Emojis Targeting Black People

2026-001-FB-UA, 2026-002-IG-UA

Summary

The Oversight Board overturns Meta's original decisions to keep up two pieces of content that use emojis to express hate, discrimination and harassment towards Black people by comparing them to monkeys. The Board has called for Meta to prevent hateful and discriminatory targeting of groups by improving its automated and human moderation to comprehensively account for "algospeak," including emojis. This should encompass ensuring its training data for automated policy enforcement is regionally appropriate and up to date, that efforts are coordinated to proactively disrupt hateful campaigns and ensure its mitigation efforts include active monitoring of emoji content inciting discrimination and hostility during major sporting events, such as the FIFA (International Federation of Association Football) World Cup.

About the Cases

These cases address two posts made in May 2025 using monkey emojis to refer to Black people.

In the first case, a user in Brazil posted a short video on Facebook featuring a scene from the movie The Hangover, in which two characters argue, dubbed in Portuguese, claiming ownership of a monkey. Text overlaying the video names the characters as the Spanish football (soccer) clubs "Barcelona" and "Real Madrid." Additional overlay text refers to boys rising to prominence in Brazilian football. The caption consists of a monkey emoji. The post was viewed over 22,000 times and 12 people reported it.

The second case involves a comment posted in response to a video on an Instagram account in Ireland. In the video, the user expresses indignation after witnessing a racist incident on the street and the caption calls to reject racism in Ireland. Another user's comment says they do not support the message, rather they want the situation to "blow up" and "to have some glorious fun with all the [monkey emojis] & out in the street." The comment additionally included several monkey, laughing and praying emojis, and underscored "glorious days ahead." The original post was viewed over 4,000 times and 62 people reported the comment.

Meta's automated systems and – after user appeals – human reviewers left both posts up. Users then appealed to the Board. After the Board selected these cases for review, Meta determined its initial decisions were wrong and removed the posts in July 2025 for violating the company's Hateful Conduct Community Standard.

Coded language through turns of phrases or emojis (called "algospeak") can be used to convey dehumanizing or hateful messages while bypassing automated content moderation systems.

Key Findings

The Board is concerned about the accuracy of the enforcement of the Hateful Conduct policy, especially in assessing emojis used as algospeak. Classifiers identified the content but took no action. Meta says reviewers should consider all aspects of the content, such as imagery, captions and text overlays, and factors beyond the immediate content, including the main post and comments. Meta also explained that its classifiers are trained on datasets of reported and labeled examples, including cases where emojis are used in potentially violating ways. However, automated and human reviews failed to accurately assess the posts.

Meta should improve automated detection of violative emoji use by periodically auditing its training data. Enforcement processes should always direct content to reviewers with appropriate language and regional expertise.

Responding to the Board's questions, Meta stated that after the company's January 7, 2025, announcement, large language models (LLMs) are now more widely integrated as an additional review layer, including for content that may violate the Hateful Conduct policy. According to Meta, the LLMs do not replace existing models, but provide a second opinion on enforcement decisions, focusing on content that has been flagged for removal. In these cases, LLMs were not involved in the review process.

The Board finds that both posts violate the Hateful Conduct Community Standard prohibiting dehumanizing comparisons to animals. Both posts utilize the monkey emoji to target Black people on the basis of their protected characteristic.

Keeping the posts up is also inconsistent with Meta's human rights responsibilities, as emojis seeking to dehumanize and incite discrimination or hostility towards protected characteristic groups should be subject to removal. It is necessary and proportionate to remove both posts.

Both posts represent forms of algospeak used to express hate, discrimination and harassment towards specific protected characteristic groups, and illustrate how emojis can be used to urge others to take discriminatory and potentially hostile action.

The Brazilian post was made in the context of widely documented systemic racism and hostility in football, particularly targeting Black players. The comment in the Irish case was shared in the context of rising racial discrimination and Afrophobia in Ireland.

To better coordinate its efforts and protect people who may not be directly named but are implicit targets of hateful campaigns, Meta should develop a framework to

harmonize its already-existing measures to proactively disrupt hateful campaigns, especially those involving the use of emojis. Meta should ensure that its time-sensitive mitigation efforts, be that through its Integrity Product Operations Center or another risk mitigation system, include active monitoring of content with emojis that incite targeted discrimination or hostility in the lead up, during and in the immediate aftermath of major sporting events, e.g., the 2026 FIFA World Cup.

The Oversight Board's Decision

The Board overturns Meta's original decision to keep up both pieces of content.

The Board also recommends that Meta:

- Audit its training data for automated systems used for Hateful Conduct policy enforcement and ensure the data is updated periodically to include examples of content with emojis in all languages, violating use of emojis and new instances of the hateful use of emojis.
- Harmonize its existing efforts to proactively disrupt hateful campaigns, especially those involving the use of emojis.to better protect people who are not directly named but the implicit targets of hateful campaigns.
- Ensure that its time-sensitive mitigation efforts, be that through its Integrity Product Operations Center or another risk mitigation system, include active monitoring of content with emojis that incite targeted discrimination or hostility in the lead up, during and in the immediate aftermath of major sporting events, such as the FIFA World Cup.

The Board reiterates the importance of its relevant previous recommendation that Meta:

- Provide users with an opportunity for self-remediation comparable to the post time friction intervention that was created as a result of the Pro-Navalny Protests in Russia recommendation no. 6. If this intervention is no longer in effect, the company should provide a comparable product intervention.

*Case summaries provide an overview of cases and do not have precedential value.

Full Case Decision

1. Case Description and Background

This decision addresses two cases involving the use of monkey emojis to refer to Black people.

The first case involves a short video posted on Facebook in May 2025 by a user in Brazil. The video features a scene from the movie The Hangover, in which two characters argue, dubbed in Portuguese, and claim ownership of a monkey. Text overlaying the video names the characters “Barcelona” and “Real Madrid,” which are Spanish football (soccer) clubs. During the argument, the character labelled “Real Madrid” briefly threatens the one labelled “Barcelona” with a gun. Additional overlay text refers to boys rising to prominence in Brazilian football and the video’s caption only contains a monkey emoji. The post was viewed over 22,000 times, and 12 people reported the content.

The second case involves a comment made in May 2025 in response to a video posted by a user in Ireland to their Instagram account. In the video, the posting user is on camera expressing indignation after witnessing a racist incident on the street in Ireland, in which a group of teenagers shouted a racist slur at a Black woman. In the caption, the

posting user expresses heartbreak for the victim and emphasizes that being both Black and Irish is possible, but that this is a conversation that "white privileged people" do not want to have. The user also urges others to speak up and to have hard conversations and encourages society to do better in combating racism. The caption ends with a hashtag call to reject racism in Ireland.

Responding to the video in the Irish case, another user commented that they do not support the message. Rather, they challenged the video's creator, asking them what they intend to do about the situation. The commenting user also expressed eagerness for the situation to "blow up" and "to have some glorious fun with all the [monkey emojis] & out in the street." The comment included several additional monkey, laughing and praying emojis and underscored "glorious days ahead." The parent post was viewed over 4,000 times and 62 people reported the comment.

Meta's automated systems detected both posts as potentially violating the Hateful Conduct policy. The classifier found the content in the Brazilian case to be non-violating, while it was unable to confidently determine the language in the Irish case (English) and, consequently, took no action on it.

Users reported the two posts for Hateful Conduct, and they were sent for review. However, they were not prioritized for human review and remained on the platforms.

In both cases, the users who reported the posts appealed Meta's decision to leave the posts up. Following human review, Meta upheld its initial decisions on appeal, after which the users appealed to the Board.

As a result of the Board selecting these cases for review, Meta determined its initial decisions were wrong and removed the posts in July 2025 for violating the company's Hateful Conduct Community Standard.

The Board notes the following context in reaching its decisions:

Racism faced by Brazilian, specifically Black, football players has attracted significant media coverage, spotlighting the wider problem of racism against players and among football fans. For example, Vinícius Júnior, a player for Real Madrid, has experienced multiple racist incidents, including comparisons to monkeys. The racism has escalated to a point where the Brazilian state of Rio de Janeiro passed the "Vinícius Júnior Law" in 2023 to combat racism during sporting events. The law mandates an interruption or even a termination of a sporting event when a racist act takes place.

Several other high-profile incidents against football players of African descent, primarily involving abuse from fans in stadiums and on social media, have been reported in European countries (for instance, in Spain, Italy, France, England). The Euro 2020 competition drew public attention to the issue as players faced an onslaught of online hate. CONMEBOL, the continental governing body of football in South America, established a taskforce in March 2025 on eradicating racism, discrimination and violence in football. This was partially in response to racism faced by Brazilian players, including monkey chants at games.

Studies have documented sustained waves of abuse following matches and news spikes. The trade union for professional footballers in England and Wales, the Professional Footballers' Association (PFA), and data science company Signify found that in 2020 more than 3,000 of the tweets sent to some players were explicitly abusive messages, of which 56% were racist. 29% of those racially abusive posts appeared in the form of emojis. The same study found that 43% of players in the Premier League, the highest level of the English football league system, received explicitly racist abuse. Similarly, recent monitoring by the Union of European Football Associations (UEFA), the governing body for football in Europe, found that 33% of posts flagged to Meta, TikTok and X, for abusive content shared around UEFA club finals were classified as racist.

As part of its research in these cases, the Board searched Meta's Content Library for content using the monkey emoji. An analysis of the top 150 most engaged public posts on both Facebook and Instagram between October 1, 2024, and October 1, 2025, shows that the emoji is most often used to accompany videos of monkeys, memes or lighthearted prank content intended to go viral. However, the research showed that when the emoji appeared in discussions about Black people, it carried various connotations in different contexts. It is deployed in a dehumanizing manner to compare Black people to animals. At other times, it is used to highlight or comment on racism encountered. And it is employed by users that appear to be from the Black community in a self-referential manner or with humor, though the Board is unable to verify the identity of these users. Many of the posts relevant to football used the monkey emoji in a dehumanizing manner, often targeting particular football players, while others, to a lesser extent, discussed or condemned the racism football players experience.

The Irish Human Rights and Equality Commission report for the 6th monitoring cycle of the European Commission against Racism and Intolerance highlighted a rise in racism, discrimination and intolerance in Ireland, noting that "the growth of far-right ideology, leading to events like the Dublin riots, has been facilitated by systemic gaps in the protection against racism and intolerance in Ireland." Ireland's shift from a prominently white nation to a more diverse, immigration-receiving country in the 1990s has led to a shift in sociopolitical relations. This includes an increase in "Afrophobia," the racism against people of African descent, throughout Ireland.

The European Union's Fundamental Rights Agency reported that almost half of people of African descent surveyed about their experiences of living in 13 EU Member States, including Ireland, "experienced racial discrimination, an increase from 39% in 2016 to 45% in 2022." Over 44% of the respondents in Ireland experienced racial harassment and 34% expressed worries about becoming a victim of a physical attack because of their ethnic or immigrant background. The Organization for Security and Cooperation in Europe's (OSCE) Office for Democratic Institutions and Human Rights' (ODIHR) 2024

Hate Crime report notes that 587 out of 676 hate crimes recorded by the Irish police were classified as motivated by racist and xenophobic bias.

Coded language through turns of phrase or emojis (also referred to as "algospeak") is used in order to bypass automated content moderation systems. Research suggests that the use of emojis "presents a common jailbreaking way that users exploit, either deliberately or unintentionally, to convey offensive meaning through stereotypical associations." Researchers further explain that although "language models have a comprehensive grasp on textual constructions, there is still a need for the models to be taught what emoji mean in various contexts, and how different emoji condition the likelihood of hatefulness in a given tweet, post or comment."

2. User Submissions

In statements to the Board, the reporting users explained that the posts contained racist language by comparing Black people to monkeys. The reporting user in the Irish case emphasized that using emojis in place of words was clearly racist and the post being on the platform highlighted flaws in Instagram's automated detection systems.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

The Hateful Conduct Community Standard prohibits content "targeting a person or group of people … on the basis of their … protected characteristic(s)," including race, ethnicity, and national origin "in written or visual form." This encompasses "dehumanizing speech in the form of comparisons to or generalizations about … animals in general or specific types of animals that are culturally perceived as inferior (including but not limited to: Black people and apes or ape-like creatures)." These

comparisons can be shown visually through the use of emojis and discerned by the content's context.

In the introduction to its Community Standards, Meta states that it may remove content that uses "ambiguous or implicit language" when additional context allows it to reasonably understand that the content goes against the Community Standards.

II. Meta's Submissions

As a result of the Board selecting these cases, Meta reversed its original decisions to keep up both posts and removed them. Following a review by the company's subject matter experts, Meta decided that both posts constituted dehumanizing speech that compares individuals to animals based on their protected characteristics, prohibited under the Hateful Conduct policy.

In the Brazilian case, Meta explained that the content appears to repurpose a movie scene to comment on European football clubs' recruitment practices, suggesting that Real Madrid and Barcelona compete over Brazilian players (who are often Black) in the same manner as the men in the scene argue over the ownership of the monkey. Given recent racial incidents in which rival supporters compared some of these recruits to monkeys, the company determined that the monkey emoji was being used to compare Black Brazilians to monkeys.

In the Irish case, Meta determined the commenting user appears to equate Black people to monkeys in sharing the monkey emoji to refer to them. According to Meta, given the video's description of a Black woman being harassed in the street, the commenter's use of the monkey emoji mirrors that behavior, comparing Black people to monkeys and expressing an intention to harass them as the youths referred to in the video did.

Meta told the Board that its at-scale reviewers allowed both posts to remain on the platforms due to a combination of contextual ambiguity, incomplete review, misapplication of policy guidance to visual indicators such as emojis, and language and tooling limitations. The company informed the Board that coaching and feedback have been provided to reviewers. Further investigation also revealed tooling issue that prevented proper translation and case routing, resulting in the Brazilian content getting assigned for initial assessment to a reviewer who did not possess Portuguese language expertise. According to Meta, this routing issue has since been resolved and all reviews should be routed to queues with relevant language and regional expertise.

Meta highlighted that its decision to remove both pieces of content seeks to protect the rights of others to be free from discrimination and it also aligns with Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), which prohibits "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." The company also considered that removal was necessary to preserve an environment where users are free from discrimination or hostility.

In response to the Board's questions, Meta stated that emojis can sometimes influence the meaning of content, but their interpretation varies widely among users and across regions and may have different meanings in different posts. Therefore, the company does not consider emojis inherently violating. Instead, Meta provides guidance and examples in the internal guidelines to human reviewers as to possible meaning of commonly used emojis such as the monkey emoji and asks reviewers to always consider the full context in which the emoji appears to determine whether its use may be violating. This involves looking at all aspects of the content, such as imagery, captions and text overlays, as well as factors beyond the immediate content, including the main post and related comments.

Similarly, Meta stated that its automated systems are designed to identify uses of emojis that may violate policies by considering the context in which they appear. To that end, according to Meta, the systems analyze the entire post, capturing all its elements and key metadata, which "allows the models to detect patterns and signals that may indicate policy violations, even when emojis are used to substitute for words or concepts." This includes consideration of previous violating patterns. Meta noted, however, that its classifiers prioritize the removal of only *explicit* Hateful Conduct violations "to minimize the risk of wrongful takedowns and overenforcement," which is particularly relevant to emojis, that can have highly context-dependent meanings.

Meta explained that following the company's announcement on January 7, 2025, it is changing its automated policy violations systems approach. Prior to that date, the system utilized proactive enforcement, entailing the automated detection and automated removal of all violating content. While Meta's automated systems can still detect potential Hateful Conduct violations, removals of such violations are now based on user reports and escalations from Trusted Partners, rather than being based solely on automated detection. The company noted that this approach applies globally, but it "may continue proactive enforcement in countries experiencing crises." Meta added that the company is continually assessing its legal obligations worldwide to determine whether these proactive efforts align with local law. Additionally, Meta explained that the company takes "a tailored approach" to the use of their automated tools, adapting their response to the specific issue at hand. For example, Meta retains flexibility to deploy its automated tools to address high-risk trends identified on the platform, as well as for targeted purposes, such as voter suppression. In certain countries experiencing crises (which can include countries designated under the Crisis Policy Protocol) or prolonged instability, where significant integrity and regulatory risk necessitate a robust monitoring system, the company may proactively remove violating content based on the context on the ground.

In response to the Board's questions, Meta stated that in the aftermath of the company's announcement on January 7, 2025, large language models (LLMs) are now more widely integrated as an additional review layer, including for content that may violate the Hateful Conduct policy. According to Meta, the LLMs do not replace existing models, but provide a second opinion on enforcement decisions, focusing on content that has been flagged for removal. In the present cases, LLMs were not involved in the review process.

Meta also stated it has several systems in place to reduce bias in the review process. For human reviewers, this includes weekly audits across all review teams, which allows Meta to understand where mistakes are being made so they can be addressed. Additionally, according to Meta, reviewers are regularly re-trained in the policies. The company said it also holds bi-weekly sessions for reviewers to seek clarification on policy details, to ensure standards are applied correctly and consistently. Meta noted that for the automated systems, the company trains its classifiers on human-reviewed reports and selects a broad range of content to ensure classifiers are learning from the most severe cases as well as content that may otherwise be overlooked.

In response to the Board's questions, Meta explained that Instagram users may be notified that their comment or post has been flagged, offering them the option to delete the content. According to Meta, as its data retention policies do not extend past 30 days, it cannot be confirmed if the user in the Irish case was notified about their comment.

Meta defines "directed mass harassment" as "attempts on- or off- platform to mobilize a large group of people to target a specific subject." Meta removes such content when it is targeting via any surface, i.e., any place on the platform, "individuals at heightened risk of offline harm," such as human rights defenders or minors. Similarly, Meta also removes such content when it is targeting any individual via their personal profiles or inbox with (1) content that violates the Bullying and Harassment policies for private individuals or (2) objectionable content that is based on a protected characteristic.

Meta noted that it does not allow Hateful Conduct content regardless of whether the content targets a public or private individual.

The Board asked questions on policy and enforcement considerations for content involving emojis; enforcement history of the posts; updates to enforcement of the Hateful Conduct policy following the January 7, 2025, announcement; and details on the enforcement mechanisms currently in place. Meta responded to all the questions.

4. Public Comments

The Board received nine public comments that met the terms for submission. Eight of the comments were submitted from the United States and Canada, and one from the Middle East and North Africa. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: the evolving use of emojis in algospeak; challenges with automated detection of coded speech as well as enforcing against content with emojis that may have multiple meanings; importance of contextual assessments and moderator trainings to detect enforcement circumvention through algospeak; and insights on racism in sports.

In October 2025, as part of ongoing stakeholder engagement, the Board consulted with representatives of advocacy organizations, academics, inter-governmental organizations and other experts on the issues of moderating content with emojis. The participants underlined the challenges of contextual analysis of content with emojis that have evolving and wide-ranging meanings. They also highlighted that while some emojis have been used as substitutes for protected characteristics groups, or intensify the harmful nature of the message, the same emojis may also be used in condemning, empowering or awareness-raising contexts. It was underlined that, though automated

systems and LLMs show promise in detecting coded language, moderation still requires human oversight to properly define context and meaning.

5. Oversight Board Analysis

The Board selected these cases to explore the use of emojis as a form of "algospeak" and online racial harassment and discrimination. It also aims to assess Meta's enforcement approaches to such evolving forms of expression, both by human moderators and automated systems, particularly following Meta's announcement on January 7, 2025, that the company is modifying its approach to automated policy violations enforcement. The cases are relevant to one of the Board's seven strategic priorities, Hate Speech Against Marginalized Groups.

The Board analyzed Meta's decisions in these cases against Meta's content policies, values, and human rights responsibilities. The Board also assessed the implications of these cases for Meta's broader approach to content governance.

5.1 Compliance With Meta's Content Policies

I. Content Rules

The Board found that both posts violate the Hateful Conduct Community Standard prohibiting dehumanizing comparisons to animals. The applicable policy line specifically references the comparison of Black people to apes and ape-like creatures as an example of dehumanizing speech. Both posts utilize the monkey emoji to target Black people on the basis of their protected characteristic. Therefore, both pieces of content violate the Hateful Conduct policy.

In the Brazilian case, the post uses a scene from a movie in which two characters argue and claim ownership over a monkey. The added overlay text over the characters in the

video suggests that football teams, such as Real Madrid and Barcelona, argue over Brazilian up-and-coming football players (who are often Black) in the same way the men in the video argue over a monkey. This reading is supported by the additional overlay text referencing boys rising to prominence in Brazilian football. The use of the monkey emoji in the caption serves to further reinforce the intention of the post, implicitly comparing Brazilian football players to monkeys. The Board recognizes the alarming trend in sports, particularly in football, of fans utilizing monkey references to racially discriminate against Black athletes. Given this established pattern and existing context, the post violates the plain meaning of Meta's Hateful Conduct policy by dehumanizing Black people in the form of comparisons to or generalizations about animals.

In the Irish case, the comment challenges the parent post's condemnation of anti-Black racism in Ireland. It expresses eagerness for the situation to "blow up" and "to have glorious fun with all the [monkey emojis] out in the street." This reference to Black people as monkeys together with the use of laughing and praying emojis and longing for such "glorious days ahead" conveys the user's intent to dehumanize through equating Black people with monkeys. Given the increase of Afrophobia in Ireland and the fact that the comment is posted under a video expressing indignation about racism in Ireland, the use of monkey emojis in the comment is a clear reference to Black people, as a protected characteristics group, equating them to monkeys. Therefore, this post violates the Hateful Conduct Policy.

II. Enforcement action

These cases raise concerns about accuracy of enforcing the Hateful Conduct policy, especially when it comes to assessing the use of emojis as a form of algospeak.

Meta instructs its reviewers to always consider the full context in which the emoji appears to determine whether its use may be violating. This means that reviewers

should look at all aspects of the content, such as imagery, captions and text overlays, as well as factors beyond the immediate content, including the main post and related comments. As the emojis may convey varying meanings, this approach is in line with the Board's prior guidance for contextual, holistic assessment of posts (see, among others, Wampum Belt).

In these cases, both automated and multiple human reviews at various levels of enforcement failed to accurately assess the posts, keeping them on the platforms. The Board is concerned that although the classifiers detected the content in both cases, they took no action on those: The Brazilian post was deemed non-violating, while the classifier was unable to confidently determine that the Irish post was in English.

Further, the initial review in the Brazilian case involved routing and translation issues, and this initial decision was not reviewed further, despite dozens of reports. To ensure adequate review, Meta's enforcement processes should be designed to always direct the content to reviewers with appropriate language and regional expertise. Finally, the Board is concerned that in both cases, notwithstanding detailed guidance for contextual and comprehensive assessment for content involving emojis, both human reviewers on appeal upheld the initial decision to keep the posts on the platforms, despite the violating nature of the posts being clear.

Meta also explained that its classifiers are trained on datasets of reported and labeled examples, including cases where emojis are used in potentially violating ways. The company should improve the ability of its automated systems to accurately detect the use of emojis in violative contexts. Given the widespread use of emojis that may bear different meanings, Meta should periodically audit the training data used for Hateful Conduct policy enforcement, especially in regard to examples with emojis in all languages, and ensure that more robust datasets are included. The company should consider the changing nature of emoji use, relying on research findings that may include information on emoji usage trends on its platforms across languages and

regions (See also PC-31493). In line with Meta's commitments to develop and enforce its global rules in a non-discriminatory manner, Meta should ensure that the datasets include examples of non-English content with emojis.

5.2. Compliance With Meta's Human Rights Responsibilities

The Board finds that keeping both posts up on the platform was not consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19, para. 2 of the ICCPR provides that "everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media." General Comment No. 34 further specifies that protected expressions include those that may be considered "deeply offensive" (para. 11).

When 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). These requirements are often referred to as the "three-part test." The Board uses this framework to interpret Meta's human rights responsibilities in line with the United Nations (UN) Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression" (A/74/486, para. 41).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality requires rules limiting expression to be accessible and clear, formulated with sufficient precision to enable an individual to regulate their conduct accordingly (General Comment No. 34, para. 25). Additionally, these rules "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and must "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 UN Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). People using Meta's platforms should be able to access and understand the rules and content reviewers should have clear guidance regarding their enforcement.

The Board finds that the rules on dehumanizing comparisons to animals in general or specific types that are culturally perceived as inferior are sufficiently clear as applied to these cases. As such, the Hateful Conduct Community Standard clearly and publicly states that the comparison of Black people to apes or ape-like creatures is prohibited, highlighting the broad recognition of this racist analogy. Furthermore, Meta's internal guidance includes an illustrative and non-exhaustive list of emojis, including the monkey and banana emojis, which could signify a visual comparison between protected characteristic groups and animals.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights of others (Article 19, para. 3, ICCPR).

The UN Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance has noted that hate speech, including online forms, "has a powerful detrimental effect at the societal level, destroying the social fabric of communities and undermining the norms of human rights and democracy, including equality and non-discrimination." (A/78/538, para. 31, (2023)). The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has warned social media companies that "inflammatory speech not only endangers individuals and encourages self-censorship, but also divides communities by fueling fear, suspicion and hostility, breaking down social trust and weakening democratic dialogue and civic participation." (A/80/341, para. 30 (2025)).

The Board has previously recognized that the Hate Speech Community Standard (now Hateful Conduct policy) pursues 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 International Convention on the Elimination of All Forms of Racial Discrimination. See also Posts Displaying South Africa's Apartheid-Era Flag and Comment on Kenyan Politics Using a Designated Slur).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "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 No. 34, para. 34).

Individual posts

The Board acknowledges that content with emojis may carry multiple meanings, including being used for condemnation, self-referential or empowering purposes.

Similar to speech that can be used in both hateful and non-hateful manners, emojis require contextual analysis to fully understand their intended meaning. However, content using emojis that seek to incite discrimination, hostility or violence towards protected characteristic groups should be subject to removal.

Article 20, para. 2 of the ICCPR provides that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence is to be prohibited by law." This prohibition is "fully compatible with the right to freedom of expression as contained in article 19 [ICCPR], the exercise of which carries with it special duties and responsibilities," (General Comment No. 11, (1983), para. 2). The prohibition under Article 20 is also subject to Article 19's three-part test (General Comment 34, para 50-52).

The Board considers that the removal of these posts is necessary and proportionate to prevent Meta's platforms from being exploited to incite discrimination or hostile acts against protected characteristics groups, in these cases Black people. The Committee on the Elimination of Racial Discrimination considers that "incitement characteristically seeks to influence others to engage in certain forms of conduct, including the commission of crime, through advocacy or threats. Incitement may be express or implied, through actions such as displays of racist symbols or distribution of materials as well as words," (General Recommendation No. 35, para. 16 (2013)). The Committee notes that while "the notion of incitement as an inchoate crime does not require that the incitement has been acted upon, in regulating the forms of incitement [...] States parties should take into account, as important elements in the incitement offences [...] the intention of the speaker, and the imminent risk or likelihood that the conduct desired or intended by the speaker will result from the speech in question." (Id.)

A 2023 UN guide for policymakers and practitioners underlines the importance of identifying and addressing "non-verbal hate speech" expressed through videos, music,

memes, and other media, as well as “coded language” that can be more difficult to detect. The Board has previously cited the phrase “malign creativity,” coined by the Wilson Center, to refer to “the use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection on social-media platforms” (see Posts in Polish Targeting Trans People decision).

In some contexts, emojis may represent forms of “algospeak” to express hate, discrimination and harassment towards specific protected characteristic groups. The Board finds that the content in both cases fits within this trend, by clearly referencing and comparing Black people with monkeys.

In the Brazilian case, monkey imagery and emoji were used to compare Brazilian football players, many of whom are Black, to monkeys. The content was posted against a broader context of widely documented systemic racism and hostility in football, with a particular focus on targeted attacks against Black football players. The content was viewed over 22,000 times and shows a worrying trend of perpetuating racist stereotypes and inciting likely and imminent discrimination and hostile action against specific protected characteristic groups, in this case Black people.

In the Irish case, the commenter used the monkey emoji to compare all Black people, referenced in the main post, to monkeys. The comment was shared amid rising discrimination and exclusion of Black people in Ireland, on a parent post that received over 293,000 likes and 9,500 comments. In Knin Cartoon, the Board found that “depicting 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.” Similarly, in this case, the commenter was depicting Black people as monkeys. The comment encouraged analogous hostile behavior to that described in the main post, thereby inciting discrimination and hostility.

These posts illustrate how emojis can be used to urge others to take discriminatory and potentially hostile action. Less severe interventions, such as labels, warning screens or other measures to reduce dissemination, would not provide adequate protection against the effects of leaving content of this nature on the platform. Therefore, their removal was warranted.

In more ambiguous cases, the Board encourages Meta to carefully continue exploring less intrusive measures, in line with the Board's recommendation in the Pro-Navalny Protests in Russia decision. These would allow users to self-remediate or foster understanding that user's content can be impacting others negatively. In developing such measures, Meta should ensure they are effective and do not lead to adverse human rights impacts (see Comment on Kenyan Politics Using a Designated Slur, recommendation no.1).

Broader issues

The Board's own research and reports illustrate that such content often targets specific individuals based on their protected characteristics, especially in the context of sports, such as football. Meta removes directed mass harassment under the Bullying and Harassment policy, when the content is escalated to its subject matter expert review. The company also introduced several user control measures to tackle abuse. For example, on Instagram, users can manage multiple unwanted comments in one go or bulk block accounts that posted them, or set up comment filters "to prevent others from leaving offensive comments that use words, phrases or emojis they don't want to see."

To better coordinate its efforts and protect users who may not be directly named but are implicit targets of hateful campaigns, Meta should develop a framework to harmonize its already-existing measures to proactively disrupt hateful campaigns, especially those involving the use of emojis. This should include campaigns for both

private and public figures through both direct/explicit and indirect/implicit mentions that trigger either the Bullying and Harassment or Hateful Conduct policies. The framework will ensure Meta has a cohesive approach to address gaps in its moderation systems, identify and evaluate coordinated and targeted hateful campaigns, and set up permanent feedback channels.

Ensuring that its systems are well equipped to handle targeted campaigns is also important in preparation for major sporting events, in particular football. Numerous documented incidents both online and at stadiums indicate an alarming trend of racist animosity between groups of supporters and spectators. Meta should ensure its time-sensitive mitigation efforts include active monitoring of content with emojis that incite targeted discrimination or hostility in the lead up, during and in the immediate aftermath of major sporting events, e.g., the 2026 FIFA World Cup. This could be achieved through establishing a cross-functional team of subject matter experts from across the company to "respond in real time to potential problems and trends," similar to an Integrity Product Operations Center; implementing a fast-track review process for appeals related to violations connected to sporting events; and real-time trends monitoring to detect spikes of content with emojis that target specific individuals or protected characteristic groups. Meta should also engage with FIFA and other professional sports associations to stay informed of relevant trends and dynamics.

6. The Oversight Board's Decision

The Board overturns Meta's original decision to keep up the content in both cases under review.

7. Recommendations

Enforcement

1. To improve the ability of its automated systems to more accurately detect the use of emojis in violative contexts, Meta should audit its training data used for Hateful Conduct policy enforcement and ensure the data is updated periodically to include examples of content with emojis in all languages, violating use of emojis and new instances of the hateful use of emojis.

   The Board will consider this recommendation implemented when Meta provides the Board with detailed results of its first audit and the necessary improvements that the company will implement as a result.

2. To better protect users who are not directly named but the implicit targets of hateful campaigns, Meta should harmonize its existing efforts to proactively disrupt hateful campaigns, especially those involving the use of emojis. This should include campaigns that involve both private and public figures through both direct/explicit and indirect/implicit mentions that trigger either the Bullying and Harassment or Hateful Conduct policies.

   The Board will consider this recommendation implemented when Meta shares with the Board its updated enforcement practices for targeted hateful campaigns.

3. To ensure its systems are well-equipped to address hateful campaigns during major sporting events, such as the FIFA World Cup, Meta should ensure that its time-sensitive mitigation efforts, be that through its Integrity Product Operations Center or another risk mitigation system, include active monitoring of content with emojis that incite targeted discrimination or hostility in the lead up, during and in the immediate aftermath of these events.

The Board will consider this recommendation implemented when Meta shares with the Board evidence confirming the deployment of its risk evaluations and mitigation efforts used during major sporting events.

The Board also reiterates the importance of its previous recommendations, noting their relevance to this issue (recommendation no. 1 from Comment on Kenyan Politics Using a Designated Slur). In line with those recommendations, Meta should:

- Provide users with an opportunity for self-remediation comparable to the post time friction intervention that was created as a result of the Pro-Navalny Protests in Russia recommendation no. 6. If this intervention is no longer in effect, Meta should provide a comparable product intervention.

Procedural Note:

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

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology.