OVERTURNED

2022-001-FB-UA

Knin cartoon

The Oversight Board has overturned Meta's original decision to leave a post on Facebook which depicted ethnic Serbs as rats

Policies and topics

Discrimination, Race and ethnicity, Freedom of expression

Hate speech

Region and countries

Europe

Croatia

Platform

Facebook

Attachments

Serbian translation

Croatian translation

Knin cartoon public comments

This decision is also available in Serbian and Croatian.

Да бисте прочитали ову одлуку на српском језику, кликните овде.

Da biste ovu odluku pročitali na hrvatskom, kliknite ovdje.

Case summary

The Oversight Board has overturned Meta's original decision to leave a post on Facebook which depicted ethnic Serbs as rats. While Meta eventually removed the post for violating its Hate Speech policy, about 40 moderators had previously decided that the content did not violate this policy. This suggests that moderators consistently interpreted the Hate Speech policy as requiring them to identify an explicit, rather than implicit, comparison between ethnic Serbs and rats before finding a violation.

About the case

In December 2021, a public Facebook page posted an edited version of Disney's cartoon "The Pied Piper," with a caption in Croatian which Meta translated as "The Player from Čavoglave and the rats from Knin."

The video portrays a city overrun by rats. While the entrance to the city in the original cartoon was labelled "Hamelin," the city in the edited video is labelled as the Croatian city of "Knin." The narrator describes how the rats decided they wanted to live in a "pure rat country," so they started harassing and persecuting the people living in the city.

The narrator continues that, when the rats took over the city, a piper from the Croatian village of Čavoglave appeared. After playing a melody on his "magic flute," the rats start to sing "their favorite song" and follow the piper out of the city. The song's lyrics commemorate Momčilo Dujić, a Serbian Orthodox priest who was a leader of Serbian resistance forces during World War II.

The piper herds the rats into a tractor, which then disappears. The narrator concludes that the rats "disappeared forever from these lands" and "everyone lived happily ever after."

The content in this case was viewed over 380,000 times. While users reported the content to Meta 397 times, the company did not remove the content. After the case was appealed to the Board, Meta conducted an additional human review, finding, again, that the content did not violate its policies.

In January 2022, when the Board identified the case for full review, Meta decided that, while the post did not violate the *letter* of its Hate Speech policy, it did violate the *spirit* of the policy, and removed the post from Facebook. Later, when drafting an explanation of its decision for the Board, Meta changed its mind again, concluding that the post violated the *letter* of the Hate Speech policy, and all previous reviews were in error.

While Meta informed the 397 users who reported the post of its initial decision that the content did not violate its policies, the company did not tell these users that it later reversed this decision.

Key findings

The Board finds that the content in this case violates Facebook's Hate Speech and Violence and Incitement Community Standards.

Meta's Hate Speech policy prohibits attacks against people based on protected characteristics, including ethnicity. The content in this case, which compares ethnic Serbs to rats and celebrates past acts of discriminatory treatment, is dehumanizing and hateful.

While the post does not mention ethnic Serbs by name, historical references in the content make clear that the rats being removed from the city represent this group. Replacing the name "Hamelin" with the Croatian city of "Knin," the identification of the piper with the Croatian village of Čavoglave (a reference to the anti-Serb song "Bojna Čavoglave" by the band 'Thompson' whose lead singer is from Čavoglave) and the image of rats fleeing on tractors are all references to Croatian military's "Operation Storm." This 1995 operation reportedly resulted in the displacement, execution, and forcible disappearance of ethnic Serb civilians. The comments on the post confirm that this connection was clear to people who viewed the content.

The Board is concerned that about 40 Croatian-speaking moderators deemed the content not to violate Facebook's Hate Speech Community Standard. This suggests that reviewers consistently interpreted the policy as requiring them to find an explicit comparison between ethnic Serbs and rats before finding a violation.

The Board also finds this content to violate Facebook's Violence and Incitement Community Standard. The Board disagrees with Meta's assessment that the content constitutes a call for expulsion without violence. By referring to the events of "Operation Storm," the post aims to remind people of past conflict and contains a violent threat. The cartoon celebrates the violent removal of Knin's ethnic Serb population and may contribute to a climate where people feel justified in attacking this group.

A serious question raised by this case is why Meta concluded that the content did not violate its policies, despite it being reviewed so many times. The fact that the content was not sent to Meta's specialized teams for assessment before it reached the Board shows that the company's

processes for escalating content are not sufficiently clear and effective. As such, the Board urges Meta to provide more information on how it escalates content.

The Oversight Board's decision

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

As a policy advisory opinion, the Oversight Board recommends that Meta:

- Clarify the Hate Speech Community Standard and the guidance provided to reviewers, explaining that even implicit references to protected groups are prohibited by the policy, when the reference would be reasonably understood.
- In line with Meta's commitment following the "Wampum Belt" case (2021-012-FB-UA), the Board recommends that Meta notify all users who have reported content when, on subsequent review, it changes its initial determination. Meta should also disclose the results of any experiments assessing the feasibility of introducing this change with the public.

*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 keep the content on Facebook. Following over 390 user reports to remove this content and Meta's additional review of the content when the Board selected the case, Meta found this content to be non-violating. However, when developing the explanation of its decision to the Board, Meta reversed its position and declared that this was an "enforcement error," removing the content for violating the Hate Speech policy. The Board finds that the content violates Meta's Hate Speech and Violence and Incitement Community Standards. It finds that the Hate Speech policy on comparing people to animals applies to content that targets groups through implicit references to protected characteristics. In this case, the content compared Serbs to rats. The Board also finds that removing the content is consistent with Meta's values and human rights responsibilities.

2. Case description and background

In early December 2021, a public Facebook page describing itself as a news portal for Croatia posted a video with a caption in Croatian. Meta translated the caption as "The Player from Čavoglave and the rats from Knin." The video was an edited version of Disney's cartoon "The Pied Piper." It was two minutes and 10 seconds long, with a voiceover in Croatian which was overlaid with the word "pretjerivač," referring to a Croatian online platform of the same name.

The video portrayed a city overrun by rats. While the entrance to the city in the original Disney cartoon was labelled "Hamelin," the city in the edited video was labelled as the Croatian city of "Knin." At the start of the video, a narrator described how rats and humans lived in the royal city of Knin for many years. The narrator continues that the rats decided that they wanted to live in a "pure rat country," so they started harassing and persecuting people living in the city. The narrator explains that when rats took over the city, a piper from the Croatian village of Čavoglave appeared. Initially, the rats did not take the piper seriously and continued with "the great rat aggression." However, after the piper started to play a melody with his "magic flute," the rats, captivated by the melody, started to sing "their favourite song" and followed the piper out of the city. Meta translated the lyrics of the song sung by the rats as: "What is that thing shining on Dinara, Dujić's cockade on his head [...] Freedom will rise from Dinara, it will be brought by Momčilo the warlord." The video then portrayed the city's people closing the gate behind the piper and the rats. The video ended with the piper herding the rats into a tractor, which then disappeared. The narrator concluded that once the piper lured all the rats into the "magical tractor," the rats "disappeared forever from these lands" and "everyone lived happily ever after."

The following factual historical background is relevant to the Board's decision. Croatia declared its independence from the Socialist Federal Republic of Yugoslavia on June 25, 1991. The remaining state of Yugoslavia (later called Federal Republic of Yugoslavia), which became predominantly of Serb ethnicity but contained many ethnic minorities, including Croats, used its armed forces in an attempt to prevent secession. The ensuing war, which lasted until 1995, resulted in extreme brutality on both sides, including forcible displacement of more than 200,000 ethnic Serbs from Croatia (Human Rights Watch Report, Croatia, August 1996). The Serb ethnic minority in Croatia, with the support of the Yugoslav National Army, opposed Croatian independence and (among other actions) established a state-like entity known as the Republic of Serbian Krajina (RSK). Knin became the capital of the RSK. During this period, many Croats were driven out of Knin. In 1995,

Croatian forces reoccupied Knin in a military operation called “Operation Storm.” This was the last major battle of the war. Because some Serbs fled on tractors, references to tractors can be used to humiliate and threaten Serbs.

Čavoglave is a village in Croatia near Knin, known as the birthplace of the lead vocalist and songwriter of the Thompson band. This Croatian band became known during the Croatian War of Independence for their anti-Serb song “Bojna Čavoglave,” which remains available online. The song, which was added to the video and has violent imagery, celebrates the retaking of Knin during Operation Storm. The piper who leads the rats out of Knin in the cartoon is identified as the “piper from Čavoglave.” The melody and lyrics that the cartoon rats sing also have specific meanings. The lyrics are from a song called “What is that thing that shines above Dinara,” which commemorates the Serbian past and Momčilo Dujić, a Serbian Orthodox priest who was a leader of Serbian resistance forces during World War II.

The page that shared the content has over 50,000 followers. The content was viewed over 380,000 times, shared over 540 times, received over 2,400 reactions, and had over 1,200 comments. The majority of the users who reacted to, commented on, or shared the content have accounts located in Croatia. Among the comments in Croatian were statements translated by Meta as: “Hate is a disease,” and “Are you thinking how much damage you are doing with this and similar stupidities to your Croatian people who live in Serbia?”

Users reported the content 397 times, but Meta did not remove the content. Of those who reported the post, 362 users reported it for hate speech. Several users appealed the leave-up decision to the Board. This decision is based on the appeal filed by one of these users, whose account appears to be located in Serbia. The user’s report was automatically rejected by an automatic system. This system resolves reports in cases that have already been examined and considered non-violating by Meta’s human reviewers a certain number of times, so that the same content is not re-reviewed. The content in this case was assessed as non-violating by several human reviewers before the automated decision was triggered. In other words, although the user report that generated this appeal was reviewed by an automatic system, previous reports of the same content had been reviewed by human reviewers, who decided that the content was not violating.

After the user who reported the content appealed Meta’s decision to take no action to the Board, an additional human review was conducted on the appeal level and again found that the content did not violate Meta policies. Meta further explained that in total, about 40 “human reviewer decisions (…) assessed the content as non-violating" and that “no human reviewer escalated the content.” Most of these human reviews took place on the appeal level. Meta added that all reviewers that reviewed the content are Croatian speakers.

When the Oversight Board included this case in its shortlist sent to Meta to confirm legal eligibility for full review, Meta did not change its assessment of the content, as it sometimes does at that stage. In late January 2022, when the Board designated the case for full review, Meta’s Content Policy team took another look. At that point, Meta determined that the Knin cartoon post did not violate the letter of the Hate Speech policy but violated the *spirit* of that policy and decided to remove it from Facebook. Meta explained that a “‘spirit of the policy’ decision is made when the policy rationale section of one the Community Standards makes clear that the policy is meant to address a given scenario that the language of the policy itself does not address directly. In those circumstances, we may nonetheless remove the content through a ‘spirit of the policy’ decision.” Later, when drafting its rationale for the Board, Meta changed its mind again, this time concluding that the post violated the *letter* of the Hate Speech policy, and that all previous reviews were in error.

According to Meta, the 397 users who reported the content were only informed about Meta’s initial determinations that the content did not violate Meta’s policies. They were not notified once Meta changed its decision and removed the content. Meta explained that due to “technical and resource limitations” it did not notify users when reported content is initially evaluated as non-violating and left up, and only later evaluated as violating and removed.

3. Oversight Board authority and scope

The Board has authority to review Meta’s decision following an appeal from the user who reported content that was then left up (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).

4. Sources of authority

The Oversight Board considered the following sources of authority:

I. Oversight Board decisions:

The most relevant prior Oversight Board decisions include:

- Case decision 2021-002-FB-UA ("Zwarte Piet"): In this case, a majority of the Board noted that "moderating content to address the cumulative harms of hate speech, even where the expression does not directly incite violence or discrimination, can be consistent with Facebook's human rights responsibilities in certain circumstances." The Board also found that "less severe interventions, such as labels, warning screens, or other measures to reduce dissemination, would not have provided adequate protection against the cumulative effects of leaving (...) content of this nature on the platform."
- Case decision 2020-003-FB-UA ("Armenians in Azerbaijan"): In the context of language which targeted a group based on national origin during conflict, the Board noted that "left up, an accumulation of such content may create an environment in which acts of discrimination and violence are more likely."
- Case decision 2021-011-FB-UA ("South Africa Slur"): The Board decided that it is in line with Meta's human rights responsibilities to prohibit "some discriminatory expression" even "absent any requirement that the expression incite violence or discriminatory acts".

II. Meta's content policies:

The policy rationale for Facebook's Hate Speech Community Standard states that hate speech is not allowed on the platform "because it creates an environment of intimidation and exclusion and, in some cases, may promote real-world violence." The Community Standard defines hate speech as a direct attack against people on the basis of protected characteristics, including race, ethnicity, and/or national origin. Meta prohibits content targeting a person or group of people based on protected characteristic(s) with "dehumanizing speech or imagery in the form of comparisons, generalizations or unqualified behavioral statements (in written or visual form) to or about: [a]nimals that are culturally perceived as intellectually or physically inferior." Meta also prohibits "[e]xclusion in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting, defined as [...] Explicit exclusion, which means things such as expelling certain groups or saying they are not allowed."

The policy rationale for 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." Specifically, Meta prohibits "coded statements where the method of violence or harm is not clearly articulated, but the threat is veiled or implicit," including where the content contains "references to historical [...] incidents of violence."

III. Meta's values:

Meta's values are outlined in the introduction to Facebook's 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 other 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.*

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 businesses' human rights responsibilities. 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; 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 and 5, International Convention on the Elimination of All Forms of Racial Discrimination (ICERD); General Recommendation 35, Committee on the Elimination of Racial Discrimination, CERD/C/GC/35, 2013;
- Responsibilities of businesses: Business, human rights, and conflict-affected regions: towards heightened action report, Report of the UN Working Group on the issue of human rights and transnational corporations and other business enterprises (A/75/212).

5. User submissions

The user who reported the content, and appealed to the Board in Croatian, states "[t]he Pied Piper symbolises the Croatian Army, which in 1995 conducted an expulsion of Croatia's Serbs, portrayed here as rats." According to this user, Meta did not assess the video correctly. They state that the content represents ethnic hate speech and that it "fosters ethnic and religious hatred in the Balkans." They also state that "this and many other Croatian portals have been stoking up ethnic intolerance between two peoples who have barely healed wounds of the war the video refers to."

When notified the Board had selected this case, the user who posted the content was invited to provide a statement. An administrator responded that they were a part of the page only as a business associate.

6. Meta's submissions

In the rationale Meta provided to the Board, Meta described its review process for this decision, but focused on explaining why its eventual removal of the content under the Hate Speech policy was justified. After repeated reports, multiple human reviewers found the content non-violating. Only after the Oversight Board selected the case did the company change its mind. Then, Meta determined that the content did not violate the *letter* of the hate speech policy, but that it made a "spirit of the policy" decision to remove the content. At this point, the Board informed Meta that it had selected the content for full review. Meta then changed its mind again, this time concluding that the content violated the *letter* of the policy. Specifically, it stated that it violated the policy line which prohibits content that targets members of a protected group and contains "[d]ehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about…[a]nimals that are culturally perceived as intellectually or physically inferior." In this revised determination, Meta stated that the content was violating as it contained a direct attack against Serbs in Knin by comparing them to rats.

Meta explained that its earlier determination that the content only violated the "spirit" of the Hate Speech policy was based on the assumption that the language of the policy did not prohibit attacks against groups on the basis of a protected characteristic identified implicitly. After additional review of this reasoning, Meta "concluded that it is more accurate to say that the policy language also prohibits attacks that implicitly identify" a protected characteristic.

Meta stated that its eventual removal was consistent with its values of "Dignity" and "Safety," when balanced against the value of "Voice." According to Meta, dehumanizing comparisons of people to animals that are culturally perceived as inferior may contribute to adverse and prejudicial treatment "in social integration, public policy, and other societally-impactful processes at institutional or cultural levels through implicit or explicit discrimination or explicit violence." Meta added that given the history of ethnic tensions and continuing discrimination against ethnic Serbs in Croatia, the video may contribute to a risk of real-world harm. In this regard, Meta referred to the Board's "Zwarte Piet" case decision.

Meta also stated that the removal was consistent with international human rights standards. According to Meta, its policy was "easily accessible" on Meta's Transparency Center website. Additionally, the decision to remove the content was legitimate to protect the rights of others from discrimination. Finally, Meta argued that its decision to remove the content was necessary and proportionate because the content "does not allow users to freely connect with others without feeling as if they are being attacked on the basis of who they are" and because of "no less intrusive means available for limiting this content other than removal."

The Board also asked Meta whether this content violated the Violence and Incitement Community Standard. Meta responded that it did not because "the content did not contain threats or statements of intent to commit violence" against ethnic Serbs and "exclusion or expulsion without violence does not constitute a violent threat." According to Meta, for the content to be removed under this policy "a more overt connection tying the rats in the video to the violent and forcible displacement" of Serbs would be necessary.

7. Public comments

The Oversight Board received two public comments related to this case. One of the comments was submitted from Asia Pacific and Oceania and one from Europe. The submissions covered the following themes: whether the content should stay on the platform, whether the comparison of Serbs to rats violates Meta's Hate Speech policy, and suggestions on how to enforce content rules on Facebook more effectively.

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: Meta's content policies, the company's values, and its human rights responsibilities.

8.1 Compliance with Meta's content policies

The Board finds that the content in this case violates the Hate Speech Community Standard. It also violates the Violence and Incitement Standard.

Hate Speech

Meta's Hate Speech policy prohibits attacks against people based on protected characteristics, including ethnicity. Here, the attacked group are ethnic Serbs living in Knin, specifically in Knin, targeted on the basis of their ethnicity. While the caption and video do not mention ethnic Serbs by name, the content of the video in its historic context, the replacement of the name "Hamelin" with "Knin," the lyrics used in the video, the identification of the piper with Čavoglave and therefore with the song by Thompson about Operation Storm, and the use of the tractor image are unmistakable references to Serb residents of Knin. Serbs are depicted as rats who must be removed from the city. The comments on the post and the many user reports confirm that this connection was abundantly clear to people who viewed the content.

The content contains two "attacks" within the definition of that term in the Hate Speech policy. First, the Hate Speech policy prohibits comparisons to "[a]nimals that are culturally perceived as intellectually or physically inferior." Meta's Internal Implementation Standards, which are guidelines provided to content reviewers, specify that comparisons to "vermin" are prohibited under this policy. The video contains a visual comparison of Serbs to rats. This constitutes a dehumanizing comparison in violation of the Hate Speech policy and the Internal Implementation Standards.

The Board finds that implied comparisons of the kind in this content are prohibited by Meta's hate speech policy.

Meta explained that previous decisions not to remove the content were based on the assumption that the letter of the policy did not apply to implicit references to protected characteristics. The Board disagrees with this assumption. The letter of the policy prohibits attacks based on protected characteristics no matter whether references to those characteristics are explicit or implicit. The Hate Speech Standard states that comparisons can take a written or visual form, such as video, and the language of the Standard does not require that references to targeted groups be explicit. While this reading of the policy is in line with both its text and rationale, the policy does not clearly formulate that implicit references are covered by the policy too.

Second, the content contains support for expelling Serbs from Knin. The rationale of the Hate Speech Standard defines attacks as "violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation." According to the policy line of the Hate Speech Community Standard applied in this case, explicit exclusion means supporting "things such as expelling certain groups or saying they are not allowed." The video in this case celebrates a historical incident where ethnic Serbs were forcibly expelled from Knin and the content states the townspeople were much better off after the rats were gone. This video contains support for ethnic cleansing in violation of the Hate Speech Standard.

Violence and Incitement

The Violence and Incitement policy prohibits "content that threatens others by referring to known historical incidents of violence." The caption and the video contain references to "Operation Storm," the 1995 military operation that reportedly resulted in displacement, execution, and

disappearance of ethnic Serb civilians. In the video, the city is named Knin and the rats flee on tractors, both references to Operation Storm. Comments to the post make clear these references are apparent to ethnic Serbs and Croatians.

The video may contribute to a climate where people feel justified in attacking ethnic Serbs. The post is designed to remind people of past conflict and to rekindle ethnic strife, with the goal of ridding the Knin area of the small remaining Serbian ethnic minority (on historical revisionism and radical nationalism in Croatia see Council of Europe Advisory Committee on the Framework Convention for the Protection of National Minorities' 2021 Fifth Opinion on Croatia, on online hate speech against ethnic Serbs see the 2018 European Commission against Racism and Intolerance report, para 30). When a threat is "veiled," according to the Facebook policy, it requires "additional context to enforce." That context is present in this post.

The Board disagrees with Meta's assessment that the content did not contain threats or statements of intent to commit violence, and that calls for exclusion or expulsion without specifying means of violence may not constitute a violent threat. The forced expulsion of people is an act of violence. The use of the Pied Piper story is not advocacy of peaceful removal but a clear reference to known historical incidents of violence, in particular with the imagery of the tractor.

As evidenced by the users who reported this post and the public comments, in the eyes of observers, rats in the cartoon represent the ethnic Serb population of the Knin area, including those who remained there. The cartoon clearly celebrates their removal. In the context of the Pied Piper story, the rats are induced to leave Knin by a magic flute rather than compelled by force, but the tractor reference refers to the actual forcible removal which is widely known about in the country. The tractor is a metaphor, but threats can be conveyed by metaphor no less effectively than by direct statements.

Meta's review process

The Board is particularly interested in why the company concluded this content was not violating so many times. It would have been helpful if Meta had focused on this at the outset, instead of focusing on why its revised decision to remove the post was correct. If the company wishes to reduce the level of violating content on its platform, it needs to treat the Board's selection of enforcement error cases as an opportunity to explore the reasons for its mistakes.

The Board notes the complexity of assessing cases such as this one and the difficulty of applying Facebook's Community Standards while accounting for context, especially considering the volume of content that human reviewers assess each day. Because of these challenges, the Board believes that it is important for Meta to improve its instructions to reviewers and pathways and processes for escalation. "Escalation" means for human reviewers to send a case to Meta's specialized teams, which then assess the content.

According to the rationale provided by Meta, to avoid subjectivity and achieve consistency in enforcement, human reviewers are instructed to apply the letter of the policy and not to evaluate intent. While objectivity and consistency are legitimate goals, the Board is concerned that the instructions provided to reviewers appear to have resulted in about 40 human reviewers erroneously qualifying the content as non-violating, and no reviewer reaching the decision which Meta ultimately believed to be the correct one, which is removal.

The possibility to escalate content is supposed to lead to better outcomes in difficult cases. Review on the escalation level may assess intent and is better equipped to account for context. This content was reported 397 times, had a wide reach, raised policy questions, required context to assess and involved content from a Croatian online platform which, according to experts consulted by the Board, was previously the subject of public and parliamentary discussion on the freedom of speech in the context of satire. Yet, no reviewer escalated the content. Meta told the Board it encourages reviewers to escalate "trending content," and to escalate when in doubt. Meta defined trending content as "anything that is repetitive in nature (…) combined with the type of action associated with the content (i.e. potential harm or community risk(…))." The fact that the content was not escalated prior to Board selection indicates that escalation pathways are not sufficiently clear and effective. The failure to escalate was a systemic breakdown.

One factor which may have prevented the escalation of content in this case is that Meta does not provide at scale reviewers with clear thresholds on when content is "trending." Another factor that may have contributed to the failure of reviewers to identify the content as "trending" – and thus to escalate – was the automated review system Meta used in this case. Meta explained that it uses automation to respond to reports when there are a certain amount of non-violation decisions over a given time period to avoid re-review.

The Board is concerned about Meta's escalation pathways, and notes that it should provide more information regarding these. It should study whether additional pathways to escalate content are necessary and whether the automated system used in this case prevents content which is

viral and often reported from being escalated.

The case also exposed flaws in Meta's reporting and appeal process. The Board is concerned about Meta not notifying users when the company changes its decision in a case. A user who reported content which is initially evaluated as non-violating and left up, and then later evaluated as violating and removed, should be updated on this.

8.2 Compliance with Meta's values

The Board finds that removing the content is consistent with Meta's values of "Voice," "Dignity," and "Safety."

The Board recognizes that "Voice" is Meta's paramount value, but the company allows for expression to be limited to prevent abuse and other forms of online and offline harm. Those targeted by dehumanizing and negative stereotypes may also see their "Voice" affected, as their use may have a silencing impact on those targeted and inhibit their participation on Facebook and Instagram. By allowing such posts to be shared, Meta may contribute to a discriminatory environment.

The Board considers the values of "Dignity" and "Safety" to be of superseding importance in this case. In this regard, the Board noted the continuing increase in cases of physical violence against ethnic Serbs in Croatia (2021 CoE Fifth Opinion on Croatia, para. 116). This justified displacing the user's "Voice" to protect the "Voice," "Dignity," and "Safety" of others.

8.3 Compliance with Meta's human rights responsibilities

The Board concludes that removing the post from the platform is consistent with 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).

Freedom of expression (Article 19 ICCPR)

The scope of the right to freedom of expression is broad. Article 19, para. 2, of the ICCPR gives heightened protection to expression on political issues and discussion of historical claims (General Comment No. 34, paras. 20 and 49). 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.

I. Legality (clarity and accessibility of the rules)

The principle of legality requires rules used by states to limit expression to be clear and accessible (General Comment 34, para. 25). The legality standard also requires that rules restricting expression "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and "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.). Individuals must have enough information to determine if and how their expression may be limited, so that they can adjust their behavior accordingly. Applied to Meta's content rules for Facebook, users should be able to understand what is allowed and what is prohibited, and reviewers should have clear guidance on how to apply these standards.

The Board finds that the Hate Speech Community Standard prohibits implicit targeting of groups on the basis of protected characteristics. This is the case for both dehumanizing comparisons to animals and for statements advocating or supporting exclusion. The errors that occurred in this case show that the language of the policy and the guidance provided to reviewers are not sufficiently clear.

In the case, about 40 human reviewers decided the content did not violate the Hate Speech Community Standard. Prior to the final determination by Meta, no human reviewer found the content to be violating. This indicates reviewers consistently interpreted the policy as requiring them to find an explicit comparison between ethnic Serbs and rats before finding a violation. The company first informed the Board that the spirit of the policy prohibited implied comparisons to animals, and later that the letter of the policy covered implied comparisons. The confusion throughout this process evidences a need for clearer policy and implementation guidance.

II. Legitimate aim

The Board has previously recognized that the Hate Speech Community Standard and the Violence and Incitement Standard pursue the legitimate aim of protecting the rights of others. Those rights include the rights to equality and non-discrimination (Article 2, para. 1, ICCPR, Article 2 and 5 ICERD) and exercise their freedom of expression on the platform without being harassed or threatened (Article 19 ICCPR).

III. Necessity and proportionality

For restrictions on expression to be considered necessary and proportionate, those restrictions "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected" (General Comment 34, para. 34). The Special Rapporteur on free expression has also noted that on social media, "the scale and complexity of addressing hateful expression presents long-term challenges" (A/HRC/38/35, para. 28). However, according to the Special Rapporteur, companies should "demonstrate the necessity and proportionality of any content actions (such as removals or account suspensions)." Moreover, companies are required "to assess the same kind of questions about protecting their users' right to freedom of expression" (ibid para. 41.).

The Facebook Hate Speech Community Standard prohibits specific forms of discriminatory expression, including comparison to animals and calls for exclusion, absent any requirement that the expression incite violence or discriminatory acts. The Board, drawing upon the UN Special Rapporteur's guidance, has previously explained that, while such prohibitions would raise concerns if imposed by a government at a broader level, particularly if enforced through criminal or civil sanctions, Facebook can regulate such expression, demonstrating the necessity and proportionality of the action (see the " South Africa Slur" decision).

The content in this case, comparing ethnic Serbs to rats and celebrating past acts of discriminatory treatment, is dehumanizing and hateful. The Board would have come to a similar conclusion about any content that targets an ethnic group in this way, especially in a region that has a recent history of ethnic conflict. The Board finds removing this content from the platform was necessary to address the serious harms hate speech on the basis of ethnicity poses.

The Board considered the factors in the Rabat Plan of Action (The Rabat Plan of Action, OHCHR, A/HRC/22/17/Add.4, 2013) to guide its analysis, while accounting for differences between international law obligations of states and human rights responsibilities of businesses . Meta has a responsibility to "seek to prevent or mitigate adverse human rights impacts that are directly linked to [its] operations, products or services" (UNGPs, Principle 13). In its analysis, the Board focused on the social and political context, intent, the content and form of the speech and the extent of its dissemination.

Regarding the context, this relates to a region that has recently experienced ethnic conflict and the backdrop of online hate speech and incidents of discrimination against ethnic minorities in Croatia (see Section 8.1. under Violence and Incitement). It intends to incite ethnic hatred, and this may contribute to individuals taking discriminatory action. The form of the expression and its wide reach is also important. The video was shared by an administrator of a Page which, according to expert briefings the Board received, is a Croatian news portal known for anti-Serb sentiments. The cartoon video form can be particularly harmful because it is especially engaging. Its reach was broad. While the video was created by someone else, it is likely that the popularity of the page (which has over 50,000 followers) would increase the reach of the video, especially as it reflects the views of the page and its followers. The content was viewed over 380,000 times, shared over 540 times, received over 2,400 reactions and had over 1,200 comments.

In the "South Africa Slur" decision, the Board decided that it is in line with Meta's human rights responsibilities to prohibit "some discriminatory expression" even "absent any requirement that the expression incite violence or discriminatory acts." The Board notes that Article 20, para. 2, ICCPR, as interpreted in the Rabat Plan of Action, requires imminent harm to justify restrictions on expression. The Board does not believe that this post would result in imminent harm. However, Meta can legitimately remove posts from Facebook that encourage violence in a less immediate way. This is justified, as the human rights responsibilities of Meta as a company differ from the human rights obligations of states. Meta can apply less strict standards for removing content from its platform than those which apply to states imposing criminal or civil penalties.

In this case, depicting the Serbs as rats and calling for their exclusion while referencing historical acts of violence, impacts the rights to equality and non-discrimination of those targeted. This justifies removing the post. Many Board Members also believed that the content had a negative impact of the freedom of expression of others on the platform, as it contributed to an environment where some users would feel threatened.

The Board finds that removing the content from the platform is a necessary and proportionate measure. Less invasive interventions, such as labels, warning screens, or other measures to reduce dissemination, would not have provided adequate protection against the cumulative effects of leaving content of this nature on the platform (for a similar analysis see the "Depiction of Zwarte Piet" case).

9. Oversight Board decision

The Oversight Board overturns Meta's original decision to leave up the content, requiring the post to be removed.

10. Policy advisory statement

Content policy

1. Meta should clarify the Hate Speech Community Standard and the guidance provided to reviewers, explaining that even implicit references to protected groups are prohibited by the policy when the reference would reasonably be understood. The Board will consider this recommendation implemented when Meta updates its Community Standards and Internal Implementation Standards to content reviewers to incorporate this revision.

Enforcement

2. In line with Meta's commitment following the "Wampum belt" case (2021-012-FB-UA), the Board recommends that Meta notify all users who have reported content when, on subsequent review, it changes its initial determination. Meta should also disclose the results of any experiments assessing the feasibility of introducing this change with the public. The Board will consider this recommendation implemented when Meta shares information regarding relevant experiments and, ultimately, the updated notification with the Board and confirms it is in use in all languages.

***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. The Board was also assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. The company Lionbridge Technologies, LLC, whose specialists are fluent in more than 350 languages and work from 5,000 cities across the world, provided linguistic expertise.