. Meta’s human rights responsibilities

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. In 2021, Meta **announced** its **Corporate Human Rights Policy**, where it reaffirmed its commitment to respecting human rights in accordance with the UNGPs.

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

- The rights to freedom of opinion and expression: Article 19, International Covenant on Civil and Political Rights (**ICCPR**), **General Comment No. 34**, Human Rights Committee, 2011; UN Human Rights Council resolution on freedom of expression and women's empowerment **A/HRC/Res/23/2** (2013); Special Rapporteur on freedom of opinion and expression, reports: **A/76/258** (2021), **A/74/486** (2019), **A/HRC/38/35** (2018), **A/68/362** (2013); and **Joint Declaration on Freedom of Expression and Gender Justice**, Special Rapporteurs on freedom of opinion and expression of The United Nations (UN), the Organization for Security and Co-operation in Europe (OSCE), the Organization of American States (OAS), the African Commission on Human and Peoples' Rights (ACHPR) (2022).
- The prohibition of incitement to discrimination, hostility or violence: Article 20, para. 2, ICCPR; Rabat Plan of Action, UN High Commissioner for Human Rights report: **A/HRC/22/17/Add.4** (2013).
- The right to non-discrimination: Article 2, para. 1 and Article 26, ICCPR.

5. User submissions

In their first appeal to the Board, the content creator said that they wanted to show women who face domestic violence that they are not alone. They also stated that removing the post stops an important discussion and keeps people from learning, and possibly sharing the post. In their second appeal, they explained that it was clear that they do not hate all men but want to discuss the problem of men committing violence against women.

6. Meta's submissions

After the Board identified the first post, Meta determined that it had been removed in error and did not violate the Hate Speech policy. Meta, however, maintained that the second post violated the Hate Speech policy.

With regards to the first post, Meta stated that the text in the caption that "men murder, rape and abuse women mentally and physically - all the time, every day" likely caused the removal. When read in isolation, Meta found it was an "unqualified behavioral statement" about men comparing them to sexual predators or violent criminals, and therefore violated the Hate Speech policy.

However, once the Board identified the case, Meta determined that, when read within the context of the post as whole, this was a "qualified behavioral statement." Meta explained that an "unqualified

behavioral statement" attributes a behavior to all or a majority of people defined by a protected characteristic, while a "qualified behavioral statement" does not.

Meta further explained that it determined the statement was qualified by looking at several factors. These included: noting the International Day for the Elimination of Violence Against Women; that the user encourages sharing the post and provides information on helplines; and that the user shares a description of an experience of violence and describes it as a social problem. Meta concluded that "the user's clear intent to raise awareness of violence against women provides further support that the content does not violate the Hate Speech policy." Meta also stated that "although [the content] does not squarely fall into Meta's allowance for raising awareness of or condemning someone else's hate speech, [its] policy is designed to allow room for raising awareness of gender-based violence."

While Meta listed the user's intent as a contextual factor in finding the content non-violating, in response to question asked by the Board, the company acknowledged that its policies generally do not grant reviewers discretion to consider intent. According to Meta, to ensure consistent and fair enforcement of its rules, it does not require at-scale reviewers "to infer intent or guess at what someone 'is really saying'" because "divining intent for hate speech invites subjectivity, bias, and inequitable enforcement." As Meta referenced criteria that are not in Meta's internal guidelines to reviewers, the Board asked Meta for any existing guidance that would help reviewers reach the correct outcome in this case. Meta then referenced additional confidential internal guidance that focused on elements not relevant to this case. The company also stated that while this case did "not fit neatly into its policies," it would expect its reviewers to understand that this content is non-violating.

With regard to the second post, Meta found that "[u]nlike the content in [the first] case, this content contained an expression of hatred directed toward men, which violates [Meta's] Hate Speech policy." The company explained that its Hate Speech policy prohibits content targeting men with expressions of contempt, which it defines as "self-admission to intolerance on the basis of protected characteristics," including expressions of hate. For Meta, the reference to being a man-hater is an expression of hate. Meta acknowledged that the user also said they do not hate all men but stated this did not negate the expression of hate.

Meta further noted that the content may violate other elements of its Hate Speech policy but stressed that its removal decision was made based solely on this expression of hate. Meta stated there was an implicit generalization about men, as the user included a phrase about knowing what men are like in general. Meta also described the part of the post that described poisonous snakes as "implicit comparison between men and snakes" and arguably violating.

The Board asked Meta 14 questions in writing, all of which Meta answered. The questions addressed issues related to the criteria, internal guidelines and automated processes for distinguishing qualified and unqualified behavioral statements; how the accumulation of strikes impacts users on Instagram; internal escalation guidelines for at-scale reviewers; and how at-scale reviewers evaluate context, intent, and the accuracy of statements.

7. Public comments

The Oversight Board received and considered 13 public comments related to these cases. One of the comments was submitted from Asia Pacific and Oceania; two were submitted from Central and South Asia; six from Europe; one from Latin America and the Caribbean; and three from the United States and Canada.

The submissions covered the following themes: the significance of gender-based violence worldwide; the frequency of incorrect removals of content shared by women condemning gender-based violence and the need for change; the lack of clarity in Meta's policies and the ineffectiveness of its appeals systems including automated moderation; and the lack of contextual approach to content governance.

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

8. Oversight Board analysis

The Board examined whether these posts should be restored by analyzing Meta's content policies, human rights responsibilities, and values. The Board also assessed the implications of these cases for Meta's broader approach to content governance.

The Board selected these appeals because they offer the potential to explore how Meta's Hate Speech rules and their enforcement allow for condemnation and awareness raising of gender-based violence, an issue the Board is focusing on through its strategic priorities of gender, hate speech, and treating users fairly.

8.1 Compliance with Meta's content policies

I. Content rules

The Board finds that the first post does not violate any Meta content policy. While the Community Guidelines apply to Instagram, Meta states that "Facebook and Instagram share content policies. ´ nt that is considered violating on Facebook is also considered violating on Instagram."

Meta ultimately found that this post was a qualified behavioral statement and did not violate Facebook's Hate Speech policy. While the statement that "men murder, rape and abuse women mentally and physically - all the time, every day" may be susceptible to different interpretations, the Board agrees with Meta's ultimate conclusion that the post taken as a whole is not violating. Rather than being a generalization about all men, or even the majority of men, the principal focus of the post is to reassure victims of gender-based violence that they are not to blame and encourage them to speak out. The user refers to the International Day for the Elimination of Violence Against Women and then provides a helpline number and shares information about local support organizations for victims of intimate partner violence.

The post discusses that women have little space to speak about these experiences, that victims are not to blame, and that men perpetrate acts of violence against women. Within this broader language, the statement that "Men murder, rape and abuse women mentally and physically - all the time, every day" describes the actions of those men who commit violence against women. In this context, this statement is also better understood as assurance to other victims of domestic violence that they are not alone. It is therefore a non-violating qualified statement.

For some Board Members, the global context of violence against women is also relevant to the analysis, as the content reflects and raises awareness of a broader worldwide societal phenomenon, further reinforcing that read within the context of the post, the statement was not an assertion that all men are rapists or murderers. On the other hand, other Board Members do not believe that such broad and contested sociological considerations such as root cause assessments or analysis of power differentials should be used to interpret the statement, believing that it could invite controversial interpretations of what constitutes hate speech. The majority of the Board, though cognizant of the societal phenomenon of violence against women and the debates around its root causes, did not rely on them in order to reach its conclusion that the statement was a "qualified" one.

Some Board Members disagreed with the majority's interpretation of this post. For these Members, the user posted a clearly unqualified behavioral statement that men "murder, rape and abuse" women "all the time, every day." Instead, they believe the post violates Meta's hate speech rules.

The Board also finds that the second post does not violate any Meta content policy. The Board finds that assessing the post as a whole, as Meta did with the first post, shows that this is not an "expression of contempt" against men, as prohibited by the Hate Speech policy. Meta argues that the user's statement that she does not hate all men does not impact the assessment of other parts post. The Board disagrees.

Again, when reading the post in its entirety, the content does not express contempt against all men, but expresses strong condemnation of violence against women and men who commit it. While the user states she is a "man-hater," she both explains that this does not mean she hates all men and describe man hating as being defined by discussing fear and condemning violence against women. The user's analogy to the fear of venomous snakes, while disturbing on the surface, actually strengthens the Board's conclusion that the post as a whole is not a condemnation of all men. Not all snakes are venomous; most are harmless. But the user is pointing out that the fear of venomous snakes brushes off onto all snakes, causing many or most human to be frightened of snakes as a class.

Some Members disagreed and thought the second post was an expression of contempt, and thus a violation of Meta's rules. A subset of these Members believed the post should remain off the platform and thus dissent from any decision to restore the post whose language, they claim, could lead to negative unintended consequences for both men and women.

The Board finds the second post to be more complex to assess than the first post. While the first post should have been more easily recognized as qualified, for the second post nuanced analysis of the entire post and its language was key to understand it was not an expression of contempt. The Board agrees that the post is ultimately a condemnation of violence against women and discusses the roots of gender-based hate, thus a majority decides to restore it to the platform.

Finally, the Board agrees that the content of these posts does not create an environment of intimidation or promote offline violence, and consequently does not violate the Hate Speech policy rationale. The Board finds this post seeks to diminish offline violence against women and falls directly within Meta's paramount value of "Voice." For this reason, the Board also finds that removing the content was not consistent with Meta's values.

II. Enforcement action

The Board notes that Meta's review and appeal process for the first post used two at-scale reviewers seven times at different levels of review. In other words, the same two people were asked to review decisions that they themselves had taken earlier, rather than refer the secondary decisions to different reviewers. The Board is concerned that the effectiveness of the appeal and HIPO reviews here may have been undermined by this approach. In the "Wampum belt" case, the Board expressed concerns about Meta's review and appeal system, and requested an evaluation of accuracy rates when content moderators are informed that they are engaged in secondary review now that the initial determination was contested. Meta **responded** that it is still exploring the most efficient way to provide reviewers with additional information to maximize the accuracy of their

reviews while ensuring consistency and scalability. Meta should consider adjusting its relevant protocol to send review jobs to different reviewers than those who previously assessed the content to improve the accuracy of decisions made upon secondary review.

The Board is further concerned about the pressure on at-scale reviewers to assess content that may require more complex policy assessment in a short amount of time, often mere seconds. The Board has previously expressed concern about the limited resources available to moderators and their capacity to prevent the kind of mistakes seen in these cases ("Wampum belt" case, "'Two buttons' meme" case).

III. Transparency

The Board **welcomes** recent changes Meta has made in response to the Board's recommendations to make its account strikes and penalty system fairer and clearer. However, Meta does not provide information in its **Transparency Center** about the consequences of Instagram strikes specifically, as it currently does for Facebook strikes. An **Instagram Help Center article** shares some penalties Meta applies to Instagram accounts when they accumulate strikes, but this is less accessible. It is also not comprehensive, as it does not mention limits on the ability to create live videos, for example. To treat users fairly, Meta should clearly explain and share Instagram-specific information in the Transparency Center alongside the information about Facebook strikes and penalties.

8.2 Compliance with Meta's human rights responsibilities

The Board finds that Meta's initial decisions to remove both posts are inconsistent with Meta's human rights responsibilities as a business.

Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including about politics, public affairs, and human rights (**General Comment No. 34** (2011), Human Rights Committee, paras. 11-12). Moreover, "the Internet has become the new battleground in the struggle for women's rights, amplifying opportunities for women to express themselves" (**A/76/258** para. 4). Empowering women to freely express themselves enables the realization of their human rights (**A/HRC/Res/23/2**; (**A/76/258** para. 5).

The **Joint Declaration on Freedom of Expression and Gender Justice**, a statement by the freedom of expression experts in the UN and regional human rights systems, discusses the importance of protecting speech that calls attention to gender-based violence. It states that "when women speak out about sexual and gender-based violence, states should ensure that such speech

enjoys special protection, as the restriction of such speech can hinder the eradication of violence against women." As social media is an important pathway to raise awareness about intimate partner violence and women's rights, and in alignment with its company values, the Board believes Meta should take a similar approach.

Where restrictions on expression are imposed by a state, they must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's voluntary human rights commitments, both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression" (**A/74/486**, para. 41).

I. Legality (clarity and accessibility of the rules)

The principle of legality requires rules that limit expression to be clear and publicly accessible (General Comment No. 34, at para. 25). Legality standards further require that rules restricting expression "provide sufficient guidance to those charged with their execution to enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (**A/HRC/38/35** at para. 46). People using Meta's platforms should be able to access and understand the rules and content reviewers should have clear guidance on their enforcement.

Meta's approach to enforce its hate speech policy raises serious legality concerns with respect to both rules analyzed by the Board. The Board's main concern is that Meta states that the Hate Speech policy is "designed to allow room for raising awareness of gender-based violence." However, neither the public-facing policy nor its internal guideline documents contain clear guidance to ensure that posts like these would not be mistakenly removed. The public-facing policy rationale mentions that someone else's hate speech can be shared to condemn it or raise awareness, but that does not apply here. The Board agrees that Meta's policies should permit expression that condemns and raises awareness of gender-based violence, when the content does not create an environment of intimidation or promote offline violence, and recommends that its policies more clearly reflect this.

For the Tier 1 hate speech rules around qualification relevant to the first post, Meta's internal guidelines mean that at-scale moderators would find it almost impossible to reach the correct one. Meta relied on a series of contextual cues to determine the first post was non-violating once it was identified by the Board, but these are not included in its internal guidance for

moderators. Meta informed the Board that “it can be difficult for at-scale content reviewers to distinguish between qualified and unqualified behavioral statements without taking a careful reading of context into account.” However, the guidance to reviewers, as currently drafted, limits the possibility of contextual analysis significantly, even when there are clear cues within the content itself that it raises awareness about gender-based violence.

Further, Meta stated that because it is challenging to determine intent at scale, its internal guidelines instruct reviewers to default to removing behavioral statements about protected characteristic groups when the user has not made it clear whether the statement is qualified or unqualified. This further reinforces the Board’s concern that moderators would remove non-violating content that condemns or raises awareness of gender-based violence. Meta states that content such as the first post on anti-gender-based violence should be allowed on its platforms, but at the same time the company’s internal guidance to human reviewers seems to lead to the opposite outcome in practice.

For the Tier 2 hate speech rules around expressions of contempt relevant to the second post, the Board finds the public guidance around expressions of hate to be clearer. However, it is similarly questionable how Meta allows for condemnation and awareness raising in relation to this rule. Meta again told the Board in its description of this case that it “allow[s] people to raise awareness of violence against women” and to “share their experiences or call out intolerance.” Meta’s position that additional language within the post that negated or nuanced the expression of contempt were not relevant reinforces the Board’s concern that there is no guidance in place to ensure that Meta’s described allowance of awareness raising of gender-based violence exists in practice.

II. Legitimate aim

Any state restriction on expression should pursue one of the legitimate aims listed in the ICCPR, which include the "rights of others." According to the Hate Speech policy rationale, it aims to protect users from an “environment of intimidation and exclusion” and to prevent offline violence. Therefore, Meta’s Hate Speech policy, which aims to protect people from the harm caused by hate speech, has a legitimate aim that is recognized by international human rights law standards.

III. Necessity and proportionality

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

companies should consider a range of possible responses to problematic content beyond deletion to ensure restrictions are narrowly tailored (**A/74/486** para. 51).

In previous hate speech cases, the Board has looked to the **Rabat Plan of Action** to assess the necessity and proportionality of removing hate speech. Although it focuses on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence, the Board applies the Plan's framework by analogy to gender-based discrimination. The **Joint Declaration on Freedom of Expression and Gender Justice**, for example, supports this approach, stating that "sex and gender should be recognized as protected characteristics for the prohibition of advocacy of hatred that constitutes incitement to discrimination, hostility or violence." In both cases, the Board considered the six Rabat Plan factors (context, identity of speaker, intent of speaker, content, extent of expression, and likelihood of harm including its imminence). The Board finds that these posts pose no risk of imminent harm and thus removal of this content was not necessary. For both cases, the Board finds that the removal of this content was not necessary to protect men from harm.

The Board finds both posts to be of public interest and non-violent, directly condemning and drawing attention to gender-based violence. The first post is a factual statement, reflecting that men commit gender-based violence. The second post contains a personal opinion and its rationale against the backdrop of global violence against women. Some of the Members that found second post policy violating would nonetheless keep it on the platform for these reasons. For this minority of Members, while the second post violated Meta's hate speech Standard, the strongly expressed views in question posed no risk of likely and imminent harm and thus removing it was inconsistent with international human rights standards. (**A/68/362** at para 52-53). Therefore, both the removals and the strikes that resulted from Meta's decisions were unnecessary.

The Board is concerned that Meta's enforcement approach to gender-based hate speech may result in the disproportionate removal of content raising awareness and condemning gender-based violence and intimate partner violence against women, as seen here.

The UN Special Rapporteur on freedom of expression has recommended that companies ensure that enforcement of hate speech rules involves an evaluation of context and the harm that the content imposes on users and the public (A/74/486, para. 58 lit. d). At the same time, the Rapporteur has noted that "the scale and complexity of addressing hateful expression presents long-term challenges and may lead companies to restrict such expression even if it is not clearly linked to adverse outcomes" (A/HRC/38/35, para. 28). While the Board understands that Meta's approach to gender-based hate speech involves complex policy and enforcement questions, and

that expression that creates an environment of intimidation or promotes offline violence could be removed, as stated in previous decisions, it is concerned that the company's current approach inhibits the discussion and condemnation of gender-based violence in posts such as these. Meta should consider how the context and prevalence of gender-based violence should influence its policy and enforcement choices.

According to **UN Women,** more than 640 million women have been subject to intimate partner violence. Most of that violence is perpetrated by current or former husbands or intimate partners, reflecting societal power imbalance worldwide. Most of the 81,000 women and girls killed in 2020 died at the hands of an intimate partner or family member, which equals to a woman or girl being killed every 11 minutes in their home. Although most people who kill women are men, Meta prohibits the phrase "Men kill women" absent additional explanation.

Multiple public comments raised the impact of gender-based violence in society worldwide (e.g., PC-11023 by Karisma Foundation (Colombia), PC-11012 by Digital Rights Foundation, PC-11026 Women's Support and Information Center). The public comment by the Digital Rights Foundation (PC-11012) states that "even in cases that use 'all men,' the intention is often to shed light on the gender hierarchy in society rather than literally condemning all men as violent perpetrators." It also reiterated that "the alarming prevalence of this phenomenon globally and that most violent crimes, including intimate partner violence towards all genders, are statistically largely perpetuated by men."

The Cyber Rights Organization stated (PC-11025) that many gender-based violence survivors that speak up and generate awareness see their discourse censored online. Additionally, the public comment by Dr. Carolina Are notes that content raising awareness of gender-based violence is often mistakenly removed while misogynistic content remains online, citing several studies (PC-10999). Experts consulted by the Board in this case state that social media policies that prohibit hate speech that are sex and gender-neutral may inadvertently result in challenges for raising awareness of violence against women and inadequate enforcement have a profound impact on victims, leading to women changing their online behavior by limiting their interactions and self-censoring.

The Board finds that within this context, it is critical that statements that condemn and raise awareness of gender-based violence, and do not create an environment of intimidation or promote offline violence, not be mistakenly removed. The Board's concern that this may be happening is particularly pronounced given that an allowance for this type of content, while highlighted by Meta, is not communicated clearly to the public and the guidance provided to moderators is confusing. To

address this, Meta should clarify its public rules and provide appropriate guidance to moderators that better reflects this allowance.

9. Oversight Board decision

The Oversight Board overturns Meta's decisions to remove both posts.

10. Recommendations

A. Content policy

1. To allow users to condemn and raise awareness of gender-based violence, Meta should include the exception for allowing content that condemns or raises awareness of gender-based violence in the public language of the Hate Speech policy. The Board will consider this recommendation implemented when the public-facing language of the Hate Speech Community Standard reflects the proposed change.

B. Enforcement

2. To ensure that content condemning and raising awareness of gender-based violence is not removed in error, Meta should update guidance to its at-scale moderators with specific attention to rules around qualification. This is important because the current guidance makes it virtually impossible for moderators to make the correct decisions even when Meta states that the first post should be allowed on the platform. The Board will consider this recommendation implemented when Meta provides the Board with updated internal guidance that shows what indicators it provides to moderators to grant allowances when considering content that may otherwise be removed under the Hate Speech policy.

3. To improve the accuracy of decisions made upon secondary review, Meta should assess how its current review routing protocol impacts accuracy. The Board believes Meta would increase accuracy by sending secondary review jobs to different reviewers than those who previously assessed the content. The Board will consider this implemented when Meta publishes a decision, informed by research on the potential impact to accuracy, whether to adjust its secondary review routing.

C. Transparency

4. To provide greater transparency to users and allow them to understand the consequences of their actions, Meta should update its Transparency Center with information on what penalties are [redacted]iated with the accumulation of strikes on Instagram. The Board appreciates that Meta has provided additional information about strikes for Facebook users in response to Board

recommendations. It believes this should be done for Instagram users as well. The Board will consider this implemented when the Transparency Center contains this information.

***Procedural note:**

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

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

Volver a Decisiones de casos y opiniones consultivas sobre políticas

RECURSOS

PREGUNTAS FRECUENTES

AVISOS DE PRIVACIDAD

CONDICIONES DE USO

POLITICA DE COOKIES