**The Case of Violence against women**

**Case Analysis**

***Meta-Data*:**

* **Case Number**: 2023-002-IG-UA & 2023-005-IG-UA
* **Date of decision**: July 12, 2023
* **Featured case**: N/A
* **Region**: Europe
* **Country**: Sweden
* **Type of expression**: Electronic / Internet-based Communication
* **Judicial Body**: Oversight Board
* **Type of law**: Meta’s content policies, International Human Rights Law
* **Main Themes**: Hate speech
* **Outcome**: Overturned
* **Status**: Closed
* **Tags**: Hate speech, Oversight Board Content Policy Recommendation, Oversight Board Enforcement Recommendation, Oversight Board Transparency Recommendation, Hate Speech Community Standard, nstagram Community Guidelines, Gender-based violence

***Analysis:***

* **Summary and Outcome**:

On July 12, 2023, the Oversight Board overturned Meta’s two decisions to remove two posts condemning and raising awareness about violence against women. Meta originally removed the content because it violated its Hate Speech policy as the first post included a statement that was characterized by two at-scale reviewers as an unqualified behavioral statement and the second post included a statement deemed as an expression of contempt against men. The Board found both posts to be non-violating of the policy as they aimed to raise awareness about violence against women and did not promote offline harm or create an environment of discrimination against men. The Board also decided that the posts aligned with Meta’s value of “Voice” as it aimed to raise awareness. Additionally, the Board decided that the removal of the posts was inconsistent with Meta’s responsibilities under human rights as it didn’t meet the elements of legality, necessity, or proportionality. The Board recommended Meta modifies its policies to include an exception that allows content raising awareness about gender-based violence.

* **Facts**:

An Instagram user in Sweden posted a video with Swedish audio and its transcription depicting a woman retiling her experience in an abusive relationship without any graphic details of the violence she endured. The video was accompanied with a caption that highlighted the blaming of victims in gender-based violence and provided a helpline for victims of domestic violence and mentioned the International Day for the Elimination of Violence against Women to show abused women they were not alone. The caption also said, “men murder, rape and abuse women mentally and physically - all the time, every day.” The post was viewed 10,000 times.

After a Meta classifier identified the content to be potentially violating Meta’s Hate Speech policy, two reviewers agreed and removed the post and gave the user a strike. After the decision was appealed by the user, it was sent to a High Impact False Positive Override (HIPO) channel which is tasked with identifying wrongfully removed content. However, the decision was upheld after the content was sent to the same human reviewers. In total, those two human reviewers examined the content seven times and found it violating each time. The content was finally restored after the Board selected this case and Meta determined the removal was in error, the strike was reserved as well.

While assessing the first post, the Board received another appeal from the same user. The second appeal concerned an Instagram video of a woman speaking Swedish and saying that although she was a man-hater, she didn’t hate all men, and her feelings of hatred stem from fear of violence. She further compared the men who committed violence against women to venomous snakes. The poster of user asked men who were “allies” to help women in their fights in the caption. The post was viewed 150,000 times.

Meta removed the content the second post for violating the Hate Speech policy and applied a strike against the account preventing it from creating live videos after user reports. Despite appeals by the user, Meta upheld the removal decision which led the user to appeal the Board for the second time.

Since 2017, several digital campaigns highlighted Facebook’s wrongful removal of phrases used to call attention to gender-based violence such as “men are trash” on grounds of being anti-men.

* **Decision Overview**:

The Oversight Board analyzed whether the removal of two posts discussing domestic violence against women and using harsh language against men followed Meta’s Hate Speech policy, Meta’s values and Meta’s responsibilities under international human rights law.

In their first appeal, the user submitted that the removal of their post hindered an important discussion about domestic violence and that their intention was solely to show abused women they were not alone. In the second appeal, they highlighted that the goal of their post was to discuss the problem of men committing violence against women and not to spread hatred against men.

As for Meta, the company determined that the first decision was made in error as the post did not violate the Hate Speech policy as the full context of the post made the sentence “men murder, rape and abuse women mentally and physically - all the time, every day” a “qualified behavioral statement”. Meta reached this later decision by considering several factors including the noting of the International Day for the Elimination of Violence Against Women, providing helplines and experience of violence being described. All those factors provide the user’s intent was to raise awareness which should have allowed the content to stay on Instagram despite not squarely following Meta’s policies.

In answering one of the Board’s questions, Meta clarified that reviewers did not consider intent as its policies do not allow it, especially in cases of hate speech, to ensure consistent enforcement and avoid bias. When the Board asked Meta to provide existing criteria that would lead the reviewers to the correct outcome in this case, Meta recognized that this case did not fit into its policies, however, the company expected its reviewers to understand the content is not violating.

However, Meta stressed its decision to remove the second post reasoning that it contained expression of hatred towards women which is violating of its Hate Speech policy. The expression of hate was the subject of the video describing herself as a man-hater. Meta noted that the content might have violated other elements of the policy, however, the removal decision was made based on the expression of hate.

The Board asked meta 14 questions in writing about 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. Meta answered all the questions.

*Compliance with Meta’s content policies and values*

1. Content Rules

The Board agreed that the first post did not violate any of Meta’s content policies. The Board further agreed that while the statement “men murder, rape and abuse women mentally and physically - all the time, every day” might be understood in many ways, it constituted a qualified behavioral statement as the post didn’t aim to be a generalization about men but to reassure women who have been victimized by gender-based violence. This was further supported by the inclusion of helpful resources and the reference to the International Day for the Elimination of Violence Against Women. The Board further stressed that the statement was used to describe the actions of abusive men and blame them within a culture that blames their victims.

Some Board members also relied on the global context of violence against women in their analysis. They decided that the content raised awareness about a worldwide issue rather than a generalization about men. However, other members declined to consider the global context as they feared it such broad and contested considerations may cause controversial interpretations on the meaning of hate speech. Majority of the Board did not consider the societal phenomenon of violence against women and the debate around its root causes to decide that the statement amounted to a qualified behavioral statement. A minority of the Board disagreed with the decision and considered the statement to be an unqualified statement that violated Meta’s Hate Speech policy.

The Board disagreed with Meta on the second post as it found non-violating of Meta’s content policies as assessing the post as a whole showed it is not an expression of contempt. The Board further decided that while the user described herself as a “man-hater”, she made it clear that she did not hate men and that her hate was a result of fear and a way to condemn violence against women rather than an expression of contempt against men. The Board considered that the venomous snakes analogy to be another indicator as most snakes are harmless, however, the fear of venomous snakes brushes off onto all snakes as the user explained. Some members disagreed with statement and identified the second post as an expression of contempt violating Meta’s policies. A subset of those members believed the post should not be resorted to avoid negative consequences for men and women.

The Board found the assessment of the second post more complex than the assessment of the first post. The first post could have been easily recognized as a qualified behavioral statement while the second post required a nuanced analysis of the whole post and its language. The Board decided that the post is a condemnation of violence against women and the majority decided to restore it.

The Board further agreed that the posts did not promote offline violence thus were not violating of the Hate Speech policy rationale. The Board found the posts to fall within Meta’s value of “Voice” as they aimed to decrease offline violence against women thus removing the posts was inconsistent with Meta’s values.

1. Enforcement action

The Board noted Meta’s review and appeal process using the same two reviewers seven different times for review, meaning they reviewed their own decisions rather than using different reviewers. The Board expressed its concern that the effectiveness of the appeal and the HIPO reviews was impacted by that approach. The Board recalled the “Wampum belt” case to reiterate its concerns regarding Meta’s review and appeal system and its accuracy.

The Board further stressed that there is pressure imposed on at-scale reviewers to assess complex content requiring nuanced analysis in a short time, often mere seconds. The Board highlighted its concern about the limited sources available to moderators and their capacity to prevent mistakes by referencing the cases of “Wampum belt” and “Two buttons meme”.

1. Transparency

While the Board welcomed Meta’s implementation of the Board’s recommendations to clarify and modify its strikes and penalty systems, it highlighted that Meta doesn’t provide information about the consequences of Instagram strikes to its users in its Transparency Center. While Meta shared some of the penalties imposed on Instagram accounts accumulating strikes in an Instagram Help Center article, the Board deemed it less accessible and not comprehensive as it didn’t mention all the penalties.

*Compliance with Meta’s human rights responsibilities*

The Board noted the importance of the internet as a tool for women to express themselves and talking about their struggles, including gender-based violence and how unlawfully restricting speech raising awareness about gender-based violence would lead to hindering the eradication of violence against women. In order to assess whether Meta has lawfully restricted the user’s freedom of expression, the Board implemented the three part test stipulated in article 19 of the International Covenant on Civil and Political Rights (ICCPR).

1. Legality (clarity and accessibility of the rules)

The principle of legality requires rules limiting expression to be clear, publicly accessible and provide guidance for those who apply them. The Board found Meta’s approach to enforce its Hate Speech policy failed to fulfill the principle of legality as neither public-facing policy nor internal guidelines include sufficiently detailed guidance on speech raising awareness that may use hate speech. While the public-facing policy includes that users can use someone else’s hate speech to condemn it or raise awareness, there is nothing in the policy that applies to the cases at hand. The Board recommended Meta should allow expression that condemns and raises awareness about gender-based violence in cases where the content doesn’t promote offline violence or create an intimidating environment.

As for the first post, Tier 1 hate speech rules of qualification were applied. When the Board referred the case to Meta, the company used contextual analysis to backtrack its original decision, however, such analysis was not available for the reviewers preventing them from ever reaching the right decision even when there were clear cues within the content that it raises awareness about gender-based violence. Meta, itself, admitted to the board that reviewers face obstacles in distinguishing between qualified and unqualified behavioral statements.

Moreover, Meta admitted that the internal guidelines instruct reviews to default to removing behavioral statements in cases where it’s unclear whether the statement is qualified or not as it is challenging to determine intent. This furthered the Board’s concern that reviewers would remove non-violating content that raised awareness about gender-based violence.

Tier 2 hate speech rules on expression of contempt were applied to the second post. While the Board recognized the public-facing policy to be clearer that Tier 1 hate speech rules, it was still concerned on Meta’s approach to deal with condemnation and raising awareness in relation to the rule. The Board stressed its concern that there was no guidance to ensure the existence of Meta’s allowance of awareness raising content due to Meta’s position that additional language present in the same video should not be used to analyze statements of contempt.

1. Legitimate aim

Article 19 of the ICCPR provide an exhaustive list of legitimate aims which includes the protection of rights of others. The Board recognized that the Hate Speech policy aims to protect Meta’s users from harm caused by hate speech such as offline violence and causing an environment of discrimination.

1. Necessity and proportionality

The principle of necessity and proportionality require restrictions on freedom of expression to be appropriate to and among the least intrusive measures to achieve their protective functions which mean that social media platforms should consider a range of ways to deal with problematic content other than deletion.

In its analysis, the Board implemented the Rabat Plan of Action which provide a six part test to assess hate speech through the elements of; context, identity of the speaker, intent of the speaker, content, extent of expression and likelihood of harm including imminence. The Board found that the posts did not fulfill the test as they didn’t pose any risk of imminent harm, thus the Board found their removal to be unnecessary to protect men from harm.

Furthermore, the Board found the posts to be of public interest and non-violent as they directly condemned and raised awareness to gender-posed violence. The Board stressed that the first post represented a factual statement that men committed gender-based violence and the second post included a personal opinion caused by the global phenomenon of violence against women. Members of the Board who found the second post to be violating of the Hate Speech policy agreed that the post should remain on Instagram as it doesn’t pose likely or imminent harm, making its removal unnecessary.

The Board stressed that these cases highlight how Meta’s enforcement approach to gender-based hate speech can result in disproportionate removal of content raising awareness and condemning violence against women. The Board recommended Meta to consider the context and prevalence of gender-based violence in its policy and enforcement choices to allow content that raises awareness and does not promote offline violence or create an environment of intimidation. An example of Meta’s enforcement approach impacting raising awareness is the prohibiting of the phrase “men kill women” when UN women reported that a woman or a girl is killed every 11 minutes in their home.

*Policy Recommendations*

1. Content policy

The Board recommended Meta includes an exception in its Hate Speech policy to allow content that condemns or raises awareness of gender-based violence.

1. Enforcement

The Board recommended Meta updates its internal guidance for reviewers to include a sufficient guide on what constitutes a qualified behavioral statement and what does not since the current guide make it impossible for reviewers to reach the correct decisions in order to avoid wrongfully removing content condemning and raising awareness about gender-based violence.

The Board also recommended Meta send secondary review jobs to different reviewers than those who took the initial decision regarding the content to improve the accuracy of decisions made upon second review.

1. Transparency

The Board recommended Meta updates its Transparency Center with information on all penalties imposed when accounts accumulate strikes on Instagram like it does for Facebook in order to provide greater transparency and foreseeability for Instagram users.

***Direction:***

* **Outcome**: Expands Expression

The Oversight’s Board decision to overturn Meta’s two original decisions expanded freedom of expression by allowing speech condemning or raising awareness about gender-based violence even if it includes harsh and offensive language to some groups as long as it doesn’t promote violence against these groups or create an environment of discrimination against them. Through this decision, the Board ensures that Meta’s platform is a safe medium for women to express themselves and raise awareness about their struggles. Furthermore, the Board negates the possibility of women self-censoring themselves due to unlawful restrictions and removals of content raising awareness about gender-based violence.

***Perspective***:

* **Related International and/or regional laws**:
* United Nations Guiding Principles on Business and Human Rights (2011)

The Board referred to this instrument to highlight Facebook’s businesses’ human rights responsibilities.

* ICCPR, art. 19

The Board assessed Meta’s responsibilities towards human rights through the lens of freedom of expression. It analysed whether Meta’s action met the three-part test stipulated in Article 19(3).

* UNHR Comm., General Comment No. 34 (CCPR/C/GC/34)

While employing the three-part test to assess if Meta’s actions allowed expression to be limited, the Board referred the General Comment for guidance. The Board also referenced the General Comment to underline how public morals differ widely.

* UN Special Rapporteur on freedom of opinion and expression, (A/76/258) (2021)

The Board reiterated the role of the internet in raising awareness about women’s struggles and in empowering women to express themselves.

* UN Special Rapporteur on freedom of opinion and expression, (A/74/486) (2019)

The Board stressed Meta should consider a range of restrictions on problematic content other than removal as provided in this report.

* UN Special Rapporteur on freedom of opinion and expression, (A/HRC/38/35) (2018)

The Board remarked that the UN Special Rapporteur on freedom of expression has proposed that the three-part test shall be used as a framework to guide platforms in content moderation practices.

* UN Special Rapporteur on freedom of opinion and expression, (A/68/362) (2013)

The Board supported its stance that removing the content would be unnecessary as it didn’t pose likely and imminent harm through referencing this report.

* UNHRC, Resolution on “The role of freedom of opinion and expression in women’s empowerment”, (A/HRC/Res/23/2) (2013)

The Board highlighted the relationship between freedom of expression of women and the realization of their human rights by referencing this resolution.

* OHCHR, Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (A/HRC/22/17/Add.4) (2011).

The Board referred to this instrument to guide its analysis on whether the content constituted hate speech that was necessary to remove in order to protect the rights of men.

* **General Law Notes**:

Oversight Board Decisions:

* “Mention of the Taliban in news reporting” case (2022-005-FB-UA)

The Board referenced this decision to highlight its abiding concern about the vagueness of Meta’s strike system for its users.

* “Wampum belt” case (2021-012-FB-UA)

The Board recalled its concerns regarding the process and accuracy of Meta’s appeal and review system as well as the limited sources and time available for reviewers to implement contextual analysis under the Hate Speech policy presented earlier in this case.

* “‘Two buttons’ meme” case ( 2021-005-FB-UA)

The Board stressed the limited sources and time available for reviewers to implement contextual analysis under the Hate Speech policy presented earlier in this case.

* “Armenians in Azerbaijan” case ( 2020-003-FB-UA)

The Board referenced this to reiterate that context is key in determining the potential for adverse outcomes under the Hate Speech policy.

***Significance***:

* **Binding or persuasive precedent within jurisdiction**; The decision establishes a binding or persuasive precedent within its jurisdiction.

According to Article 2 of the Oversight Board Charter, “For each decision, any prior board decisions will have precedential value and should be viewed as highly persuasive when the facts, applicable policies, or other factors are

substantially similar.” In addition to Article 4 of the Oversight Board Charter, “The board’s resolution of each case will be binding and Meta will implement it promptly, unless implementation of a resolution could violate the law. In instances where Meta identifies that identical content with parallel context — which the board has already decided upon — remains on Meta, it will take action by analyzing whether it is technically and operationally feasible to apply the board’s decision to that content as well. When a decision or a policy advisory opinion includes recommendations, Meta will take further action by analyzing the operational procedures required to implement the recommendations, considering those recommendations in the formal policy development process of Meta, and transparently communicating about actions taken as a result.”

* **Related Cases**: Self-generated
* **Date updated**: N/A

***Docs***:

* **Official Case Documents**:

**Examples:**

Judgment (in English)

https://www.oversightboard.com/decision/IG-H3138H6S

Press Release issued by the Oversight Board (in English)

https://oversightboard.com/news/1664046764100847-oversight-board-overturns-meta-s-decisions-in-the-violence-against-women-cases/

* **Amicus Briefs and Other Legal Authorities**

Public Comments Appendix

https://oversightboard.com/attachment/7127012427314701/