Criticism of EU Migration Policies and Immigrants

(2025-003-FB-UA, 2025-004-FB-UA)

Summary

The majority of the Board has found that two pieces of immigration-related content, posted on Facebook ahead of the June 2024 European Parliament elections, violate the Hateful Conduct policy and Meta should take them down. The Board recognizes the right to free expression is paramount when assessing political discussions and commentary. However, content such as these two posts contributed to heightened risks of violence and discrimination in the run-up to an election, in which immigration was a major political issue and anti-migrant sentiment was on the rise. For the majority, it is necessary and proportionate to remove them. One post by a Polish political party intentionally uses racist terminology to harness anti-migrant sentiment. The other post generalizes immigrants as gang rapists, a claim that, when repeated, whips up fear and hatred.

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 Case

The first case involves a meme posted on the official Facebook page of Poland's far-right political alliance, Confederation. In the meme, Polish Prime Minister Donald Tusk looks into a door's peephole, while a Black man walks up behind him. Polish text says:

“Good evening, did you vote for Platform? I’ve brought the murzyn from the immigration pact.” Platform is Tusk’s political party, the Civic Platform coalition, while the pact is the European Union’s Pact on Migration and Asylum. The Polish word, “murzyn,” used to describe Black people, is widely considered to be derogatory. The caption criticizes the EU pact and encourages people to vote for Confederation in the European elections to stop “uncontrolled immigration.” This content has been viewed around 170,000 times.

In the second case, a German Facebook page describing itself as against left-leaning groups posted an AI-generated image of a blonde-haired, blue-eyed woman holding up her hand in a stop gesture. German text says people shouldn’t come to the country anymore because no more “gang rape specialists” are needed due to the Green Party’s immigration policy. There is also a non-hyperlinked address for an article, titled “Non-German suspects in gang rapes,” on the German Parliament’s website. This post has been viewed around 9,000 times.

Both posts were reported for hate speech. Meta found no violations, leaving them on Facebook. Users then appealed the cases to the Board.

Key Findings

The majority of the Board finds that both posts violate the renamed Hateful Conduct policy, while a minority finds no violations in either.

The Polish post contains the word “murzyn,” which the majority considers to be a discriminatory slur, used to attack Black people based on race. Meta’s January 7 changes did not impact its rule on slurs, defined as “words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic.” Implying the inferiority or uncleanliness of Black people, the term’s offensive nature is recognized in the main Polish language dictionaries. It is also notable that Black-led and Polish-speaking civil society movements have played an important role in raising awareness of the term’s discriminatory and harmful impacts.

The majority notes that Meta does not currently include “murzyn” as a slur, recommending this be changed and calling on the company to more accurately enforce its slurs policy.

A minority of the Board disagrees, finding the term does not meet Meta’s definition of a slur and clearer evidence is needed that it inherently creates an atmosphere of exclusion and intimidation.

The majority also finds that the German post is violating because it contains a Tier 1 attack, generalizing that the majority of immigrants are “gang rape specialists.” This rule, which does not allow allegations of “serious immorality and criminality” based on immigration status, including by calling people “sexual predators,” remains unchanged since January 7. For this rule to apply, posts must target more than 50% of a group, with Meta’s internal guidance (not available publicly) advising that reviewers leave up content when it is unclear if this condition has been met. This is why Meta left up the German post. The majority of the Board disagrees with Meta’s assessment. It recommends the company change this rule to require users to clearly indicate they are targeting less than half of a group, for example, by using qualifiers such as “some.”

A minority of the Board disagrees, finding the German post does not state or imply that all or most immigrants are gang rapists.

Finally, the majority notes it is appropriate for Meta to consider the effects on human rights of such hateful conduct accumulating on its platforms. A minority disagrees with the majority, finding that removals would only have been justified if the posts constituted incitement to likely and imminent violence and discrimination. These two posts called for no action, other than participation in an election and discussion of public interest around immigration.

The Oversight Board’s Decision

The Oversight Board overturns Meta in both cases.

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 the rights of immigrants, in particular refugees and asylum seekers, with a focus on markets where these populations are at heightened risk. 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.
- Add the term “murzyn” to its Polish market slur list.
- Ensure it carries out broad external engagement with relevant stakeholders, including consulting with impacted groups and civil society, when auditing its slur lists.
- Update its internal guidance, making it clear that Tier 1 attacks (including those based on immigration status) are prohibited unless it is clear from the content that it refers to a defined subset of less than half the group. This would reverse the current presumption that content refers to a minority, unless it specifically states otherwise.

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

Full Case Decision

1. Case Description and Background

The Oversight Board has reviewed two cases involving content posted on Facebook ahead of the June 2024 European Parliament elections, in which immigration was a key issue. In May of that year, the European Union (EU)’s Pact on Migration and Asylum was adopted, establishing new rules to manage migration in Europe.

The first case involves a meme posted by an administrator of the official Facebook page of Poland's far-right political alliance, Confederation (Konfederacja Wolność i Niepodległość). The image shows the country's Prime Minister Donald Tusk looking into a door viewer (or peephole), as a Black man walks up behind him. Polish text over the image says: "Good evening, did you vote for Platform? I've brought the murzyn from the immigration pact." Platform refers to Tusk's centrist Civic Platform coalition, which came into power in December 2023. "Murzyn," the Polish word used to describe Black people in the text, is widely considered to be a derogatory slur in Poland, although Meta does not prohibit it. The caption criticizes the EU pact and encourages people to vote for Confederation in the European elections to stop immigrants being allowed into Poland and the EU. The post has been viewed around 170,000 times, shared less than 500 times and has under 500 comments.

In the second case, the administrator of a German Facebook page described as being against left-leaning groups posted an image that appears to be AI-generated. The image shows a blonde-haired, blue-eyed woman holding up her hand in a stop gesture, with both a stop sign and the German flag in the background. German text over the image says people should no longer come to Germany as they don't need any more "gang rape specialists," due to the Green Party's immigration policy. This is followed, in much smaller text, by a non-hyperlinked website address for an article on the German Parliament's website titled "Non-German suspects in gang rapes." The post has been viewed about 9,000 times and shared less than 500 times.

Ten Facebook users reported the Polish post and one reported the German post, all for hate speech. Meta left both posts on Facebook and, after each decision was unsuccessfully appealed to Meta, both cases were 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).

2. User Submissions

The user who appealed against the Polish post cited academic references to support their position that "murzyn" is a pejorative and derogatory term that perpetuates racial stereotypes and discrimination. The user who appealed against the German post noted that it appears to claim all refugees are criminals and rapists.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Hateful Conduct (previously named Hate Speech) Community Standard

Meta defines hateful conduct in the same way that it previously defined "hate speech," as "direct attacks against people" on the basis of protected characteristics, including national origin, race and ethnicity. The policy continues to treat immigration status as a "quasi-protected characteristic." This means Meta only protects immigrants from the most severe attacks under Tier 1 of the policy. On January 7, Meta added an explanation to the policy rationale that people sometimes "call for exclusion or use insulting language in the context of discussing political or religious topics," including immigration. Meta explicitly states that its "policies are designed to allow room for these types of speech."

Tier 1 prohibits "allegations of serious immorality and criminality," giving sexual predators and violent criminals as examples. The policy previously prohibited allegations about less serious forms of criminality, but this has been moved from Tier 1 to Tier 2. Tier 2 does not provide such protections to migrants – therefore, Meta now allows assertions that most migrants are, for example, thieves. Tier 2 continues to prohibit calls for exclusion but this protection also does not extend to migrants.

Tier 1 states that its prohibitions do not apply if content targets less than half of a group. Meta's internal guidance to moderators explains how to treat direct attacks that refer to less than 100% of a target group, including on the basis of immigration status. If the content contains a quantifier like "most" indicating it refers to more than 50% of the group, then Tier 1 prohibitions apply. If it is unclear whether the content refers to more than 50% of the group, then the content is permitted. Accordingly, content asserting that all or most migrants in a country are rapists or violent criminals is prohibited, but content asserting that some of them are rapists or violent criminals is allowed.

Tier 1 of the Hateful Conduct policy continues to prohibit "content that describes or negatively targets people with slurs." Slurs are defined as "words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression and violence." Meta sets out how it develops, enforces and updates its slur list on its Transparency Center.

II. Meta's Submissions

Meta left both posts up, finding neither violated its renamed Hateful Conduct policy. Meta confirmed the January 7 changes did not impact its decisions because its policies on racial slurs and generalizations comparing migrants to sexual predators or violent criminals have not changed.

Meta stated that the Polish post did not violate the policy because it does not contain a violating attack under Tier 1. Meta does not designate "murzyn" as a slur in the Polish market. Meta explained that the term was last considered for categorization in 2023, but was not added because Meta determined its use was historically neutral and, though it can be used contemptuously, its similarity to other words could lead to overenforcement.

Regarding the German post, Meta found the content not to be violating as "it is unclear whether the content is calling all, most, or some migrants gang rape specialists." For Meta, the content does not state or imply that all or most migrants will commit gang rape. Meta also noted that the article referred to in the post does not support the conclusion that it is attacking the majority of immigrants coming to Germany.

Finally, while Meta acknowledged both posts may be read as exclusionary, the company explained that neither violates its prohibition on "calls for exclusion" as Tier 2 prohibitions do not provide protections on the basis of immigration status.

The Board asked questions on Meta's Hateful Conduct policy, the company's slur lists, and how it assesses content from political parties and anti-migrant speech in the context of elections. Meta responded to all questions.

4. Public Comments

The Oversight Board received 18 public comments that met the terms for submission. Of these, 15 were submitted from Europe, two from the United States and one from Sub-Saharan Africa. 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: whether "murzyn" is a discriminatory slur; anti-immigrant rhetoric on social media; links between online hate speech and offline violence; the importance of being able to discuss immigration issues; and the rise of conspiracy theories in political rhetoric on migration issues.

5. Oversight Board Analysis

The Board selected these cases to examine how Meta ensures freedom of expression in discussions around immigration, while also respecting the human rights of migrants

in the context of an election. These cases fall within the Board's strategic priorities of Hate Speech Against Marginalized Groups and Elections and Civic Space.

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

The majority of the Board finds that both the Polish and German posts violate the Hateful Conduct policy and should be removed from Facebook. A minority finds no violations, however, in either post under the Hateful Conduct policy. The Board's outcome did not change as a result of Meta's January 7 changes.

The majority of the Board finds that "murzyn" is a discriminatory slur within the meaning of Meta's policy because it is used to attack Black people on the basis of their race, inherently creating an atmosphere of discriminatory exclusion and intimidation. The Board notes it is overwhelmingly used online as part of derogatory statements about Black people (also see public comments, including from the Institute for Strategic Dialogue – PC-30797, PC-30795 and PC-30790). Experts consulted by the Board explained the term is used in idioms and proverbs that imply the inferiority or uncleanliness of Black people, based on race. Black-led and Polish-speaking civil society movements in Poland played a key role in raising awareness of the term's discriminatory and harmful impacts. Harms, including perpetuating negative stereotypes and legitimizing discriminatory treatment by portraying Black people as the "other" within society, result from the term's derogatory nature and associations with inferiority. For the majority, it is especially compelling that a term is viewed as both derogatory and harmful by the marginalized group that it refers to. For this reason, Meta should be more systematic and thorough in its consultations with impacted groups when auditing its slur list, and more broadly when updating its policies.

The Board notes that contemporary understandings of the term matter. While some Polish speakers maintain that the term is neutral, the Polish Language Council issued guidance in 2020 that it is archaic, pejorative and should not be used in the public sphere. Experts also noted that while the term may have been perceived as neutral in the 20th century, it had negative and pejorative connotations prior to this. For example, the word was previously used to mean "a slave," tying it directly to one of history's worst examples of discrimination, oppression and violence, clearly meeting Meta's definition of a slur. The main Polish language dictionaries have now updated their definitions of the term to recognize it as offensive. For these reasons, the majority finds that use of the term creates an atmosphere of exclusion and intimidation. As such, the Board issues a recommendation to ensure Meta more accurately enforces its slurs policy moving forward. The majority also notes that, had the post not used this slur, it would have been permissible under Meta's content policies (see the Board's Armenians in Azerbaijan decision).

A minority of the Board disagrees that the Polish post is violating, finding the term does not meet Meta's definition of a slur. While the term may be seen as offensive and derogatory, this is insufficient to find that it should be considered a banned term. For the minority, Meta's policy requires clearer evidence that the use of the term *inherently* creates an atmosphere of exclusion and intimidation. There should be more than correlative ties to periods of historic discrimination, oppression and violence (in other times and places), but evidence that its use has been and continues to be intrinsic to the infliction of those harms.

The majority of the Board finds that the German post constitutes a Tier 1 attack by generalizing that the majority of immigrants are "gang rape specialists." This prohibition remains unchanged following Meta's January 7 policy changes.

For the majority, the characterization of immigrants entering the country as "gang rape specialists," without any qualifying language (e.g., "some" or "too many"), clearly conveys a generalized attack on all immigrants. Contrary to Meta's assessment, the fact that the post includes the website address (which is not hyperlinked and appears

in smaller text) of an article titled “Non-German suspects in gang rapes,” does not affect this conclusion. Instead, it supports the majority’s conclusion. The text in the post only includes the title of the article, which, rather than conveying the nuances discussed in the article’s fuller analysis, implies that “non-Germans” are generally the suspects of gang rapes.

For more accurate enforcement of the Hateful Conduct policy, the majority of the Board recommends Meta should reverse its default presumption that unless content clearly refers to more than 50% of a group, it will be considered non-violating (e.g., “immigrants are gang rapists” should be presumed as a generalization and therefore be violating). Meta should require users posting content that could violate the Hateful Conduct policy to clearly indicate they are targeting less than 50% of a group (e.g., “some immigrants are gang rapists”).

A minority of the Board finds that, while the German post is deeply offensive, it is not a generalization prohibited by Meta’s revised Hateful Conduct policy or the pre-January 7 version. The content does not state or imply that all or most immigrants are gang rapists. This group of Board Members is also concerned that the majority’s recommendation would place an undue burden on users having to explain their positions. The article referenced in the post, “Non-German suspects in gang rapes,” does not support a conclusion that the post is attacking the majority of immigrants, as it includes a nuanced discussion of why immigrants may be over-represented in official statistics on the perpetration of gang rapes. The minority notes that the post addresses a valid subject of discussion, especially in the context of an election where immigration, and in particular the relationship between migrants and crime, is a pivotal issue. Meta’s January 7 changes to the Hateful Conduct policy rationale make it clear the company intends its policies to provide more space for freedom of expression when discussing immigration.

5.2Compliance With Meta’s Human Rights Responsibilities

The majority of the Board finds that the removal of both posts, as required by a proper interpretation of Meta’s content policies, is also consistent with Meta’s human rights

responsibilities. A minority of the Board disagrees, finding that removal is not consistent with these responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including views about politics, public affairs and human rights (General Comment No. 34, paras. 11-12). 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 (UNGPs), 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. Under UNGPs Principle 13, companies should "avoid causing or contributing to adverse human rights impacts through their own activities" and "prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services." 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). At the same time, when company rules differ from international standards, companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation (ibid., at para 48).

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.*). 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 concludes there are no legality issues with the Hateful Conduct rules as applied to these cases. However, the Board is concerned that a recent version of this policy, following a December 2023 update, was being enforced for many months globally while only available in U.S. English, until the Board questioned Meta on this. Users accessing the Transparency Center from any other market would, by default, be accessing an outdated translation of the policy. The Board again encourages Meta to pay greater attention to ensuring its rules are accessible in all languages as swiftly as possible following any policy changes (see Punjabi Concern Over the RSS in India).

II. Legitimate Aim

Any restriction on freedom of expression should pursue one or more of the legitimate aims of the ICCPR, which include the "rights of others" (Article 19, para. 3, ICCPR). In several decisions, the Board has found that Meta's Hate Speech (renamed Hateful Conduct) policy aims to protect the right to equality and non-discrimination, a legitimate aim that is recognized by international human rights standards (see e.g., Knin Cartoon and Myanmar Bot). This continues to be the legitimate aim of the Hateful Conduct policy.

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 value of expression is particularly high when discussing matters of public concern and the right to free expression is paramount in the assessment of political discourse and commentary on public affairs. People have the right to seek, receive and impart ideas and opinions of all kinds, including those that may be controversial or deeply offensive (General Comment 34, para. 11). In the Politician's Comments on Demographic Changes decision, the Board found that, while controversial, the expression of this opinion on immigration did not include direct dehumanizing or hateful language towards vulnerable groups, or a call for violence.

The majority of the Board finds that removal of both posts is necessary and proportionate. This is guided by the six factors outlined in the Rabat Plan of Action in assessing risks posed by potential hate speech.

For the Polish post, the word "murzyn" is used generally and in this case to denigrate people on the basis of their race. The term's repeated use on Meta's platforms creates an environment in which discrimination and violence against Black people is more likely. Here, the slur is not used in a permissible context, either self-referentially in an empowering way, or to condemn or raise awareness of someone else's hate speech. For the majority, the cumulative effects of repeated use of this slur on Meta's platforms are comparable to the dehumanizing use of "blackface," as discussed by the majority in the Zwarte Piet case. It is much more obvious, however, in this post that the user is intentionally invoking racist terminology to harness anti-migrant sentiment by mobilizing anti-Black stereotypes (whereas in Zwarte Piet removal was justified without there being hostile intent).

For these reasons, the majority of the Board finds that removal of the post would be necessary regardless of when it was shared. It additionally notes that in this instance, in the run-up to an election with high levels of anti-migrant sentiment, there were heightened risks of violence and discrimination. Experts consulted by the Board

highlighted that vigilante groups in Poland have organized on social media to form "civic patrols," which target foreigners and people with foreign accents with offline violence and intimidation, including attacks on migrant accommodation. According to the OSCE, the police recorded 893 hate crimes in Poland in 2023, with racist and xenophobic motivation the highest recorded category. Research has also previously found that hate crimes in Poland were most often experienced by people of African descent. In this context, it is notable that the speaker is a political party with a sizable following and vote share in Poland. It has a broad reach (this post had around 170,000 views) and the ability to influence supporters to take action and attract media coverage. While it is of course important that a political party can freely campaign in an election, including by raising concerns about immigration, it can do this without using racial slurs (see Armenians in Azerbaijan).

The German post shares a similar context to the Polish post. It was also shared immediately before elections during which immigration was a major political issue, with high levels of anti-migrant sentiment present. Consistent with Meta's policies, the majority considers it necessary and proportionate to remove statements generalizing that the majority of immigrants as gang rapists. Crimes against migrants and anti-migrant online discourse were on the rise in Germany at the time. The United Nations High Commissioner for Human Rights has previously "expressed alarm at the often extraordinarily negative portrayal in many countries of migrants, but also of minority groups by the media, politicians and other actors in society [calling] for measures to curb growing xenophobic attitudes," (A/HRC/22/17/ADD.4, para. 3). Experts consulted by the Board noted that anti-immigrant rhetoric in Germany, often voiced and amplified on social media, may have contributed to attacks on immigrants and minorities (also see public comments PC-30803, PC-30797 and PC-30790). The 2024 riots in the United Kingdom also highlighted how social media content on topics like race and immigration can contribute to offline violence. The German post intentionally generalizes immigrants as sexual predators, a claim that repeated over again whips up fear and hatred, laying the foundations for inciting discrimination and violence against this group.

The users in both these cases could have contributed to the political debate without using racial slurs or engaging in degrading generalizations if Meta had given them notifications as to why their posts were potentially violating. Specificity in notifications when content is removed is important, but Meta should also explore increasing the use of prompts to invite users prior to posting to reconsider language that may potentially violate the company's policies. In the Pro-Navalny Protests in Russia case, the Board recommended that Meta notify users of the reason their content was violating, so they could repost without the violating part. In response to this recommendation, Meta has introduced notifications to users that their posts might be violating, giving them the opportunity to delete and repost content before any enforcement action is taken. Meta shared that over, a 12-week period in 2023, users opted to delete their post more than 20% of the time, decreasing the amount of violating content through self-remediation.

The majority emphasizes that in reaching its decisions on both posts, the standards for content moderation by a social media company should not be compared so directly to the standards limiting states' application of punitive law. Meta is not engaged in an after-the-fact detailed investigation of whether a crime was committed but is operating in real-time with incomplete information. Were it to wait until violence or discrimination is imminent before acting, it would be too late for it to prevent harm in accordance with its responsibilities under the UNGPs. Both the challenge of assessing the impact of each piece of content at scale and the unpredictable nature of online virality justify Meta taking a more cautious approach to moderation.

The majority reiterates that Meta as a private actor may remove hate speech that falls short of the threshold of incitement to imminent discrimination or violence, where this meets the ICCPR Article 19(3) requirements of necessity and proportionality (see South Africa Slurs). Meta allowing all hate speech that falls short of incitement as foreseen under Article 20 of the ICCPR would make Meta's platforms an intolerable and unsafe place for minorities and marginalized groups to express themselves. In these cases, it may cause not only migrants but anyone who is not white to withdraw

from public discourse, having a chilling effect that diminishes the value of pluralism and access to information for all people. It is therefore appropriate that Meta's approach to content moderation considers the effects on human rights of hateful content accumulating on its platforms, even when in isolation those posts do not incite imminent violence or discrimination (see Depiction of Zwarte Piet, Communal Violence in Indian State of Odisha, Armenians in Azerbaijan and Knin Cartoon).

The majority notes that less severe interventions, such as labels, warning screens or other measures to reduce dissemination, would not provide adequate protection against the cumulative effects of leaving content of this nature on the platform (see Depiction of Zwarte Piet and Knin Cartoon).

A minority of the Board finds that the removal of neither the Polish or German post is necessary and proportionate. They note that both posts may be offensive, but neither reaches the threshold of incitement to likely and imminent acts of violence, discrimination or hostility. For the minority, the concept of cumulative harms is not based on principles flowing from international freedom of expression standards. Rather, it is so elastic as to depart from requirements of basic causation, emptying the necessity and proportionality evaluation of substance. Compared to using the Rabat Plan of Action in a strict sense to assess the necessity and proportionality of content removal based on whether speech poses the likelihood of imminent harm, the cumulative harms concept essentially abandons this key factor. With respect to these posts, it is significant that neither called for action other than participation in an election, and/or a discussion of public interest matters around immigration. It is essential that users are able to express their opinions on the most pressing political issues facing their countries, including immigration. The minority notes that a wide array of content moderation tools are available to Meta beyond the binary "leave up/take down" choice, with less intrusive means than removals available to mitigate potential harms. When faced with the binary up/down choice, a minority would accord more weight to the importance of the electorate having full access to the views of political candidates and parties in the context of an election, and the heightened risks to expression that private censorship can have on democratic processes.

Perceptions of unfairness and bias in the moderation of political views threaten the legitimacy of platform governance more broadly. Meta should take inspiration from the Rabat Plan, which also has a focus on positive policy measures, to consider less intrusive means than censorship to ensure potential harms are averted.

Access to Remedy

The users who reported these posts were not informed that those reports (or appeals) were not prioritized for review. The Board reiterates concerns raised previously (see Explicit AI Images of Female Public Figures) that users may be unaware that their report or appeal was not prioritized for review. Given Meta's January 7 announcement that it now plans to focus automated systems on tackling "illegal and high-severity violations," and rely more on user reports for "less severe" policy violations, the demands on reviewing user reports may increase. It will be crucial that Meta is able to accurately prioritize and actually review the volume of reports it receives so that its policies are fairly enforced. When user reports are not prioritized for review, users should be informed that no review has taken place.

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 groups may be differently impacted, including immigrants, refugees and asylum seekers. 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).

6. The Oversight Board's Decision

The Oversight Board overturns Meta's decisions to leave up the content in both cases.

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, 2025, updates to the Hateful Conduct Community Standard. First, it should identify how the policy and enforcement updates may adversely impact the rights of immigrants, in particular refugees and asylum seekers, with a focus on markets where these populations are at heightened risk. 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.

Enforcement

2. Meta should add the term “murzyn” to its Polish market slur list.

   The Board will consider this recommendation implemented when Meta informs the Board this has been done.

3. When Meta audits its slur lists, it should ensure it carries out broad external engagement with relevant stakeholders. This should include consulting with impacted groups and civil society.

   The Board will consider this recommendation implemented when Meta amends its explanation of how it audits and updates its market-specific slur lists on its Transparency Center.

4. To reduce instances of content that violates its Hateful Conduct policy, Meta should update its internal guidance to make it clear that Tier 1 attacks (including those based on immigration status) are prohibited, unless it is clear from the content that it refers to a defined subset of less than half of the group. This would reverse the current presumption that content refers to a minority unless it specifically states otherwise.

   The Board will consider this recommendation implemented when Meta provides the Board with the updated internal rules.

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