FB-L1LANIA7

Case decision 2021-012-FB-UA

Dec. 9, 2021

Case summary

Note: On October 28, 2021, Facebook announced that it was changing its company name to Meta. In this text, Meta refers to the company, and Facebook continues to refer to the product and policies attached to the specific app.

The Oversight Board has overturned Meta's original decision to remove a Facebook post from an Indigenous North American artist that was removed under Facebook's Hate Speech Community Standard. The Board found that the content is covered by allowances to the Hate Speech policy as it is intended to raise awareness of historic crimes against Indigenous people in North America.

About the case

In August 2021, a Facebook user posted a picture of a wampum belt, along with an accompanying text description in English. A wampum belt is a North American Indigenous art form in which shells are woven together to form images, recording stories and agreements. This belt includes a series of depictions which the user says were inspired by "the Kamloops story," a reference to the May 2021 discovery of unmarked graves at a former residential school for Indigenous children in British Columbia, Canada.

The text provides the artwork's title, "Kill the Indian/ Save the Man," and identifies the user as its creator. The user describes the series of images depicted on the belt: "Theft of the Innocent, Evil Posing as Saviours, Residential School / Concentration Camp, Waiting for Discovery, Bring Our Children Home." In the post, the user describes the meaning of their artwork as well as the history of wampum belts and their purpose as a means of education. The user states that the belt was not easy to create and that it was emotional

to tell the story of what happened at Kamloops. They apologize for any pain the art causes survivors of Kamloops, noting their “sole purpose is to bring awareness to this horrific story.”

Meta’s automated systems identified the content as potentially violating Facebook’s Hate Speech Community Standard the day after it was posted. A human reviewer assessed the content as violating and removed it that same day. The user appealed against that decision to Meta prompting a second human review which also assessed the content as violating. At the time of removal, the content had been viewed over 4,000 times, and shared over 50 times. No users reported the content.

As a result of the Board selecting this case, Meta identified its removal as an “enforcement error” and restored the content on August 27. However, Meta did not notify the user of the restoration until September 30, two days after the Board asked Meta for the contents of its messaging to the user. Meta explained the late messaging was a result of human error.

Key findings

Meta agrees that its original decision to remove this content was against Facebook’s Community Standards and an "enforcement error.” The Board finds this content is a clear example of ‘counter speech’ where hate speech is referenced to resist oppression and discrimination.

The introduction to Facebook’s Hate Speech policy explains that counter speech is permitted where the user’s intent is clearly indicated. It is apparent from the content of the post that it is not hate speech. The artwork tells the story of what happened at Kamloops, and the accompanying narrative explains its significance. While the words ‘Kill the Indian’ could, in isolation, constitute hate speech, in context this phrase draws attention to and condemns specific acts of hatred and discrimination.

The Board recalls its decision 2020-005-FB-UA in a case involving a quote from a Nazi official. That case provides similar lessons on how intent can be assessed through indicators other than direct statements, such as the content and meaning of a quote, the timing and country of the post, and the substance of reactions and comments on the post.

In this case, the Board found that it was not necessary for the user to expressly state that they were raising awareness for the post to be recognized as counter speech. The Board noted internal "Known Questions" to moderators that a clear statement of intent will not always be sufficient to change the meaning of a post that constitutes hate speech. Moderators are expected to make inferences from content to assess intent, and not rely solely on explicit statements.

Two separate moderators concluded that this post constituted hate speech. Meta was not able to provide specific reasons why this error occurred twice.

The Oversight Board decision

The Oversight Board overturns Meta's original decision to take down the content.

In a policy advisory statement, the Board recommends that Meta:

- Provide users with timely and accurate notice of any company action being taken on the content their appeal relates to. Where applicable, including in enforcement error cases like this one, the notice to the user should acknowledge that the action was a result of the Oversight Board's review process.
- Study the impacts on reviewer accuracy when content moderators are informed they are engaged in secondary review, so they know the initial determination was contested.
- Conduct a reviewer accuracy assessment focused on Hate Speech policy allowances that cover artistic expression and expression about human rights

violations (e.g., condemnation, awareness raising, self-referential use, empowering use). This assessment should also specifically investigate how the location of a reviewer impacts the ability of moderators to accurately assess hate speech and counter speech from the same or different regions. Meta should share the results of this assessment with the Board, including how results will inform improvements to enforcement operations and policy development and whether it plans to run regular reviewer accuracy assessments on these allowances. The Board also calls on Meta to publicly share summaries of the results of these assessments in its quarterly transparency updates on the Board to demonstrate it has complied with this recommendation.

*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 decision to remove a post by an Indigenous North American artist that included a picture of their art along with its title, which quotes an historical instance of hate speech. Meta agreed that the post falls into one of the allowances within the Facebook Community Standard on Hate Speech as it is clearly intended to raise awareness of historic crimes against Indigenous people in North America.

2. Case description

In early August 2021, a Facebook user posted a picture of a wampum belt, along with an accompanying text description in English. A wampum belt is a North American Indigenous art form in which shells are woven together to form images, recording stories and agreements. This belt includes a series of depictions which the user says were inspired by

“the Kamloops story,” a reference to the May 2021 discovery of unmarked graves at a former residential school for Indigenous children in British Columbia, Canada.

The text provides the artwork’s title, “Kill the Indian/ Save the Man,” and identifies the user as its creator. The user then provides a list of phrases that correspond to the series of images depicted on the belt: “Theft of the Innocent, Evil Posing as Saviours, Residential School / Concentration Camp, Waiting for Discovery, Bring Our Children Home.” In the post, the user describes the meaning of their artwork as well as the history of wampum belts and their purpose as a means of education. The user states that the belt was not easy to create and that it was very emotional to tell the story of what happened at Kamloops. They go on to say that the story cannot be hidden from the public knowledge again and that they hope the belt will help prevent that happening. The user concludes their post by apologizing for any pain the artwork causes to survivors of the residential school system, saying that their “sole purpose is to bring awareness to this horrific story.”

Meta’s automated systems identified the content as potentially violating the Facebook Community Standard on Hate Speech the day after it was posted. A human reviewer assessed the content as violating and removed it that same day. The user appealed against that decision to Meta prompting a second human review which also assessed the content as violating. At the time of removal, the content had been viewed over 4,000 times, and shared over 50 times. No users reported the content. As a result of the Board selecting this case, Meta identified its removal as an “enforcement error” and restored the content on August 27. However, Meta did not notify the user of the restoration until September 30, two days after the Board asked Meta for the contents of its messaging to the user. Meta explained the late messaging was a result of human error. The messaging itself did not inform the user that their content was restored as a consequence of their appeal to the Board and the Board’s selection of this case.

A public comment by the Association on American Indian Affairs (Public Comment-10208) points out that the quote used as the title of the artwork is from Richard Henry Pratt, an

army officer who established the first federal Indian boarding school in the United States of America. The phrase summarized the policies behind the creation of boarding schools that sought to forcefully 'civilize' Native peoples and 'eradicate all vestiges of Indian culture.' Similar policies were adopted in Canada and have been found to amount to cultural genocide by the Truth and Reconciliation Commission of Canada.

The user's reference to what happened at "Kamloops" is a reference to the Kamloops Indian Residential School, a former boarding school for First Nations children in British Columbia, Canada. In May 2021, leaders of the Tk'emlúps te Secwépemc First Nation announced the discovery of unmarked graves in Kamloops. Authorities have confirmed 200 probable burial sites in the area.

The Canadian government estimates that a minimum of 150,000 Indigenous children went through the residential school system before the last school was shut down in 1997. Indigenous children were often forcibly removed from their families and prohibited from expressing any aspect of Indigenous culture. The schools employed harsh and abusive corporal punishment, and staff committed or tolerated sexual abuse and serious violence against many students. Students were malnourished, the schools were poorly heated and cleaned, and many children died of tuberculosis and other illnesses with minimal medical attention. The Truth and Reconciliation Commission concluded that at least 4,100 students died while attending the schools, many from mistreatment or neglect, others from disease or accident.

3. Authority and scope

The Board has authority to review Meta's decision following an appeal from the user whose content was removed (Charter Article 2, Section 1; Bylaws Article 3, Section 1). The Board may uphold or reverse Meta's decision, and its decision is binding on the company (Charter Article 4; Article 3, Section 5). The Board's decisions may include policy advisory

statements with non-binding recommendations that Meta must respond to (Charter Article 4; Article 3, Section 4).

When the Board selects cases like this one, where Meta subsequently agrees that it made an error, the Board reviews the original decision to help increase understanding of why errors occur, and to make observations or recommendations that may contribute to reducing errors and to enhancing due process. After the Board's decision in *Breast Cancer Symptoms and Nudity* (2020-004-IG-UA, Section 3), the Board adopted a process that enables Meta to identify any enforcement errors prior to a case being assigned to a panel (see: transparency reports, page 30). It is unhelpful that in these cases, Meta focuses its rationale entirely on its revised decision, explaining *what should have happened* to the user's content, while inviting the Board to uphold this as the company's "ultimate" decision. In addition to explaining why the decision the user appealed against was wrong, the Board suggests that Meta explain how the error occurred, and why the company's internal review process failed to identify or correct it. The Board will continue to base its reviews on the decision a user appealed.

4. Relevant standards

The Oversight Board considered the following standards in its decision:

I. Facebook Community Standards:

The Facebook Community Standards define hate speech as "a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability." Under "Tier 1," prohibited content includes "violent speech or support in written or visual form." The Community Standard also includes allowances to distinguish non-violating content:

We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.

II. Meta's values:

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

The goal of our Community Standards has always been to create a place for expression and give people a voice. […] We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Meta limits "Voice" in service of four values, and two are relevant here:

"Safety": *We are committed to making Facebook a safe place. Expression that threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook.*

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

III. Human rights standards:

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

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

- Freedom of expression: Article 19, International Covenant on Civil and Political Rights (ICCPR); General Comment No. 34, Human Rights Committee, 2011; Article 5, International Convention on the Elimination of All Forms of Racial Discrimination (ICERD); UN Special Rapporteur Report on Hate Speech, A/74/486, 2019; UN Special Rapporteur Report on Online Content Moderation, A/HRC/38/35, 2018.
- Equality and non-discrimination: Article 2, para. 1 and Article 26 (ICCPR); Article 2, ICERD; General Recommendation 35, Committee on the Elimination of Racial Discrimination, 2013.
- Cultural rights: Article 27, ICCPR; Article 15, International Covenant on Economic, Social and Cultural Rights (ICESCR); UN Special Rapporteur in the field of cultural rights, report on artistic freedom and creativity, A/HRC/23/34, 2013.
- Rights of Indigenous peoples: UN Declaration on the Rights of Indigenous People, Article 7, para. 2; Article 8, para. 1, and Article 19.

5. User statement

The user stated in their appeal to the Board that their post was showcasing a piece of traditional artwork documenting history, and that it had nothing to do with hate speech. The user further stated that this history "needed to be seen" and in relation to Meta's removal of the post stated that "this is censorship."

6. Explanation of Meta's decision

Meta told the Board that the phrase "Kill the Indian" constituted a Tier 1 attack under the Facebook Community Standard on Hate Speech, which prohibits "violent speech" targeting people on the basis of a protected characteristic, including race or ethnicity.

However, Meta acknowledged the removal of the content was wrong because the policy permits sharing someone else's hate speech to "condemn it or raise awareness." Meta noted that the user stated in the post that their purpose was to bring awareness to the horrific story of what happened at Kamloops.

Meta noted that the phrase "Kill the Indian/Save the Man" originated in the forced assimilation of Indigenous children. By raising awareness of the Kamloops story, the user was also raising awareness of forced assimilation through residential schools. In response to a question from the Board, Meta clarified that a content reviewer would not need to be aware of this history to correctly enforce the policy. The user's post stated they were raising awareness of a horrific story and therefore a reviewer could reasonably conclude that the post was raising awareness of the hate speech it quoted.

Meta informed the Board that no users reported the content in this case. Meta operates machine learning classifiers that are trained to automatically detect potential violations of the Facebook Community Standards. In this case, two classifiers automatically identified the post as possible hate speech. The first classifier, which analyzed the content, was not very confident that the post violated the Community Standard. However, another classifier determined, on the basis of a range of contextual signals, that the post might be shared widely and seen by many people. Given the potential harm that can arise from the widespread distribution of hate speech, Meta's system automatically sent the post to human review.

Meta clarified in response to the Board's questions that a human reviewer based in the Asia-Pacific region determined the post to be hate speech and removed it from the platform. The user appealed and a second human reviewer in the Asia-Pacific region reviewed the content and also determined it to be hate speech. Meta confirmed to the Board that moderators do not record reasoning for individual content decisions.

7.Third-party submissions

The Oversight Board considered eight public comments related to this case: four from the United States and Canada, two from Europe, one from Sub-Saharan Africa, and one from Asia-Pacific and Oceania. The submissions addressed themes including the significance of the quote the user based the title of their artwork on, context about the use of residential schools in North America, and how Meta's content moderation impacts artistic freedoms and the expression rights of people of Indigenous identity or origin.

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

8.Oversight Board analysis

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

8.1 Compliance with Community Standards

Meta agreed that its original decision to remove this content was against the Facebook Community Standards and was an "enforcement error." The Board finds that the content in this case is unambiguously not hate speech. This content is a clear example of 'counter speech,' where hate speech is referenced or reappropriated in the struggle against oppression and discrimination.

The Hate Speech Community Standard explicitly allows "content that includes someone else's hate speech to condemn it or raise awareness." Two separate moderators nevertheless concluded that this post constituted hate speech. Meta was not able to provide the specific reasons why this particular error occurred twice.

In the *Nazi Quote* case (2020-005-FB-UA), the Board noted that the context in which a quote is used is important to understand its meaning. In that case, the content and

meaning of the quote, the timing of the post and country where it was posted, as well as the substance of reactions and comments to the post, were clear indications that the user did not intend to praise a designated hate figure.

The Board finds that it was not necessary for the user to expressly state that they were raising awareness for the intent and meaning of this post to be clear. The pictured artwork tells the story of what happened at Kamloops, and the accompanying narrative explains its significance. While the words 'Kill the Indian' could, in isolation, constitute hate speech, assessing the content as a whole makes clear the phrase is used to raise awareness of and condemn hatred and discrimination. The content used quotation marks to distinguish the hateful phrase of its title, which in full was "Kill the Indian / Save the Man." This should have given a reviewer pause to look deeper. The way the user told the Kamloops story and explained the cultural significance of the wampum belt made clear they identified with the victims of discrimination and violence, and not its perpetrators. Their narrative clearly condemned the events uncovered at Kamloops. It was clear from comments and reactions to the post that this intent to condemn and raise awareness was understood by the user's audience.

The Board notes that Facebook's internal "Known Questions," which form part of the guidance given to moderators, instruct moderators to err on the side of removing content that includes hate speech where the user's intent is not clear. The Known Questions also state that a clear statement of intent will not always be sufficient to change the meaning of a post that constitutes hate speech. This internal guidance provides limited instruction to moderators on how to properly distinguish prohibited hate speech from counter speech that quotes hate speech to condemn it or raise awareness. As far as the Board is aware, there is no guidance on how to assess evidence of intent in artistic content quoting or using hate speech terms, or in content discussing human rights violations, where such content is covered by the policy allowances.

8.2 Compliance with Meta's values

The Board finds that the original decision to remove this content was inconsistent with Meta's values of "Voice" and "Dignity" and did not serve the value of "Safety." While it is consistent with Meta's values to limit the spread of hate speech on its platforms, the Board is concerned that Meta's moderation processes are not able to properly identify and protect the ability of people who face marginalization or discrimination to express themselves through counter speech.

Meta has stated its commitment to supporting counter speech:

As a community, a social platform, and a gathering of the shared human experience, Facebook supports critical Counterspeech initiatives by enforcing strong content policies and working alongside local communities, policymakers, experts, and changemakers to unleash Counterspeech initiatives across the globe.

Meta claims that "Voice" is the company's most important value. Art that seeks to illuminate the horrors of past atrocities and educate people on their lasting impact is one of the most important and powerful expressions of the value of "Voice," especially for marginalized groups who are expressing their own culture and striving to ensure their own history is heard. Counter speech is not just an expression of "Voice" but also a key tool for the targets of hate speech to protect their own dignity and push back against oppressive, discriminatory, and degrading conduct. Meta must ensure that its content policies and moderation practices account for and protect this form of expression.

For a user who is raising awareness about mass atrocities to be told that their speech is being suppressed as hate speech is an affront to their dignity. This accusation, in particular when confirmed by Meta on appeal, may lead to self-censorship.

8.3 Compliance with Meta's human rights responsibilities

The Board concludes that the removal of this post contravened Meta's human rights responsibilities as a business. Meta has committed itself to respect human rights under the UN Guiding Principles on Business and Human Rights (UNGPs). Its Corporate Human Rights Policy states that this includes the International Covenant on Civil and Political Rights (ICCPR).

This is the Board's first case concerning artistic expression, as well as its first case concerning expression where the user self-identifies as an Indigenous person. It is one of several cases the Board has selected where the user was seeking to bring attention to serious human rights violations.

Freedom of expression (Article 19 ICCPR)

International human rights standards emphasize the value of political expression (Human Rights Committee General Comment 34, para. 38). The scope of protection for this right is specified in Article 19, para. 2, of the ICCPR, which gives special mention to expression "in the form of art." The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) also provides protection from discrimination in the exercise of the right to freedom of expression (Article 5), and the Committee tasked with monitoring states' compliance has emphasized the importance of the right with respect to assisting "vulnerable groups in redressing the balance of power among the components of society" and to offer "alternative views and counterpoints" in discussions (CERD Committee, General Recommendation 35, para. 29).

Art is often political, and international standards recognize the unique and powerful role of this form of communication in challenging the status quo (UN Special Rapporteur in the field of cultural rights, A/HRC/23/34, at paras 3 – 4). The internet, and social media platforms like Facebook and Instagram in particular, have special value to artists in reaching new and larger audiences. Their livelihoods may depend on access to social platforms that dominate the Internet.

The right to freedom of expression is guaranteed to all people without discrimination (Article 2, para. 1, ICCPR). The Board received submissions that the rights of Indigenous people to free, prior and informed consent where states adopt legislative or administrative measures that affect those communities imply a responsibility for Meta to consult with these communities as it develops its content policies (Public Comment-10240, Minority Rights Group; see also UN Declaration on the Rights of Indigenous Peoples, Article 19). The UN Special Rapporteur on freedom of opinion and expression has raised a similar concern in the context of social media platforms' responsibilities (A/HRC/38/35, para. 54).

The content in this case engages a number of other rights as well, including the rights of persons belonging to national, ethnic or linguistic minorities to enjoy, in community with other members of their group, their own culture (Article 27, ICCPR), and the right to participate in cultural life and enjoy the arts (Article 15, ICESCR). The art of creating a wampum belt that sought to record and bring awareness to human rights atrocities and their continued legacy receives protection under the UN Declaration on Human Rights Defenders, Article 6(c), as well as the right to truth about atrocities (UN Set of Principles to Combat Impunity). The UN Declaration on the Rights of Indigenous Peoples expressly recognizes the forcible removal of children can be an act of violence and genocide (Article 7, para 2) and provides specific protection against forced assimilation and cultural destruction (Article 8, para 1).

ICCPR Article 19 requires that 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). The UN Special Rapporteur on freedom of expression has encouraged social media companies to be guided by these principles when moderating online expression, mindful that regulation of expression at scale by private companies may give rise to concerns particular to that context (A/HRC/38/35, paras. 45 and 70). The Board has employed the three-part test based on Article 19 of the ICCPR in all of its decisions to date.

I. Legality (clarity and accessibility of the rules)

The Community Standard on Hate Speech clearly allows content that condemns hate speech or raises awareness. This component of the policy is sufficiently clear and accessible for the user to understand the rules and act accordingly (General Comment 34, para. 25). The legality standard also requires 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" (*Ibid.*) The failure of two moderators to properly assess the application of policy allowances to this content indicates that further internal guidance to moderators may be required.

II. Legitimate aim

Any state restriction on freedom of expression must pursue one of the legitimate aims listed in Article 19, para. 3 of the ICCPR. In its submissions to the Board, Meta has routinely invoked aims from this list when justifying action it has taken to suppress speech. The Board has previously recognized that Facebook's Hate Speech Community Standard pursues the legitimate aim of protecting the rights of others. Those rights include the right to equality and non-discrimination, freedom of expression, and the right to physical integrity.

III. Necessity and proportionality

The clear error in this case means that the removal was obviously not necessary, which Meta has accepted. The Board is concerned that such an unambiguous error may indicate deeper problems of proportionality in Meta's automated and human review processes. Any restrictions on freedom of expression should be appropriate to achieve their protective function and should be the least intrusive instrument amongst those that might achieve their protective function (General Comment 34, para. 34). Whether Meta's content moderation system meets the requirements of necessity and proportionality depends

largely on how effective it is in removing *actual* hate speech while minimizing the number of erroneous detections and removals.

Every post that is wrongly removed harms freedom of expression. The Board understands that mistakes are inevitable, for both humans and machines. Hate speech and responses to it will always be context specific, and its boundaries are not always clear. However, the types of mistakes and the people or communities who bear the burden of those mistakes reflect design choices that must constantly be assessed and examined. This requires further investigation of the root causes of the mistake in this case, and broader evaluation of how effectively counter speech is moderated.

Given the importance of critical art from Indigenous artists in helping to counter hatred and oppression, the Board expects Meta to be particularly sensitive to the possibility of wrongful removal of the content in this case and similar content on Facebook and Instagram. It is not sufficient to evaluate the performance of Meta's enforcement of Facebook's Hate Speech policy as a whole. A system that performs well on average could potentially perform quite poorly on subcategories of content where incorrect decisions have a particularly pronounced impact on human rights. It is possible that the types of errors that occurred in this case are rare; the Board notes, however, that members of marginalized groups have raised concerns about the rate and impact of false positive removals for several years. The errors in this case show that it is incumbent on Meta to demonstrate that it has undertaken human rights due diligence to ensure its systems are operating fairly and are not exacerbating historical and ongoing oppression (UNGPs, Principle 17).

Meta routinely evaluates the accuracy of its enforcement systems in dealing with hate speech. This assessment is not broken down into assessments of accuracy that specifically measure Meta's ability to distinguish hate speech from permitted content that condemns hate speech or raises awareness.

Meta's existing processes also include *ad-hoc* mechanisms to identify error trends and investigate their root causes, but this requires large samples of content against which to measure system performance. The Board enquired whether Meta has specifically assessed the performance of its review systems in accurately evaluating counter speech that constitutes artistic expression and counter speech raising awareness of human rights violations. Meta told the Board that it had not undertaken specific research on the impact of false positive removals on artistic expression or on expression from people of Indigenous identity or origin.

Meta has informed the Board of obstacles to beginning such assessments, including the lack of a system to automate the collation of a sample of content that benefits from policy allowances. This was because reviewers mark content as violating or non-violating, and are not required to indicate where non-violating content engages a policy allowance. A sample of counter speech that fits within this allowance would need to be assembled manually.

While the Board was encouraged by the level of detail provided on how Meta evaluates performance during a Question and Answer session held at the Board's request, it is clear that more investment is needed in assessing the accuracy of enforcement of Hate Speech policy allowances and learning from error trends. Without additional information about Meta's design decisions and the performance of its human and automated systems, it is difficult for the Board or Meta to assess the proportionality of Meta's current approach to hate speech.

When assessing whether it is necessary and proportionate to use the specific machine learning tools at work in this case to automatically detect potential hate speech, understanding the accuracy of those tools is key. Machine learning classifiers always involve trade-offs between rates of false positives and false negatives. The more sensitive a classifier is, the more likely it is to correctly identify instances of hate speech, but it is also more likely to wrongly flag material that is not hate speech. Differently trained classifiers

and different models vary in their utility and effectiveness for different tasks. For any given model, different thresholds can be used that reflect a judgment about the relative importance of avoiding different types of mistakes. The likelihood and severity of mistakes should also inform decisions about how to deploy a classifier, including whether it can take action immediately or whether it requires human approval, and what safeguards are put into place.

Meta explained that the post in question in this case was sent for review by its automated systems because it was likely to have a large audience. This approach can limit the spread of harmful material, but it is also likely to increase the risk that powerful art that counters hate is wrongly removed. Meta told the Board that it regularly evaluates the rate of false positives over time, measured against a set of decisions by expert reviewers. Meta also noted that it was possible to assess the accuracy of the particular machine learning models that were relevant in this case and that it keeps information about its classifiers' predictions for at least 90 days. The Board requested information that would allow us to evaluate the performance of the classifier and the appropriateness of the thresholds that Meta used in this case. Meta informed the Board that it could not provide the information the Board sought because it did not have sufficient time to prepare it for us. However, Meta noted that it was considering the feasibility of providing this information in future cases.

Human review can provide two important safeguards on the operation of Meta's classifiers: first before the post was removed, and then again upon appeal. The errors in this case indicate that Meta's guidance to moderators assessing counter speech may be insufficient. There are any number of reasons that could have contributed to human moderators twice reaching the wrong decision in this case. The Board is concerned that reviewers may not have sufficient resources in terms of time or training to prevent the kind of mistake seen in this case, especially in respect of content permitted under policy allowances (including, for example, "condemning" hate speech and "raising awareness").

In this case, both reviewers were based in the Asia-Pacific region. Meta was not able to inform the Board whether reviewer accuracy rates differed for moderators assessing potential hate speech who are not located in the region the content originates from. The Board notes the complexity of assessing hate speech, and the difficulty of understanding local context and history, especially considering the volume of content that moderators review each day. It is conceivable that the moderators who assessed the content in this case had less experience with the oppression of Indigenous peoples in North America. Guidance should include clear instruction to evaluate content in its entirety and support moderators in more accurately assessing context to determine evidence of intent and meaning.

The Board recommended in its *Two Buttons Meme* decision (2021-005-FB-UA) that Meta let users indicate in their appeal that their content falls into one of the allowances to the Facebook Community Standard on Hate Speech. Currently, when a user appeals one of Meta's decisions that goes to human review, the reviewer is not informed that the user has contested a prior decision and does not know the outcome of the prior review. Whereas Meta has informed the Board that it believes this information will bias the review, the Board is interested in whether this information could increase the likelihood of more nuanced decision-making. This is a question that could be empirically tested by Meta; the results of those tests would be useful in evaluating the proportionality of the specific measures that Meta has chosen to adopt.

Under the UNGPs, Meta has a responsibility to perform human rights due diligence (Principle 17). This should include identifying any adverse impacts of content moderation on artistic expression and the political expression of Indigenous peoples countering discrimination. Meta should further identify how it will prevent, mitigate and account for its efforts to address those adverse impacts. The Board is committed to monitoring Meta's performance and expects to see the company prioritize risks to marginalized groups and show evidence for continual improvements.

9.Oversight Board decision

The Oversight Board overturns Meta's original decision to take down the content.

10. Policy advisory statement

Enforcement

1. Provide users with timely and accurate notice of any company action being taken on the content their appeal relates to. Where applicable, including in enforcement error cases like this one, the notice to the user should acknowledge that the action was a result of the Oversight Board's review process. Meta should share the user messaging sent when Board actions impact content decisions appealed by users, to demonstrate it has complied with this recommendation. These actions should be taken with respect to all cases that are corrected at the eligibility stage of the Board's process.

2. Study the impacts of modified approaches to secondary review on reviewer accuracy and throughput. In particular, the Board requests an evaluation of accuracy rates when content moderators are informed that they are engaged in secondary review, so they know the initial determination was contested. This experiment should ideally include an opportunity for users to provide relevant context that may help reviewers evaluate their content, in line with the Board's previous recommendations. Meta should share the results of these accuracy assessments with the Board and summarize the results in its quarterly Board transparency report to demonstrate it has complied with this recommendation.

3. Conduct accuracy assessments focused on Hate Speech policy allowances that cover artistic expression and expression about human rights violations (e.g., condemnation, awareness raising, self-referential use, empowering use). This assessment should also specifically investigate how the location of a reviewer impacts the ability of moderators to accurately assess hate speech and counter speech from the same or different regions.

The Board understands this analysis likely requires the development of appropriate and accurately labelled samples of relevant content. Meta should share the results of this assessment with the Board, including how these results will inform improvements to enforcement operations and policy development and whether it plans to run regular reviewer accuracy assessments on these allowances, and summarize the results in its quarterly Board transparency report to demonstrate it has complied with this recommendation.

*Procedural note:

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

For this case decision, independent research was commissioned on behalf of the Board. An independent research institute headquartered at the University of Gothenburg and drawing on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world and Duco Advisers, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology, provided expertise on socio-political and cultural context.

- Date published: Dec 9, 2021
- Platform: Facebook
- Relevant Community Standard: Hate speech
- Related topics: Art / Writing / Poetry, Culture, Marginalized communities
- Region: United States & Canada
- Country(s) affected: United States, Canada
- Attachments: Public Comments 2021-012-FB-UA