Posts Displaying South Africa's Apartheid-Era Flag

(2025-001-FB-UA, 2025-002-FB-UA)

Summary

Following a review of two Facebook posts containing images of South Africa's 1928-1994 flag, the majority of the Board has upheld Meta's decisions to keep them up. Board Members acknowledge the long-term consequences and legacy of apartheid on South Africa. However, these two posts do not clearly advocate for exclusion or segregation, nor can they be understood as a call for people to engage in violence or discrimination. The deliberation in these cases also resulted in recommendations to improve conflicting language in the Dangerous Organizations and Individuals policy.

Additional Note: Meta's January 7, 2025, revisions did not change the outcome in these cases, though the Board took the rules at the time of posting and the updates into account during deliberation. On the broader policy and enforcement changes hastily announced by Meta in January, the Board is concerned that Meta has not publicly shared what, if any, prior human rights due diligence it performed in line with its commitments under the UN Guiding Principles on Business and Human Rights. It is vital Meta ensures adverse impacts on human rights globally are identified and prevented.

About the Cases

Shared ahead of South Africa's general elections in May 2024, the first Facebook post shows a photo of a white male soldier holding the country's old flag, in use during the apartheid era. A caption urges others to share the post if they "served under this flag." This post was viewed more than 500,000 times. Reported by three users, Meta decided the content did not break its rules.

The second post, also on Facebook, comprises stock photos from the apartheid era, including, the former flag, white children standing next to a black man on an ice cream bicycle, a public whites-only beach and a toy gun. The post's caption says these were the good old days, asks others to "read between the lines," and includes winking face and "OK" emojis. Viewed more than two million times, the content was reported by 184 users, mostly for hate speech. Meta's human reviewers decided the post did not violate the Community Standards.

In both cases, users who reported the content to Meta then appealed to the Board.

Key Findings

The majority of the Board has found that neither post violates the Hateful Conduct policy, while a minority finds both are violating.

The policy does not allow "direct attacks" in the form of "calls or support for exclusion or segregation" based on a protected characteristic. Neither post advocates for bringing back apartheid or any other form of racial exclusion, according to the majority. While the soldier post uses the flag in a positive context, it does not advocate racial exclusion or segregation. For the photo grid post, the images combined with the emojis and the message to "read between the lines" indicate a racist message, but they do not rise to the level needed to violate this policy.

A minority disagrees, pointing out the flag is an unambiguous and direct symbol of apartheid, which when shared with positive or neutral references, can be understood as support for racial segregation. For example, there is no doubt that the photo grid post, with its images of segregated life, messages and the "OK" emoji – understood by white supremacists globally as covert hate speech – supports racial exclusion.

The Board has found unanimously that both posts violate the Dangerous Organizations and Individuals policy although the majority and a minority disagree over why. The company removes content that glorifies, supports or represents hateful

ideologies, including white supremacy and separatism, as well as “unclear references” to these ideologies. The Board agrees the 1928-1994 flag cannot be decoupled from apartheid, a form of white separatist ideology. For the majority, both posts represent unclear references to white separatism, while for the minority, they explicitly glorify this ideology.

It is not necessary and proportionate to remove the content, according to the majority, because the likelihood of imminent discrimination or violence from these posts is low. Banning such speech does not make intolerant ideas disappear and other content moderation tools less intrusive than removals could have been applied. A minority disagrees, noting that removal is necessary to ensure respect for equality and non-discrimination for non-white South Africans. They also point out the chilling effects of such hatred accumulating on Meta’s platforms on the freedom of expression of those targeted.

All Board Members recognize conflicting language around “references” to hateful ideologies under the Dangerous Organizations and Individuals Community Standard. During the deliberation, there were questions over why Meta does not list apartheid as a standalone designation. Some Board Members asked why Meta’s list centers on those ideologies that may present risks in Global Minority regions but remains silent on comparable hateful ideologies in the Global Majority.

The Oversight Board’s Decision

The Oversight Board upholds Meta’s decisions to leave up both posts.

The Board recommends that Meta:

- In respect of the January 7, 2025, updates to the Hateful Conduct Community Standard, Meta should identify how the policy and enforcement updates may adversely impact populations in Global Majority regions. It should adopt measures to prevent and/or mitigate these risks and monitor their

effectiveness. Finally, Meta should update the Board every six months on its progress, reporting on this publicly at the earliest opportunity.

- Adopt a single, clear and comprehensive explanation of how its prohibitions and exceptions under the Dangerous Organizations and Individuals Community Standard apply to designated hateful ideologies.
- List apartheid as a standalone designated hateful ideology in the rules of the Dangerous Organizations and Individuals Community Standard.
- Provide more global examples to its reviewers of prohibited glorification, support and representation of hateful ideologies, including examples that do not directly name the listed ideology.

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

Full Case Decision

1. Case Description and Background

These cases involve two Facebook posts shared in the run-up to South Africa's general election in May 2024.

The first post shows a photo of a white male soldier holding South Africa's pre-1994 flag, which was the country's flag under apartheid. The English caption urges users to share the content if they "served under this flag." The post was viewed around 600,000 times and shared around 5,000 times. Three users reported the content to Meta for hate speech and violence. As Meta's human reviewers found the content to be non-violating, it was kept up. One of the users who reported the content then appealed to the Board.

The second post is a photo grid containing stock images taken during the apartheid era, including: the country's former flag; an adult Black man on an ice cream bicycle with three white children standing next to him in a seemingly whites-only neighborhood; a public whites-only beach with a theme park; a South

African board game; a packet of white candy cigarettes; and a silver toy gun. The caption states these were the “good old days” and asks the audience to “read between the lines,” followed by winking face and “OK” emojis. It was viewed around two million times and shared around 1,000 times. Within a week of posting, 184 users reported the content, mostly for hate speech. Some of the reports were assessed by human reviewers, who determined the content did not violate the Community Standards. The remaining reports were processed through a combination of automated systems and prior human review decisions. Like the soldier post, Meta found this content to be non-violating and kept it up on the platform. One of the users who reported the content then appealed to the Board.

On January 7, 2025, Meta announced revisions to its Hate Speech policy, renaming it the Hateful Conduct policy. These changes, to the extent relevant to these cases, will be described in Section 3 and analyzed in Section 5. The Board notes content is accessible on Meta’s platforms on a continuing basis, and updated policies are applied to all content present on the platform, regardless of when it was posted. The Board therefore assesses the application of policies as they were at the time of posting, and, where applicable, as since revised (see also the approach in Holocaust Denial).

The Board notes the following context in reaching its decision:

From 1948 to 1994, South Africa was under a state-sanctioned apartheid regime involving the racial segregation of white and non-white South Africans, although discriminatory laws had existed in the country before apartheid was formally adopted. During this time, South Africa was represented by an orange, white and blue flag. In 1994, following the end of apartheid, South Africa adopted the six-color flag that it uses today. Despite the end of apartheid, socioeconomic inequality continues to afflict the non-white population of the country in particular, contributing to racial tensions in politics and public discourse.

In 2018, the Nelson Mandela Foundation took legal action in South Africa seeking to ban the "gratuitous display" of the apartheid-era flag following its use in protests the previous year. The action alleged that it amounted to "hate speech, unfair discrimination and harassment," and that it celebrated the system's atrocities. In 2019, South Africa's Equality Court held that the flag's gratuitous display amounted to hate speech and racial discrimination that can be prosecuted under domestic law. The court ruling clarified that displaying the flag is not illegal if used for artistic, academic, journalistic or other public interest purposes. The Supreme Court of Appeal (SCA) upheld this decision in April 2023.

On May 29, 2024, South Africa held elections for the National Assembly. The African National Congress (ANC), the political party led by Nelson Mandela after the end of apartheid, lost its parliamentary majority. However, incumbent party leader Cyril Ramaphosa retained his presidency by forming a coalition government with opposition parties.

2. User Submissions

The authors of the posts were notified of the Board's review and provided with an opportunity to submit a statement. No response was received.

In their statement to the Board, the user who reported the soldier post stated that South Africa's former flag is comparable to the German Nazi flag. They said "brazenly displaying" it incites violence because the country is still reeling from the impact of apartheid as a crime against humanity. The user also stated that sharing such images during an election period can encourage racial hatred and endanger lives. Similarly, the user who reported the photo grid post explained that the use of the flag is illegal and taken as a whole, it suggests apartheid was a "better time" for South Africans. They emphasized how the former flag represents oppression and is "derogatory" and "painful" for the majority of South Africans.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Hateful Conduct (previously named Hate Speech) Community Standard

Meta's Hateful Conduct policy states that "people use their voice and connect more freely when they don't feel attacked on the basis of who they are." Meta defines "hateful conduct" in the same way that it previously defined "hate speech," as "direct attacks against people" based on protected characteristics, which include race, ethnicity and national origin. As a result of the Board's recommendation to clarify its approach in the Knin Cartoon case, Meta states in the introduction to its Community Standards that the company 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.

Tier 2 of the Hateful Conduct policy prohibits as a form of direct attack "calls or support for exclusion or segregation or statements of intent to exclude or segregate," whether in written or visual form. Meta prohibits the following types of calls for or support for exclusion: (i) general exclusion, which means calling for general exclusion or segregation, such as "No X allowed!"; (ii) political exclusion, which means denying the right to political participation or arguing for incarceration or denial of political rights; (iii) economic exclusion, which generally means denying access to economic entitlements and limiting participation in the labor market; and (iv) social exclusion, which means things like denying access to physical and online spaces and social services. Prior to January 7, the prohibition on "general exclusion" was called "explicit exclusion."

Dangerous Organizations and Individuals

Meta's Dangerous Organizations and Individuals policy seeks to "prevent and disrupt real-world harm."

Under the policy rationale, Meta states that it removes content that glorifies, supports or represents "hateful ideologies."

Meta explains it designates prohibited ideologies, which the policy lists as "including Nazism, white supremacy, white nationalism [and] white separatism" because they are "inherently tied to violence" and attempt "to organize people around calls for violence or exclusion of others based on their protected characteristics." Directly alongside this listing, the company states it removes *explicit* glorification, support and representation of these ideologies (emphasis added).

Meta twice states it also removes "unclear references" to hateful ideologies, once in the policy rationale and again under the description of Tier 1 organizations.

Meta explains in the policy rationale that it requires users to "clearly indicate their intent" when creating or sharing such content. If a user's intent is "ambiguous or unclear," Meta defaults to removing content.

II. Meta's Submissions

Meta left both posts on Facebook, finding no violations of its policies. Meta confirmed that its analysis of the content was not affected by the January 7 policy changes.

Meta stated that the posts did not violate the Hateful Conduct policy, as there were no calls for exclusion of a protected group under Tier 2, nor any other prohibited direct attack. None of the statements in the posts mentioned a protected group, nor did the posts advocate for a particular action. According to Meta, for the policy to be operable at-scale, there must be a "direct" and explicit attack, not an implicit attack. Neither post had a direct attack.

Meta's internal enforcement guidance to reviewers contains an illustrative list of emojis that are violating if used in a context that allows a reviewer to confirm intent to directly attack a person or group on the basis of a protected characteristic. Photos, captions, text overlay on photos and the content of videos can help indicate what an emoji means. The list is global and does not contain the "OK" emoji.

Meta decided the posts did not violate the Dangerous Organizations and Individuals policy. Meta noted that the flag shown in the posts was used in South Africa between 1928 and 1994, including the apartheid era and the years preceding it. The company acknowledged that since the end of apartheid, this flag has sometimes been used in historical commemoration but is most often used as a symbol of Afrikaner heritage and apartheid. However, it also recognized that the flag represents other meanings, including South Africans' connections to different aspects of that period such as personal experiences, military service and other aspects of citizenship.

Regarding Meta's prohibition on explicit glorification, support or representation of hateful ideologies, the company noted in its guidance to reviewers that only Nazism, white supremacy, white nationalism and white separatism are named as hateful ideologies. Meta did, however, explain to the Board that it removes "praise of segregation policies" like those implemented during apartheid in South Africa as white separatism. In response to Board requests for examples, Meta said it would remove a statement like "apartheid was wonderful" in most instances, but this is not an example provided to reviewers in the enforcement guidance. Examples of policy violations provided to reviewers include, among others, "white supremacy is the right thing" and "yes, I am a white nationalist."

Meta considered that the soldier post's statement, "Share if you served under this flag," did not glorify or support a designated hateful ideology. Likewise, the photo grid post's caption describing the apartheid era as the "good old days" and asking users to "read between the lines" [wink emoji, "OK" emoji], combined with the apartheid flag and historical images of that era, do not, by themselves, glorify or

support a hateful ideology. While Meta acknowledges the "OK" emoji is in some contexts associated with the white power movement, Meta's view is it predominantly means "okay," including in South Africa. Meta concluded its use here is not meant to glorify or support a hateful ideology.

As part of its integrity efforts for the May 2024 South African elections, Meta ran anti-hate speech and misinformation campaigns on its platforms and local radio in the election lead-up. These campaigns were designed to educate people about identifying and reporting hate speech and misinformation online.

The Board asked questions on the renamed Hateful Conduct and Dangerous Organizations and Individuals policies and their enforcement, which symbols and ideologies could violate these policies and Meta's electoral integrity efforts in South Africa. Meta responded to all questions.

4. Public Comments

The Oversight Board received 299 public comments that met the terms for submission. Of those, 271 were submitted from sub-Saharan Africa, 10 from Europe, four from Central and South Asia, five from the US and Canada, seven from the Middle East and North Africa, and two from Asia-Pacific and Oceania. Because the public comments period closed before January 7, 2025, none of the comments address the policy changes Meta made on that date. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: what the apartheid-era flag meant in South African history and politics; the impact of displaying it on non-whites and efforts to build a multi-cultural South Africa, and whether it should be allowed on Meta's platforms; and, coded uses of online symbols and recommended approaches to moderating visual images that may constitute implicit attacks against protected groups.

5. Oversight Board Analysis

The Board selected these cases to address Meta's respect for freedom of expression and other human rights in the context of an election, and how it treats imagery associated with South Africa's recent history of apartheid. These cases fall within the Board's strategic priorities of Elections and Civic Space and 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

Hateful Conduct (formerly Hate Speech) Community Standard

The Board notes that Meta's prohibition on "calls or support for exclusion or segregation" based on a protected characteristic is open to at least two interpretations, neither of which is impacted by the January 7 policy changes. The majority of the Board, noting Meta's paramount value of voice, favors a narrow reading of the rule requiring advocacy for exclusion or segregation. A minority, noting Meta's value of dignity, applies a broader reading, interpreting the prohibition to also encompass support for exclusion or segregation more generally.

The majority of the Board finds that neither post violates this prohibition. While the posts appear to display nostalgia for the apartheid era, they do not advocate reinstituting apartheid or any other form of racial exclusion.

Considering the soldier post, the majority recognizes that many people see the 1928–1994 flag as a symbol of apartheid. However, the flag itself, combined with a statement about military service, does not advocate exclusion or segregation. Additional elements would need to be present in the post to make it violating. While

the flag is invoked positively in this post, that context is specific to military service and there is no sufficiently clear statement or reference that apartheid or similar policies should be reinstituted. Notwithstanding how divisive or insensitive sharing this flag may be to many in present-day South Africa, it would be incorrect to presume, without more evidence, that this post advocates racial exclusion or segregation that would violate this policy.

The majority similarly finds that the photo grid post, with the image of the 1928–1994 flag alongside photographs of apartheid-era South Africa and the caption, does not advocate segregation or exclusion. They feasibly evoke general nostalgia for the period they depict. The majority acknowledges that the phrases "the good old days" and "read between the lines," and the winking face and "OK" emojis are all, in combination with the photographs, indicators of a racist message that change how the images alone would be perceived. Nevertheless, Meta's Hateful Conduct policy does not prohibit the expression of all racially insensitive or even racist viewpoints. The post, taken as a whole, does not rise to the level of advocacy for the reinstitution of apartheid or other forms of racial segregation or exclusion and is therefore permitted.

For a minority, the 1928–1994 flag is an unambiguous and direct symbol of apartheid. When shared with a positive or neutral reference (rather than with condemnation), it is contextually understood in South Africa as support for racial segregation and exclusion and therefore is violating. For this minority, an innocuous display of the flag is not possible and can only be interpreted as support for the racial exclusion of the apartheid era (also see public comments, including from the South African Human Rights Commission and Nelson Mandela Foundation, noting the 2023 SCA decision, PC-30759; PC-30771; PC-30768; PC-30772; PC-30774). The apartheid-era flag has also been co-opted by white nationalist movements in other parts of the world (PC-30769).

For these reasons, the minority finds that both posts constitute support for racial exclusion. The soldier post, encouraging others to reshare the flag, can only be

understood as support for the segregationist policy the flag represents. Considering the photo grid post as a whole, the images and caption make the post's support for racial exclusion and segregation clear. As the post includes the flag without condemning or awareness-raising context, it violates on this basis alone. In addition, the other photographs appear to be stock images of aspects of life that were segregated; they do not tell a personal story of nostalgia, as the caption also makes clear. The use of the white power "OK" emoji in the caption is significant. It is understood by white supremacists globally as covert hate speech and a dog whistle, literally spelling out the letters "W" (for white) with three fingers and a "P" (for power) with the connecting thumb and index finger (see PC-30768). Its inclusion here was not in isolation. When accompanied with images of apartheid, a reference to "the good old days" and an invitation to users to "read between the lines," together with a wink emoji, even a person not accustomed to white supremacist symbology is left in no doubt that this post supports racial exclusion and is therefore violating. For the minority, in reaching this conclusion, it is important to understand how the use of racist language and symbols online has adapted to evade content moderation, and how more subtle (but nevertheless direct) expressions of support for exclusion can be used to connect like-minded people. As here, coded or indirect hate speech can be alarmingly non-ambiguous, even when it leaves literal statements of intended meaning unsaid.

Dangerous Organizations and Individuals Community Standard

Through questions the Board asked Meta, the Board understands that the company's designation of white separatism as a hateful ideology includes South African apartheid. However, internal guidance to Meta's reviewers could make this more explicit by providing broader examples of violations. As addressed in Section 5.2 (legality) below, Meta's rules on designated ideologies are vague. The Board unanimously finds that both posts violate the Dangerous Organizations and Individuals Community Standard, but for different reasons. For the majority, both posts meet Meta's definition of unclear references to white separatism, which the

policy prohibits. For a minority, both posts rise to the level of glorification of white separatism.

The Board notes that the 1928–1994 South African flag cannot be decoupled from apartheid as a form of white separatist ideology. It was the national flag during two decades of legalized racial discrimination preceding apartheid and from when apartheid was instituted in 1948.

For the majority, the soldier post, which encourages others to reshare if they served in the military under the flag, does not explicitly glorify apartheid as a form of white supremacy in its express and positive reference to military service. Similarly, the photo grid post does not indicate whether it is alluding to personal experiences during the apartheid era or glorifying it. As noted above, however, there are several indicators of a racist message in this post, most notably the use of the "OK" emoji alongside the flag. For the majority, the positive but indirect indicators in both posts constitute a violating "unclear reference" to white separatism but are not sufficiently explicit to amount to "glorification."

For a minority of the Board, both posts meet the threshold for explicit glorification of white separatist ideology for the same reasons they constitute support for racial exclusion or segregation under the Hateful Conduct policy. In the soldier post, positive reference to the apartheid-era flag as an inherent symbol of white separatism, including in the context of military service, constitutes glorification of that ideology, even without apartheid policies being specifically mentioned. For the photo grid post, the combination of the white power symbol ("OK" emoji), the flag and the phrase "the good old days," also explicitly glorifies this ideology. For these Board Members, users reporting both posts and the comments the posts attracted confirm that the content's glorification of apartheid was well understood by audiences. Many reactions to the posts demonstrate how white separatists' crude communications can creatively evade content moderation. They also show how networked hateful actors can exploit the design of Meta's platforms to spread their message, identify new members and expand their numbers.

5.2 Compliance With Meta's Human Rights Responsibilities

The majority of the Board finds that keeping both posts up on the platform was consistent with Meta's human rights responsibilities. A minority disagrees, finding that removal would be consistent with these responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including views about politics, public affairs and human rights (General Comment No. 34, paras. 11-12). The UN Human Rights Committee has highlighted that the value of expression is particularly high when discussing political issues (General Comment No. 34, paras. 11, 13). It has emphasized that freedom of expression is essential for the conduct of public affairs and the effective exercise of the right to vote (General Comment No. 34, para. 20; also see General Comment No. 25, paras. 12 and 25). 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 UN Guiding Principles on Business and Human Rights, which Meta itself has committed to in its Corporate Human Rights Policy. The Board does this 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 enable them to ascertain what sorts of expression are properly restricted and what sorts are not" (*Ibid.*). 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.

These cases highlight two problems of clarity regarding Meta's Hateful Conduct prohibitions. First, Meta provided conflicting responses to the Board on whether "direct attacks" include implicit statements or not (similar concerns were raised in the Board's Knin Cartoon case). Additionally, whether the rule on "calls or support for exclusion or segregation" is limited to advocacy for exclusion or encompasses any broader support of exclusion or segregation is also unclear. This is compounded by a lack of global examples of violations provided to reviewers, with none encompassing apartheid.

The Dangerous Organizations and Individuals Community Standard also presents conflicting language on Meta's approach to hateful ideologies. In some parts, it specifies that unclear references to hateful ideologies are prohibited, while in others it implies that only "explicit glorification, support or representation" is prohibited. The internal guidance provided to reviewers states that "[r]eferences, [g]lorification, [s]upport, or [r]epresentation" are all prohibited. The list of prohibitions under the "we remove" section of the policy does not refer to the rule on hateful ideologies at all, creating further confusion.

While the Board finds "white separatism" should implicitly include apartheid as implemented in South Africa, Meta's internal guidance to reviewers does not make this explicit nor include sufficient examples relevant to the South African context.

The examples of violating content provided to the Board by Meta in response to questions (e.g., "apartheid was wonderful," "white supremacy is the right thing") do not reflect the realities of how racial supremacist messaging is often framed. At the same time, the Board notes that while apartheid, as implemented in South Africa, is inherently intertwined with white separatism and white supremacy, the concept of apartheid in international law applies to the intentional dominion of any racial group over another to systematically oppress them (Rome Statute of the International Criminal Court, Article 7(2)(h); Apartheid Convention, Article 2). This raises questions as to why apartheid is not listed as a standalone designation. As Meta's policies are global, several Board Members also questioned why Meta's listing is centered around ideologies that may present risks in Global Minority regions while remaining silent on many comparable hateful ideologies in Global Majority regions.

II. Legitimate Aim

Any restriction on freedom of expression should 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 Board has previously recognized that the Hate Speech Community Standard 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 ICERD). This is true also of the revised Hateful Conduct policy.

Similarly, the Board considers that the Dangerous Organizations and Individuals policy, seeking to "prevent and disrupt real-world harm," pursues the legitimate aim of protecting the rights of others, such as the right to life (ICCPR, Article 6) and the right to non-discrimination and equality (ICCPR, Articles 2 and 26).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality require 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).

The majority of Board Members finds that keeping up both posts is in line with Meta's human rights responsibilities, and removal would not be necessary and proportionate. These Board Members acknowledge that apartheid's legacy persists and has had long-term consequences felt across South Africa today. At the same time, international human rights law affords heightened protection to freedom of expression that relates to political participation, including in the context of elections (General Comment No. 25, paras. 12 and 25). In the Politician's Comments on Demographic Changes case, the Board affirmed that expressions of controversial opinion are protected by international human rights standards. Both these posts constitute protected expression. Even if considered to be "deeply offensive," this does not convert them to incitement to likely and imminent discrimination (see General Comment No. 34, (2011), para. 11; see also para. 17 of the 2019 report of the UN Special Rapporteur on freedom of expression, A/74/486).

The majority emphasizes that the UN Special Rapporteur on freedom of expression has made it clear that speech bans can be justified when there is imminent and likely concrete harm. But when the harms are not likely or imminent, other measures can be deployed (see A/74/486, paras. 13, 54). Similarly, the UN Human Rights Committee has stated: "Generally, the use of flags, uniforms, signs and banners is to be regarded as a legitimate form of expression that should not be restricted, even if such symbols are reminders of a painful past. In exceptional cases, where such symbols are directly and predominantly associated with incitement to discrimination, hostility or violence, appropriate restrictions should apply," (General Comment No. 37 on the right of peaceful assembly, CCPR/C/GC/37, para. 51).

For the majority, there are concerns about the excessive breadth of the Dangerous Organizations and Individuals Community Standard's prohibition on "unclear references." As Meta looks to reduce mistakes in its content moderation, as announced on January 7, these Board Members encourage an examination of how accurate and precise the enforcement of the "unclear references" rule is, as well as the compatibility of removals with Meta's human rights responsibilities.

The Board has often used the Rabat Plan of Action's six-factor test to assess if incitement to violence or discrimination is likely and imminent. The majority finds this was not met by either post. For the majority, the likelihood of imminent discrimination or violence posed by the content is low for a variety of reasons. As noted above, the historical context of apartheid in South Africa and its continuing legacy is important to the interpretation of these posts. At the same time, the country's relatively stable representative democracy since the end of apartheid and its robust legal framework for protecting human rights, are also relevant, particularly as it underwent elections at the time of these posts. Experts consulted by the Board noted that white supremacist rhetoric was not a major issue during the May 2024 elections. They said the period leading up to those elections was not characterized by interracial violence nor calls for violence from the white minority against other racial or ethnic groups. Neither post is from a high-profile or influential speaker, reducing the risk that either post would persuade anyone to engage in imminent acts of discrimination or violence. Neither post includes calls for action. The posts do not contain a clear intent to advocate for future acts of discrimination or violence nor would they be understood as a call to people to engage in such acts. Given these various factors, the majority determines that it was not likely nor imminent that violence or discrimination would have resulted from these posts.

Banning highly offensive speech that does not incite imminent and likely harm does not make intolerant ideas disappear. Rather, people with those ideas are driven to other platforms, often with like-minded people rather than a broader range of individuals. This may exacerbate intolerance instead of enabling a more transparent, public discourse about the issues.

The majority believes a variety of other content moderation tools short of removals could have served as a less intrusive means to achieve legitimate aims in these cases. The majority acknowledges the potential negative emotional ramifications of content in these cases as well as Meta's legitimate aim of seeking to prevent discrimination. As the Board stated in one of its very first opinions (Claimed Covid Cure), the company should first seek to achieve legitimate aims by deploying measures that do not infringe on speech. If that is not possible, the company should select the least intrusive tool for achieving the legitimate aim. Then, it should monitor that the selected tool is effective. Meta should use this framework in publicly justifying its rules and enforcement actions.

Indeed, the UN Special Rapporteur on freedom of expression has noted (A/74/486, para 51): "Companies have tools to deal with content in human rights-compliant ways, in some respects a broader range of tools than that enjoyed by States." The Board urges Meta to transparently explore expanding its enforcement toolkit and introduce intermediate measures in enforcing its Hateful Conduct Community Standard, instead of defaulting to a binary choice of keep up or take down. In the Myanmar Bot case, the Board found that "heightened responsibilities should not lead to default removal, as the stakes are high in both leaving up harmful content and removing content that poses little or no risk of harm." The Board urges Meta to examine how removing content can be an extreme measure that adversely impacts freedom of expression online. It also urges the company to consider other tools, such as the removal of content from recommendations or reduced distribution in users' feeds, in appropriate circumstances.

A minority of Board Members finds that removing both posts would be a necessary and proportionate limit on freedom of expression to ensure respect for the right to equality as well as freedom from discrimination for non-white South Africans. The minority is guided by the Rabat Plan factors to assess the risks posed by potential hate speech, including the harms these posts contributed to (op. cit).

In particular, a minority notes the public comments from the Nelson Mandela Foundation and the South African Human Rights Commission, among others. These confirm the various ways in which expression on Meta's platforms, supporting, justifying or otherwise glorifying segregation, contributes to the persistence of discrimination following apartheid (PC-30759; PC-30771; PC-30768; PC-30772; PC-30774). Comments beneath each post, largely in Afrikaans, which reveal a sense of white supremacy rooted in colonialism, confirm for this minority that the intent of the speaker to advocate hatred in an environment of severe discrimination was successful. The minority note that in the Depiction of Zwarte Piet case, the majority of the Board upheld the removal of a post based on its effects on the self-esteem and mental health of Black people, even when those effects may not have been directly intended by the speaker. This case is relevant beyond South Africa. Experts the Board consulted noted that symbols of apartheid, including the 1928–1994 flag, have been co-opted by white nationalist movements in other parts of the world too. This includes Dylann Roof, who gunned down nine members of a Black congregation church in the United States in 2015. A photo of Roof included on his social media shows him wearing a jacket with a patch of the apartheid-era flag (PC-30769).

A minority, moreover, reiterates that Meta, as a private actor, may remove hate speech that falls short of the threshold of incitement to imminent discrimination or violence when this meets the ICCPR Article 19(3) requirements of necessity and proportionality (report A/HRC/38/35, para. 28). In the South Africa Slurs case, the Board upheld Meta's removal of a racial slur relying heavily on the particularities of the South African context. For a minority in this case, the removal of both posts is necessary not only to prevent discrimination but also to ensure that the accumulation of hatred on the platform does not have a chilling effect on the freedom of expression of people repeatedly targeted by hate speech (see also Depiction of Zwarte Piet, Communal Violence in Indian State of Odisha, Armenians in Azerbaijan and Knin Cartoon). For the minority, the consequences on users' human rights from content moderation (specifically, the removal of speech and feature limits or suspensions for recurring violations) are significantly different from those enforcing laws on hate speech (such as fines or imprisonment). For these

reasons, a minority finds that removing both posts in accordance with the Hateful Conduct rule on exclusion, as well as the Dangerous Organizations and Individuals prohibition on "glorification," would be necessary and proportionate. A minority notes that accurately scaled enforcement of the Dangerous Individuals and Organizations exception on social and political discourse should ensure this set of rules is not overenforced.

Human Rights Due Diligence

Principles 13, 17 (c) and 18 of the UNGPs require Meta to engage in ongoing human rights due diligence for significant policy and enforcement changes, which the company would ordinarily do through its Policy Product Forum, including engagement with impacted stakeholders. The Board is concerned that Meta's January 7, 2025, policy and enforcement changes were announced hastily, in a departure from regular procedure, with no public information shared as to what, if any, prior human rights due diligence it performed.

Now these changes are being rolled out globally, it is important that Meta ensures adverse impacts of these changes on human rights are identified, mitigated and prevented, and publicly reported. This should include a focus on how communities may be differently impacted, including in Global Majority regions. In relation to enforcement changes, due diligence should be mindful of the possibilities of both overenforcement (Call for Women's Protest in Cuba, Reclaiming Arabic Words) as well as underenforcement (Holocaust Denial, Homophobic Violence in West Africa, Post in Polish Targeting Trans People).

The Board notes that many of these changes are being rolled out worldwide, including in Global Majority countries like South Africa and others with a recent history of crimes against humanity, not limited to apartheid. It is especially important Meta ensures that adverse impacts of these changes on human rights in such regions are identified, mitigated, prevented and accounted for publicly as soon as possible, including through robust engagement with local stakeholders. The

Board notes that in 2018, Meta cited the failure to remove hate speech from Facebook in crisis situations like Myanmar as motivation for increasing reliance on automated enforcement. In many parts of the world, users are less likely to engage with Meta's in-app reporting tools for a variety of reasons, making user reports an unreliable signal of where the worst harms could be occurring. It is therefore crucial that Meta considers fully how the effects of any changes to automated detection of potentially violating content, both for under- and overenforcement, may have uneven effects globally, especially in countries experiencing current or recent crises, war or atrocity crimes.

6. The Oversight Board's Decision

The Oversight Board upholds Meta's decisions to leave up both pieces of content.

7. Recommendations

Content Policy

1. As part of its ongoing human rights due diligence, Meta should take all of the following steps in respect of the January 7 updates to the Hateful Conduct Community Standard. First, it should identify how the policy and enforcement updates may adversely impact populations in global majority regions. Second, Meta should adopt measures to prevent and/or mitigate these risks and monitor their effectiveness. Third, Meta should update the Board on its progress and learnings every six months, and report on this publicly at the earliest opportunity.

   The Board will consider this recommendation implemented when Meta provides the Board with robust data and analysis on the effectiveness of its prevention or mitigation measures on the cadence outlined above, and when Meta reports on this publicly.

2. To improve the clarity of its Dangerous Organizations and Individuals Community Standard, Meta should adopt a single, clear and comprehensive explanation of how its prohibitions and exceptions under this Community Standard apply to designated hateful ideologies.

   The Board will consider this recommendation implemented when Meta adopts a single, clear and comprehensive explanation of its rule and exceptions related to designated hateful ideologies (under "we remove").

3. To improve the clarity of its Dangerous Organizations and Individuals Community Standard, Meta should list apartheid as a standalone designated hateful ideology in the rules.

   The Board will consider this recommendation implemented when Meta adds apartheid to its list of designated hateful ideologies.

Enforcement

4. To improve clarity to reviewers of its Dangerous Organizations and Individuals Community Standard, Meta should provide more global examples to reviewers of prohibited glorification, support and representation of hateful ideologies, including examples that do not directly name the listed ideology.

   The Board will consider this recommendation implemented when Meta provides updated internal guidance to the Board including more global examples, including ones that do not directly name the listed ideology.

***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.*