OVERTURNED

2023-011-IG-UA, 2023-012-FB-UA, 2023-013-FB-UA

United States posts discussing abortion

The Oversight Board has overturned Meta’s original decisions to remove three posts discussing abortion and containing rhetorical uses of violent language as a figure of speech.

Policies and topics

- Sex and gender equality, Health, Freedom of expression
- Violence and incitement

Region and countries

- United States and Canada
- United States

Platform

- Instagram

Attachments

United States Abortion Cases Public Comments Appendix

Case summary

The Oversight Board has overturned Meta's original decisions to remove three posts discussing abortion and containing rhetorical uses of violent language as a figure of speech. While Meta acknowledges its original decisions were wrong and none of the posts violated its Violence and Incitement policy, these cases raise concerns about whether Meta's approach to assessing violent rhetoric is disproportionately impacting abortion debates and political expression. Meta should regularly provide the Board with the data that it uses to evaluate the accuracy of its enforcement of the Violence and Incitement policy, so that the Board can undertake its own analysis.

About the cases

The three abortion-related pieces of content considered in this decision were posted by users in the United States in March 2023.

In the first case, a user posted an image of outstretched hands, overlaid with the text, "Pro-Abortion Logic" in a public Facebook group. The post continued, "We don't want you to be poor, starved or unwanted. So we'll just kill you instead." The group describes itself as supporting the "sanctity of human life."

In the other two cases, both users' posts related to news articles covering a proposed bill in South Carolina that would apply state homicide laws to abortion, meaning the death penalty would be allowed for people getting abortions. In one of these posts, on Instagram, the image of the article headline was accompanied by a caption referring to the South Carolina lawmakers as being "so pro-life we'll kill you dead if you get an abortion." The other post, on Facebook, contained a caption asking for clarity on whether the lawmakers' position is that "it's wrong to kill so we are going to kill you."

After Meta's automated systems, specifically a hostile speech classifier, identified the content as potentially harmful, all three posts were sent for human review. Across the three cases, six out of seven human reviewers determined the posts violated Meta's Violence and Incitement Community Standard because they contained death threats. The three users appealed the removals of their content. When the Board selected these cases, Meta determined its original decisions were wrong and restored the posts.

decisions were wrong and restored the posts.

Key findings

The Board concludes that none of the three posts can be reasonably interpreted as threatening or inciting violence. While each uses some variation of “we will kill you,” expressed in a mock first-person voice to emphasize opposing viewpoints, none of the posts expresses a threat or intent to commit violence. In these three cases, six out of seven human moderators made mistakes in the application of Meta's policies. The Board has considered different explanations for the errors in these cases, which may represent, as Meta’s responses suggest, a small and potentially unavoidable subset of mistaken decisions on posts. It is also possible that the reviewers, who were not from the region where the content was posted, failed to understand the linguistic or political context, and to recognize non-violating content that used violent words. Meta’s guidance may also be lacking, as the company told the Board that it does not provide any specific guidance to its moderators on how to address abortion-related content as part of its Violence and Incitement policy.

Discussion of abortion policy is often highly charged and can include threats that are prohibited by Meta. Therefore, it is important Meta ensure that its systems can reliably distinguish between threats and non-violating, rhetorical uses of violent language.

Since none of these cases are ambiguous, the errors suggest there is scope for improvement in Meta’s enforcement processes. While such errors may limit expression in individual cases, they also create cyclical patterns of censorship through repeated mistakes and biases that arise from machine-learning models trained on present-day abusive content. Additionally, these cases show that mistakenly removing content that does not violate Meta’s rules can disrupt political debates over the most divisive issues in a country, thereby complicating a path out of division.

Meta has not provided the Board with sufficient assurance that the errors in these cases are outliers, rather than being representative of a systemic pattern of inaccuracies.

The Board believes that relatively simple errors like those in these cases are likely areas in which emerging machine learning techniques could lead to marked improvements. It is also supportive of Meta’s recent improvement to the sensitivity of its violent speech enforcement workflows. However, the Board expects more data to assess Meta’s performance in this area over time.

The Oversight Board's decision

The Oversight Board overturns Meta’s original decisions to remove three posts discussing abortion.

The Board recommends that Meta:

- Provide the Board with the data that it uses to evaluate the enforcement accuracy of its Violence and Incitement policy. This information should be sufficiently comprehensive to allow the Board to validate Meta’s arguments that the type of errors in these cases are not a result of any systemic problems with Meta’s enforcement processes.

* 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 original decisions to remove two Facebook posts and one Instagram post, all of which discussed abortion. The Board finds that the three posts did not violate Meta’s Violence and Incitement policy, as they did not incite or threaten violence but were rhetorical comments about abortion policy. Meta has acknowledged that its original decisions were wrong, and that the content did not violate its Violence and Incitement policy. The Board selected these cases to examine the difficult content moderation problem of dealing with violent rhetoric when used as a figure of speech as well as its potential impact on political expression.

2. Case description and background

In March 2023, three users in the United States posted abortion-related content, two on Facebook and one on Instagram. The posts reflect different perspectives on abortion. In the first case *(Facebook group case)*, a Facebook user posted an image showing outstretched hands with a text overlay saying, “Pro-Abortion Logic.” It continues, “We don’t want you to be poor, starved or unwanted. So we’ll just kill you instead,” and has the caption, “Psychopaths...” The post was made in a public group with approximately 1,000 members. The group describes itself as supporting traditional values and the “sanctity of human life,” while opposing, among other things, the “liberal left.”

The other two cases related to users posting news articles covering a proposed bill in South Carolina that would apply state homicide laws to abortion, making people who get abortions eligible for the death penalty.

In the second case *(Instagram news article case)*, an Instagram user posted an image of a news article headline stating, “21 South Carolina GOP Lawmakers Propose Death Penalty for Women Who Have Abortions.” The caption describes the lawmakers as being “so pro-life we'll kill you dead if you get an abortion.”

In the third case *(Facebook news article case)*, a Facebook user posted a link to an article entitled “South Carolina GOP lawmakers propose death penalty for women who have abortions.” The caption asks for clarity on whether the lawmakers’ position is that “it’s wrong to kill so we are going to kill you.” Each of the pieces of content in the three cases had fewer than 1,000 interactions.

Meta uses automated systems to identify potentially violating content on its platforms. These include content classifiers that use machine learning to screen for what Meta considers “hostile” speech. In all three cases, one of these hostile speech classifiers identified the content as potentially harmful and sent it for human review. Meta informed the Board that, in each case, a human reviewer determined the post violated the Violence and Incitement Community Standard’s prohibition on death threats. Each of the three users appealed the removals. In both the Facebook group and Instagram news article cases, an additional human review upheld the original removals for violating the Violence and Incitement policy. In the Facebook news article case, on appeal, a second human reviewer found the content was non-violating. This post was then reviewed for a third time, as Meta has told the Board it generally requires two reviews to overturn an initial enforcement decision. The third reviewer found the content violated the prohibition on death threats and Meta therefore upheld its initial decision to remove the content. In total, seven human moderators were involved in assessing the content across the three cases. Four of them were located in the Asia Pacific region and three were located in the Central and South Asia region.

The three users appealed the cases to the Board. As a result of the Board selecting these cases, Meta determined that its previous decisions to remove the three pieces of content cases, Meta determined that its previous decisions to remove the three pieces of content were in error and restored the posts. Meta stated that, while the policy prohibits threats that could lead to death, none of the pieces of content included a threat.

As relevant context, the Board notes that in June 2022, the United States Supreme Court issued its decision in Dobbs v. Jackson Women's Health Organization. The decision determined that the United States Constitution does not protect the right to abortion, overruling the precedent set in Roe v. Wade and leaving the question of whether, and how, to regulate abortion to individual states. Since then, legislation has been proposed and passed in multiple states, and abortion regulation is a high-profile political issue. As mentioned above, two of the posts refer to one such proposed bill in South Carolina.

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). The Board monitors the implementation of its recommendations, and may follow up on prior recommendations in its case decisions.

When the Board selects cases like these three in which 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.

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 in respect of 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:

- Iran protest slogan (case decision 2022-013-FB-UA)
- Russian poem (case decision 2022-008-FB-UA)
- UK drill music (case decision 2022-007-IG-MR)
- Tigray Communications Affairs Bureau (case decision 2022-006-FB-MR)
- Knin cartoon (case decision 2022-001-FB-UA)
- Wampum belt (case decision 2021-012-FB-UA)
- Two buttons meme (case decision 2021-005-FB-UA)
- Protests in India against France (case decision 2020-007-FB-FBR)

II. Meta's content policies

Meta seeks to prohibit threats of violence while permitting joking or rhetorical uses of threatening language. The policy rationale for Facebook's Violence and Incitement Community Standard explains: "We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence." It further states that Meta removes content "when [it] believe[s] there is a genuine risk of physical harm or direct threats to public safety." Meta says that it tries "to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat."

Meta's rules specifically prohibit "threats that could lead to death" and "threats that lead to serious injury" of private individuals, unnamed specified persons, or minor public figures. It defines threats as including "statements of intent to commit violence," "statements advocating for violence," or "aspirational or conditional statements to commit violence."
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, when 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 Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019); Rabat Plan of Action.
- The right to life: Article 6, ICCPR.
- The right to participation in public affairs and the right to vote: Article 25, ICCPR; General Comment No. 25 (1996).

5. User submissions

All three users submitted a statement as part of their appeals to the Board.

- The user in the *Facebook group case* said that they were not inciting violence against anyone but merely pointing out the "flawed logic" of groups supporting abortion.
- The user in the *Instagram news article case* explained that they were being sarcastic and were echoing the perspective of groups opposing abortion. They further stated that Meta does not enforce the Violence and Incitement policy to protect LGBTQIA+ people who receive credible death threats.
- The user in the *Facebook news article case* said that censoring articles on issues of women's rights harms public discourse and that they did not call for anyone to be harmed. They complained that Facebook was known to miss context around posts and explained that some content creators use words like "de-life" or "unalive" rather than "kill" in attempts to avoid moderation.

6. Meta's submissions

Meta explained that while human reviewers initially assessed the content in these three cases to be violating, after the cases were appealed to the Oversight Board, the company determined they did not violate the Violence and Incitement policy, and should remain on the platforms.

In the *Facebook group case*, Meta found that the user was not making a threat against any target, but rather characterizing how they believe groups supporting abortion rationalize their position. In the *Instagram news article case*, Meta explained that it was clear that the user did not threaten violence when the post was considered holistically. Finally, in the *Facebook news article case*, Meta similarly explained that the post did not contain a threat against any target when read in the context of the entire post. The user was instead using satire to express their political views on the proposed legislation.

Meta said that it did not have further information about why six out of seven of the human reviewers involved in these cases incorrectly found the content violating. This is because Meta does not require its at-scale reviewers to document the reasons for their decisions. Meta conducted a root cause analysis, an internal exercise to determine why a mistake was made, into the removal of all three pieces of content. In each analysis, Meta determined that the mistakes were "the result of human review error, where a reviewer made a wrong decision despite correct protocols in place."

The Board asked Meta 10 questions in writing. The questions related to the challenge of interpreting the non-literal use of violent language at scale, Meta's hostile speech classifier, the training for moderators regarding the Violence and Incitement policy, and the guidance to moderators to address content that relates to abortion and/or capital punishment. All questions were answered.

7. Public comments

The Oversight Board received 64 public comments relevant to this case: 4 comments were submitted from Asia Pacific and Oceania, 5 from Central and South Asia, 6 from Europe, 4 from Latin America and the Caribbean, and 45 from the United States and Canada.

The submissions covered the following themes: abortion discourse in the United States and recent legal developments; the central role of social media in facilitating public discourse about abortion; the impact of Meta's Violence and Incitement policy and moderation practices on abortion discourse and freedom of expression; the use of violent rhetoric in political debate; and the potential effects of content labelling and fact-checking of medical misinformation, among others.

In June 2023, as part of ongoing stakeholder engagement, the Board consulted representatives of advocacy organizations, academics, and other experts on issues relating to the moderation of abortion-related content and hostile speech. Roundtable meetings were held under the Chatham House rule. Participants raised a variety of issues, including the contextually relevant use of the word "kill" in abortion discourse, the importance of context when assessing possible death threats, fact checking medical misinformation, and moderating satire and humor in abortion discourse, among others. The insights provided at this meeting were valuable, and the Board extends its appreciation to all participants.

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

8. Oversight Board analysis

The Board selected these cases to examine the difficult content moderation problem of addressing violent rhetoric when used as a figure of speech. It selected content that reflects different positions to better assess the nature of this problem. These cases fall into the Board's strategic priority of "Gender."

These cases are clear enforcement errors but represent a difficult set of problems. In one sense, they are easy to resolve as Meta and the Board agree that none of the posts threaten or advocate violence. The posts in question have already been restored, and the Board agrees that they do not violate Meta's Community Standards. However, the Board is concerned that Meta's approach to assessing violent rhetoric could have a disproportionate impact on debates around abortion. Through its analysis of these cases, the Board aims to determine whether they indicate the existence of a systemic problem in this area.

8.1 Compliance with Meta's content policies

The Board finds that the posts in these cases do not violate the Violence and Incitement Community Standard. That policy prohibits "threats that could lead to death (and other forms of high-severity violence) … targeting people or places," including "statements of intent to commit high-severity violence." None of the posts in these three cases can be reasonably interpreted as threatening or inciting violence. Each of the three posts uses some variation of "we will kill you" expressed in a mock first-person voice to characterize opposing viewpoints in the abortion debate. When read in full, none of these posts advocates violence or expresses a threat or intent to commit violence. In fact, all three posts appear to be intended to criticize the violence that the authors perceive in their opponents' positions. None of these cases are ambiguous. In each case, the violent language used is political commentary or a

caricature of views that the user opposes.
It is important Meta ensures that its systems can reliably distinguish between threats and non-violating, rhetorical uses of violent language. Discussion of abortion policy is often highly charged and divisive and can include threats that are prohibited by Meta. As threats are often directed at activists and vulnerable women, as well as medical practitioners and public figures like judges, they can have serious negative impacts on participation and political expression. At the same time, public debates about abortion often invoke violent speech in non-literal ways.

Mistaken removals of non-violating content (false positives) negatively impact expression, while mistakenly leaving up violent threats and incitement (false negatives) presents major safety risks and can suppress the participation of those targeted because of their identity or opinion. These mistakes may limit expression in individual cases but also create cyclical patterns of censorship. To the extent that content moderation machine learning models are trained on present-day abusive content, mistakes and biases will be repeated in the future. These cases regarding abortion also show that taking down false positives can disrupt political debates over the most divisive issues before a nation, thereby complicating or even precluding a path out of division.

The Board recognizes that understanding context and the non-literal use of violent language at scale is a difficult challenge. Classic examples in content moderation debates of false positives include phrases like " I will kill you for sending me spoilers!" False negatives include threats expressed in coded language that are often misunderstood and dismissed by social media platforms, an issue frequently raised by gender-rights advocates. For example, advocates complain that threats and abuse that target them are not taken seriously enough. The Board previously addressed this in the hate speech context in the Knin cartoon case. The Board recognizes this as a crucial issue because real threats of violence and death must not remain on Facebook and Instagram.

Over a series of cases, the Board has repeatedly emphasized that violent words, used rhetorically, do not necessarily convey a threat or incite violence. In the Iran protest slogan case, the Board found that the protest slogan, "Marg bar Khamenei" (literally "death to Khamenei") should not be removed given its usage in Iran. The Board restored a post that suggested "[taking the sword] out of its sheath," with a majority of the Board interpreting the post as a criticism of President Macron's response to religiously motivated violence, and not a veiled threat (Protest in India against France). The Board overturned Meta's decision to remove a poem comparing the Russian army in Ukraine to Nazis, which included the call to remove a poem comparing the Russian army in Ukraine to Nazis, which included the call to "kill the fascist... Kill him! Kill him! Kill!" (Russian Poem), finding that the quotes are an artistic and cultural reference employed as a rhetorical device. The Board also overturned Meta's removal of a UK drill music clip that referred to gun violence ("Beat at the crowd, I ain't picking and choosing (No, no). Leave man red, but you know...," UK drill music), finding that Meta should have given more weight to the content's artistic nature. The Board also criticized Meta's original decision to remove a post featuring art by an Indigenous North American artist entitled "Kill the Indian / Save the Man," a phrase used to justify the forced removal and assimilation of Indigenous children as part of historic crimes and acts of cultural genocide carried out in North America (Wampum belt). On the other hand, when evaluating content posted during escalating violence in Ethiopia, the Board upheld Meta's decision to remove a post urging the national army to "turn its gun towards the fascist" (Tigray Communication Affairs Bureau). In each of these cases, the Board has held that the meaning of posts must be evaluated in context; overly literal interpretations of speech may often lead to errors in moderation.

In each of the three cases in this decision, Meta has accepted that its initial findings were incorrect and that none of the three posts violated its policies. Meta has already restored the content to Facebook and Instagram, and the Board finds that the decision to reinstate the posts was correct.

8.2 Compliance with Meta's human rights responsibilities

Article 19, para. 2 of the International Covenant on Civil and Political Rights (ICCPR) protects "the expression and receipt of communications of every form of idea and opinion capable of transmission to others," including about politics, public affairs, and human rights (General Comment No. 34, paras. 11-12). Moreover, the UN Human Rights Committee has stated that "free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential" (General Comment No. 34, para. 20). In addition, restrictions on speech may not discriminate on the basis of political opinion (General Comment 34, para. 26). In these cases, all three pieces of content discuss abortion, a key political issue in the United States. Facebook and Instagram have become important sites for political discussion, and Meta has a responsibility to respect the freedom of expression of its users on controversial political issues.

For Meta to meet its voluntary commitments to respect human rights, its rules and procedures must meet the requirements of legality, legitimate aim, and necessity and proportionality (Article 19, para. 3, ICCPR). The Board has acknowledged that while the ICCPR does not create obligations for Meta as it does for states, Meta has committed to respect human rights as set out in the UNGPs (A/74/486, paras. 47-48).

I. Legality

The condition of legality, which requires that rules are clear and accessible to both the people subject to them and those enforcing them, is satisfied in these cases. The policy rationale of the Violence and Incitement policy makes it clear that non-threatening uses of violent language and casual statements are not prohibited.

II. Legitimate aim

Meta's rules prohibiting threats are also addressed at achieving a legitimate aim. The Violence and Incitement policy aims to "prevent potential offline harm" by removing content that poses "a genuine risk of physical harm or direct threats to public safety." This policy serves the legitimate aim of respecting the rights of others, such as the right to life (Article 6, ICCPR), as well as public order and national security (Article 19, para. 3, ICCPR). In the context of political speech, the policy may also pursue the legitimate aim of respecting others' right to participate in public affairs (Article 25, ICCPR).

III. Necessity and proportionality

The Violence and Incitement policy can only comply with Meta's human rights responsibilities if it meets the principle of necessity and proportionality, which requires that any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; [and] they must be proportionate to the interest to be protected" (General Comment No. 34, para. 34).

The Board is concerned that the rhetorical use of violent words may be linked to disproportionately high rates of errors by human moderators. In response to the Board's questions, Meta said that the posts in these cases were sent for human review seven times, and in six of those, human moderators wrongly concluded that the post contained a death threat. The "Root Cause Analysis" that Meta carried out internally to determine why these cases were wrongly decided led the company to conclude that all the mistakes were simply a result of human error, with no indication that its protocols are deficient.

In general, mistakes are inevitable among the hundreds of millions of posts that Meta moderates every month. In the first quarter of 2023, Meta either removed or placed behind a warning screen 12.4 million posts on Facebook and 7.7 million posts on Instagram under its Violence and Incitement policy. Approximately 315,000 of those Facebook posts and 96,000 of those Instagram posts were later restored, most following an appeal by the user. At this scale, even a very low error rate represents tens of thousands of errors. The true rate of mistakes is hard to estimate for a number of reasons, including that not all decisions will be appealed, not all appeals are correctly resolved, and violating posts that are not identified (false negatives) are not quantified.

The Board expects that well-trained human moderators, with language proficiency and access to clear guidance, should not often make mistakes in clear situations like the content in these cases. However, the Board does not have adequate data from Meta to conclude whether the erroneous decisions taken by these human moderators represent a systemic problem for content discussing abortion.

Rates of false positives and false negatives are usually inversely correlated. The more that Meta tries to ensure that all threats violating the Violence and Incitement policy are detected and removed, the more likely it is to make the wrong call on posts that use violent language rhetorically in a non-threatening way. Meta has told the Board that "distinguishing between the literal and non-literal use of violent language is particularly challenging because it requires consideration of multiple factors like the user's intent, market-specific language nuances, sarcasm and humor, the nature of the relationship between people, and the adoption of a third-party 'voice' to make a point." While Meta explained that it recognizes the harm that over-enforcement can do to free expression, it believes that the risk to people's safety that threatening speech poses justifies its approach.

The Board agrees that Meta must be careful in making changes to its violent threats policy and enforcement processes. It could be potentially disastrous to allow more rhetorical uses of violent language (thereby reducing false positives) without understanding the impact on targets of veiled and explicit threats. This is particularly the case in the context of posts about abortion policy in the United States, where women, public figures like judges, and medical providers have reported experiencing serious abuse, threats, and violence.

In these three cases, the removals were not necessary or proportional to the legitimate aim

In these three cases, the removals were not necessary or proportional to the legitimate aim pursued. However, the Board does not yet have sufficient information to conclusively determine whether Meta's Violence and Incitement policy and enforcement processes are necessary and proportionate. The Board has considered different explanations for the mistakes in these cases. These cases may represent, as Meta's responses suggest, a small and potentially unavoidable subset of mistaken decisions on posts. It is also possible that the reviewers, who were not from the region where the content was posted, failed to understand the linguistic or political context, and to recognize non-violating content that used violent words. Meta's guidance may also be lacking, as the company told the Board that it does not provide any specific guidance to its moderators on how to address abortion-related content as part of its Violence and Incitement policy.

Given the scale of the risks to safety that are involved, the Board is wary of recommending major changes to Meta's policies or enforcement processes without better understanding the distribution of errors and the likely human rights impacts of different options. However, the Board remains concerned by the potential wider implications raised by these apparently simple enforcement errors that are missed by Meta and appealed to the Board. Meta has not been able to provide the Board with sufficient assurance that the errors in these cases are outliers and do not represent a systemic pattern of inaccuracies in dealing with violent rhetoric in general or political speech about abortion in particular. The Board therefore requests and recommends that Meta engages in a collaborative process to identify and assemble more specific information that the Board can use in its ongoing work to help Meta align its policies with human rights norms.

IV. A future of continuous improvement and oversight

The Oversight Board expects Meta to demonstrate continual improvement in the accurate enforcement of its policies. Meta has said that these errors are not attributable to any shortcoming in its policies, training, or enforcement processes. If this is the case, then the Board suggests that these cases are useful examples of areas where improvements in Meta's automated tools may help better align its enforcement processes with human rights. In general, in implementing technological improvements to content moderation, Meta should strive to reduce the number of false positives as much as possible without increasing the number of false negatives. Meta has explained to the Board that the classifiers used in these cases were not highly confident that the content was likely to violate its policies, and sent the content for human review. While the Board accepts that the automated interpretation of context and nuance, particularly sarcasm and satire, is difficult, this is an area where the pace of advancement across the industry is extraordinarily rapid. The Board believes that relatively simple errors like those in these cases are likely areas in which emerging machine learning techniques could lead to marked improvements.

Meta told the Board that it has improved the sensitivity of its violent speech enforcement workflows to reduce the instances of over-enforcement when people jokingly use explicit threats with their friends. After assessing a sample of content automatically removed by the hostile speech classifier, Meta tested no longer proactively deleting some types of content that the sample showed resulted in the most over-enforcement and it is assessing the results. This sort of progress and improvement is exactly what social media companies should be doing on a continual basis. The Board expects to continue helping Meta identify potential areas of improvement and to help evaluate the human rights impacts of potential changes to its rules and enforcement processes.

The Oversight Board expects Meta to share more data to enable the Board to assess improvements in performance over time, including regular detailed analyses of its experiments and the changes it makes in its efforts to improve. Previously, such as in the Two buttons meme decision, when the Board recommended that Meta develop better processes to assess sarcasm, the company stated that it implemented the recommendation without demonstrating the progress it claimed to have made. To assess the necessity and proportionality of content moderation at scale, the Board must be able to reliably evaluate whether Meta's rules and processes are appropriate to achieving Meta's aims. In this case, we have highlighted our concerns about the potential uneven enforcement of Meta's Violence and Incitement policy on political speech and noted that moderation at scale involves complex challenges and difficult trade-offs. These cases raise concerns about the possibility of disproportionate removal of political speech, specifically about abortion, when posts are more likely to use words and phrases that present an increased risk of being mistaken for violent threats.

In these cases, the Board recommends that Meta demonstrate its position with data sufficient to facilitate an analysis of the proportionality of the policy. As an achievable first step, the Board recommends that Meta begins to regularly share the data that it already holds and generates for its own internal evaluation processes, including the data it relied on to substantiate claims that its policies and procedures are necessary and proportionate. The Board expects Meta to engage in a collaborative process to identify the information that would enable the Board to analyze Meta's policies in light of the trade-offs and likely impacts of potential alternatives.

9. Oversight Board decision

The Oversight Board overturns Meta's original decisions to take down the content in all three cases.

10. Recommendations

Enforcement

1. In order to inform future assessments and recommendations to the Violence and Incitement policy, and enable the Board to undertake its own necessity and proportionality analysis of the trade-offs in policy development, Meta should provide the Board with the data that it uses to evaluate its policy enforcement accuracy. This information should be sufficiently comprehensive to allow the Board to validate Meta’s arguments that the type of enforcement errors in these cases are not a result of any systemic problems with Meta’s enforcement processes. The Board expects Meta to collaborate with it to identify the necessary data (e.g., 500 pieces of content from Facebook and 500 from Instagram in English for US users) and develop the appropriate data sharing arrangements.

The Board will consider this implemented when Meta provides the requested data.

***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.