OVERTURNED

2022-004-FB-UA

Colombian police cartoon

The Oversight Board has overturned Meta's original decision to remove a Facebook post of a cartoon depicting police violence in Colombia.

Policies and topics

Freedom of expression, Governments, Mistreatment

Dangerous individuals and organizations

Region and countries

Latin America and the Caribbean

Colombia

Platform

Facebook

Attachments

Colombian police cartoon public comments

Case summary

The Oversight Board has overturned Meta's original decision to remove a Facebook post of a cartoon depicting police violence in Colombia. The Board is concerned that Media Matching Service banks, which can automatically remove images that violate Meta's rules, can amplify the impact of incorrect decisions to bank content. In response, Meta must urgently improve its procedures to quickly remove non-violating content from these banks.

About the case

In September 2020, a Facebook user in Colombia posted a cartoon resembling the official crest of the National Police of Colombia, depicting three figures in police uniform holding batons over their heads. They appear to be kicking and beating another figure who is lying on the ground with blood beneath their head. The text of the crest reads, in Spanish, "República de Colombia - Policía Nacional - Bolillo y Pata." Meta translated the text as "National Police – Republic of Colombia – Baton and Kick."

According to Meta, in January 2022, 16 months after the user posted the content, the company removed the content as it matched with an image in a Media Matching Service bank. These banks can automatically identify and remove images which have been identified by human reviewers as violating the company's rules. As a result of the Board selecting this case, Meta determined that the post did not violate its rules and restored it. The company also restored other pieces of content featuring this cartoon which had been incorrectly removed by its Media Matching Service banks.

Key findings

As Meta has now recognized, this post did not violate its policies. Meta was wrong to add this cartoon to its Media Matching Service bank, which led to a mass and disproportionate removal of the image from the platform, including the content posted by the user in this case. Despite 215 users appealing these removals, and 98% of those appeals being successful, Meta still did not remove the cartoon from this bank until the case reached the Board.

This case shows how, by using automated systems to remove content, Media Matching

Service banks can amplify the impact of incorrect decisions by individual human reviewers. The stakes of mistaken additions to such banks are especially high when, as in this case, the content consists of political speech criticizing state actors.

In response, Meta should develop mechanisms to quickly remove any non-violating content which is incorrectly added to its Media Matching Service banks. When decisions to remove content included in these banks are frequently overturned on appeal, this should immediately trigger reviews which can remove this content from the bank.

The Board is particularly concerned that Meta does not measure the accuracy of Media Matching Service banks for specific content policies. Without this data, which is crucial for improving how these banks work, the company cannot tell whether this technology works more effectively for some Community Standards than others.

The Oversight Board's decision

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

The Board recommends that Meta:

- Ensure that content with high rates of appeal and high rates of successful appeal is reassessed for possible removal from its Media Matching Service banks.
- Limit the time between when banked content is identified for additional review and when, if deemed non-violating, it is removed from the bank. This would ensure that content which is not violating is quickly removed from Media Matching Service banks.
- Publish error rates for content mistakenly included in Media Matching Service banks of violating content, broken down by content policy, in its transparency reporting.

*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 Facebook post of a cartoon depicting police violence in Colombia. The Board finds that the cartoon, which did not violate any Facebook Community Standard, was wrongly entered into one of Meta's Media Matching Service banks leading to the incorrect removal of the post. These banks can automatically identify and remove images which have been previously identified as violating the company's rules. After the Board selected the case, Meta found this content to be non-violating and acknowledged that it was removed in error. The Board is concerned that the wrongful inclusion of non-violating content into the Media Matching Service banks results in disproportionate wrongful enforcement. Moreover, the failure to prevent or remedy this error quickly compounds this problem over time. Hence, Meta must urgently improve its review mechanisms to remove non-violating content from these banks quickly, monitor the performance of these mechanisms and publicly release information as a part of its transparency efforts.

2. Case description and background

In September 2020, a Facebook user in Colombia posted a picture of a cartoon as a comment on another user's post. The cartoon resembles the official crest of the National Police of Colombia and depicts three figures wearing police uniforms and holding batons over their heads. The figures appear to be kicking and beating another figure who is lying on the ground with blood beneath their head. A book and a pencil are shown next to the figure on the ground. The text on the crest reads in Spanish, "República de Colombia - Policía Nacional - Bolillo y Pata." Meta's regional markets team translated the text as "National Police – Republic of Colombia – Baton and Kick." The post was made during a time of widespread protest in the country following a police killing.

According to Meta, in January 2022, 16 months after the content was originally posted to Facebook, the company removed the content as it matched with an image in a Media Matching Service bank of content that violates Facebook's Dangerous Individuals and Organizations Community Standard. The user appealed, and Meta maintained its decision to remove the content, but based its removal decision on the Violence and Incitement Community Standard instead. At the time of removal, the content had been viewed three times and received no reactions or user reports.

As a result of the Board selecting this case, Meta reviewed the content again and determined it did not violate the Dangerous Individuals and Organizations Community Standard or the

Violence and Incitement Community Standard. The content was then restored to the platform about one month after it had been removed.

Meta also informed the Board that because the image was in a Media Matching Service bank, identical content, including this case, had been removed from the platform. 215 of those removals were appealed by users, and 210 of those appeals were successful, meaning that the reviewers in this set of cases decided the content was not violating. All remaining removals, as well as any corresponding strikes and feature limits, were also reversed by Meta after the Board selected the case.

3. Oversight Board 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 overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4).

When the Board selects cases like this one, where Meta, subsequently acknowledges that it made an error, the Board reviews the original decision to increase understanding of the content moderation process which led to the error and to make structural recommendations to reduce errors and treat users more fairly in the future.

4. Sources of authority

The Oversight Board considered the following sources of authority:

I.Oversight Board decisions:

The Board's most relevant decision to this case is the "Colombia protests" decision (2021-010-FB-UA). In this decision the Board highlighted the public interest in allowing content criticizing the government during protests, in particular in contexts where states are accused of violating human rights.

II. Meta's content policies:

II.Meta´s content policies:

Facebook’s Dangerous Individuals and Organizations Community Standard states that Meta “remove[s] praise, substantive support and representation of various dangerous organizations.”

Facebook's Violence and Incitement Community Standard states that Meta "aim[s] to prevent potential offline harm that may be related to Facebook” and that it restricts expression “when [it] believe[s] there is a genuine risk of physical harm or direct threats to public safety.”

Facebook's Violent and Graphic Content Community Standard states that Meta “remove[s] content that glorifies violence or celebrates the suffering or humiliation of others.”

III.Meta’s values:

Meta's values are outlined in the introduction to the Facebook Community Standards, where the value of "Voice" is described as "paramount." Meta limits "Voice" in service of four values, namely “Authenticity,” “Safety,” “Privacy,” and “Dignity.” “Safety” and “Dignity” are the most relevant here.

IV.International human rights standards:

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

The Board's analysis of Meta’s human rights responsibilities in this case was informed by the following human rights standards which are applied in Section 8 of this decision:

- 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.

5. User submissions

In their statement to the Board, the user expressed confusion as to why the content was removed by Meta. The user explained that the content reflected reality in Colombia which was important for those who were interested in or affected by the situation.

6. Meta's submissions

Meta explained in its rationale that it removed the content because it matched with an image that had been mistakenly entered by a human reviewer into a Dangerous Individuals and Organizations Media Matching Service bank. On appeal, Meta upheld the removal but decided the content violated its Violence and Incitement policy rather than its Dangerous Individuals and Organizations policy. Meta later confirmed that both decisions to remove the content were wrong. The company stated that the Violent and Graphic Content policy could also be relevant to this case as it depicts a violent attack, but that it does not apply to images such as cartoons.

According to Meta, its Media Matching Service banks identify and act on media, in this case images, posted on its platforms. Once content is identified for banking, it is converted into a string of data, or "hash." The hash is then associated with a particular bank. Meta's Media Matching Service banks align with particular content policies such as Dangerous Individuals and Organizations, Hate Speech, and Child Sexual Exploitation, Abuse and Nudity, or specific sections within a policy. In this case, Meta stated the content was in a Media Matching Service bank specifically for criminal organizations, a prohibited category within the Dangerous Individuals and Organizations policy.

Depending on what Meta uses the specific bank for, it can be programmed to take different actions once it identifies banked content. For example, Meta might delete content that is violating, add a warning screen, or ignore it if it has been banked as non-violating content. Media Matching Service banks can also provide guidance to content reviewers when they are reviewing content that is banked.

Meta also stated that its Media Matching Service banks can act at different points in time. For example, Meta can scan images at the point of upload to prevent some violating content from being posted. Banks can also be configured to only detect and take action on newly uploaded content, or they can be used to scan existing content on the platform.

The Board asked Meta about how it identifies and adds content to Media Matching Service

The Board asked Meta about how it identifies and adds content to Media Matching Service banks. Meta explained that at-scale human reviewers identify content eligible for banking. According to Meta, for Media Matching Service banks associated with the Dangerous Individuals and Organizations policy, the content is then sent to a process called “Dynamic Multi-Review,” where “two reviewers must agree on one decision in order for the media to be sent to the bank.” Meta describes this as a guardrail to help prevent mistakes.

The Board also asked Meta about what mechanisms it has to identify erroneously banked content. Meta stated it has different “anomaly alerting systems.” These include a system that is triggered if a bank contains content that Meta considers to be viral. It also includes a system to identify banked content with a high level of successful appeals after removal. These systems were active for this bank when the content was removed.

Meta stated that the number of removals and successful appeals on this content generated an alert to Meta’s engineering team around the time the content was removed. However, a month later, when the Board brought this case to Meta’s attention, the company had still not reviewed and removed the content from the Media Matching Service bank. It is unclear whether this engineering team alerted other teams that would be responsible for re-reviewing and determining whether the content was violating. While Meta indicated that some time lag is expected when reviewing reports, it gave the Board no indication of when it would have addressed this alert. The Board is concerned about the length of this timeframe. Meta should be monitoring how long it takes to remove mistakenly added content from banks and reverse wrongful enforcement decisions after alerts are triggered and set a concrete goal to minimize that time.

Lastly, the Board asked Meta about the metrics it uses to audit the performance of its Media Matching Service banks. Meta states it generally monitors when its banks make more incorrect enforcement decisions and tries to identify what might be causing those increased errors. However, currently these audit assessments usually focus on individual banks. Because these banks can target specific policy lines and be programmed to take a variety of different actions, it can be difficult to generate meaningful data on overall enforcement accuracy for a specific community standard or for Media Matching Service banks in general from these analyses. Meta stated it is working to create a unified metric to monitor the accuracy of Media Matching Service banks.

The Board asked Meta a total of 26 questions, 24 of which were answered fully and two of which were answered partially. The partial responses were about measuring accuracy and error rates for the company's Media Matching Service banks and technical issues related to messages Meta sent to the user. Meta also provided a virtual briefing to a group of Board Members about Media Matching Service banks.

7. Public comments

The Board received four public comments related to this case. Two of the comments were submitted from the United States and Canada, and two from Latin America and the Caribbean. The submissions covered the use of media matching technology in content moderation, the socio-political context in Colombia, the importance of social media in recording police violence in handling protests, and how Meta's content policies should protect freedom of expression.

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

8. Oversight Board analysis

The Board selected this case as it involves artistic expression about police violence in the context of protests, a pressing issue across Latin America. The case also provided an opportunity for the Board to analyze Meta's use of Media Matching Service technology in content moderation. The Board looked at the question of whether this content should be restored through three lenses: Meta's content policies, the company's values, and its human rights responsibilities.

8.1 Compliance with Meta's content policies

I.Content rules

The Board finds that Meta's actions were not consistent with Facebook's content policies. As Meta acknowledges, the content did not violate any Meta policy. The decision to add it to a Media Matching Service bank and the failure to overturn the automated removal on appeal were wrong.

II.Enforcement action

Meta only restored the content, along with other pieces of content also removed because of incorrect Media Matching Service banking, after the Board selected the case. In response to a incorrect Media Matching Service banking, after the Board selected the case. In response to a question from the Board, Meta stated it had feedback mechanisms to identify errors and stop acting on mistakenly banked content. Nevertheless, 215 users appealed removals and 98% of those appeals were successful, and no feedback mechanism resulted in the content being removed from the Media Matching Service bank before the case reached the Board.

8.2 Compliance with Meta’s values

The Board finds that the original decision to remove this content was inconsistent with Meta's value of "Voice,” and that its removal was not supported by any other Meta values.

8.3 Compliance with Meta’s human rights responsibilities

The Board concludes that Meta’s initial decision to remove the content was inconsistent with its 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). Facebook’s Corporate Human Rights Policy states that this includes the International Covenant on Civil and Political Rights (ICCPR).

Freedom of expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, and expression about social or political concerns receives heightened protection (General Comment 34, paras. 11 and 20). 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 Board uses this framework to guide its analysis of Meta’s content moderation.

I.Legality (clarity and accessibility of the rules)

The requirement of legality provides that any restriction on freedom of expression is accessible and clear enough to provide guidance as to what is permitted and what is not. In this case, the incorrect removal of the content was not attributable to the lack of clarity or accessibility of relevant policies.

II.Legitimate aim

Any restriction on expression should pursue one of the legitimate aims listed in the ICCPR, which include the "rights of others." According to Meta, the Dangerous Individuals and Organizations and Violence and Incitement policies seek to prevent offline violence. The Board has consistently found that these aims comply with Meta's human rights responsibilities.

III.Necessity and proportionality

The principle of necessity and proportionality provides that any restrictions on freedom of expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; [and] they must be proportionate to the interest to be protected" (General Comment 34, para. 34).

In this case, removing the user's content was not necessary because it did not serve any legitimate aim. Additionally, the design of Meta's Media Matching Service banks enabled reviewers to mistakenly add content to a bank that resulted in the automatic removal of identical content, despite it being non-violating. The Board finds this was extremely disproportionate, based on the significant number of removals in this case, even considering Meta's scale of operation. Despite completing a lengthy data validation process to verify the number of content removals, Meta barred the Board from disclosing this number, citing a concern that the number was not an accurate reflection of the quality of Meta's Media Matching Service systems.

The Board finds the removal of the content in this case particularly concerning as the content did not violate any Meta policy but contained criticism of human rights violations which is protected speech. Police violence is an issue of major and pressing public concern. The Board has also previously noted the importance of social media in sharing information about protests in Colombia in case 2021-010-FB-UA.

The Board finds that adequate controls on the addition, auditing, and removal of content in such banks, as well as appeals opportunities, are essential. These banks greatly increase the scale of some enforcement decisions, resulting in disproportionate consequences for mistakes. Regarding the addition of content, given the consequences of removal that can be amplified to a disproportionate scale when automated, there is a need for Meta to strengthen procedures to ensure that non-violating content is not added. The stakes of mistaken additions to Media Matching Service banks are especially high when, as in this case, the content consists of political speech criticizing state actors or actions. Media Matching Service banks automate and amplify the impacts of individual incorrect decisions, and it is important for Meta to continually consider what error mitigation measures best help its human reviewers, including, for example, additional staffing, training, and time to review content.

Regarding auditing and feedback, the use of Media Matching Service banks to remove content with limited or flawed feedback mechanisms raises concerns of disproportionate erroneous enforcement, where one mistake is amplified to a much greater scale. Despite Meta informing the Board it had feedback mechanisms to identify and stop acting on mistakenly banked content, 215 users appealed the removal with a 98% success rate and the content remained banked. Meta should ensure that content acted on due to its inclusion in a Media Matching Service bank with high rates of overturn immediately trigger reviews with the potential to remove this content from the bank.

The Board is particularly concerned with the lack of performance metrics for the accuracy of Media Matching Service banks for particular content policies. Without objective metrics to monitor the accuracy of Media Matching Service banks, there is no effective governance over how this technology may be more or less effective for certain content policies. There are also no concrete benchmarks for improvement.

While the Board notes that Meta is in the process of creating a unified metric to monitor the accuracy of Media Matching Service banks, it urges the company to complete this process as soon as practicable. To enable the establishment of metrics for improvement, Meta should publish information on accuracy for each content policy where it uses Media Matching Service technology.

This should include data on error rates for non-violating content mistakenly added to Media Matching Service banks of violating content. It should also include the volume of content impacted by incorrect banking and key examples of errors. This data would allow Meta to understand the broader impacts of banking errors and set targets to reduce them. Further, it will allow users to better understand and respond to errors in automated enforcement.

9. Oversight Board decision

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

10. Policy advisory statement

Enforcement

1. To improve Meta's ability to remove non-violating content from banks programmed to identify or automatically remove violating content, Meta should ensure that content with high rates of appeal and high rates of successful appeal is re-assessed for possible removal from its Media Matching Service banks. The Board will consider this recommendation implemented when Meta: (i) discloses to the Board the rates of appeal and successful appeal that trigger a review of Media Matching Service-banked content, and (ii) confirms publicly that these reassessment mechanisms are active for all its banks that target violating content.

2. To ensure that inaccurately banked content is quickly removed from Meta's Media Matching Service banks, Meta should set and adhere to standards that limit the time between when banked content is identified for re-review and when, if deemed non-violating, it is removed from the bank. The Board will consider this recommendation implemented when Meta: (i) sets and discloses to the Board its goal time between when a re-review is triggered and when the non-violating content is restored, and (ii) provides the Board with data demonstrating its progress in meeting this goal over the next year.

Transparency

3. To enable the establishment of metrics for improvement, Meta should publish the error rates for content mistakenly included in Media Matching Service banks of violating content, broken down by each content policy, in its transparency reporting. This reporting should include information on how content enters the banks and the company's efforts to reduce errors in the process. The Board will consider this recommendation implemented when Meta includes this information in its Community Standards Enforcement Report.

*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. Linguistic expertise was provided by Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world.