Symbols Adopted by Dangerous Organizations

2025-015-IG-MR, 2025-016-IG-MR, 2025-017-IG-MR

Summary

The Oversight Board has considered three cases involving symbols often used by hate groups, but which can also have other uses. The Board calls on Meta to explain how it creates and enforces its designated symbols list under its Dangerous Organizations and Individuals Community Standard. This will provide greater transparency for users.

The Board is concerned about the potential overenforcement of references to designated symbols. Meta should develop a system to automatically flag when non-violating content is being removed in large volume.

The Board upholds Meta's decisions to remove violating content in two of the cases and to leave the content up in the third case.

About the Cases

Meta referred to the Board three Instagram posts involving symbols often used by hate groups, but which can have other uses. The first post involved an image showing a woman, with the words “Slavic Army” and a Kolovrat symbol superimposed on her face covering. In the caption, the user expressed pride in being Slavic and hoped their “people will wake up.”

The second post was a carousel of photographs of a woman wearing an iron cross necklace with a swastika on it and a T-shirt with an AK-47 assault rifle and “Defend Europe,” written in Fraktur font, printed on it. The caption contained the Odal (or Othala) rune, the hashtag #DefendEurope, symbols outlining an M8 rifle and heart emojis.

On referring the two posts to the Board, Meta removed them because they violated the Dangerous Organizations and Individuals policy.

The third post is a carousel of drawings of an Odal rune wrapped around a sword with a quotation about blood and fate by Ernst Jünger, a German author, philosopher and soldier. The caption repeats the quotation, shares a selective early history of the rune and states that prints of the image are for sale. Meta concluded that this post does not breach any of its rules.

Key Findings

The majority of the Board finds that the Kolovrat symbol post glorified white nationalism. A minority disputes any automatic link between Slavic pride and white nationalism. The Board finds the defend Europe post glorified white supremacy. The two posts should be taken down for violating the Dangerous Organizations and Individuals policy. Under the policy rationale, Meta removes content that glorifies, supports or represents ideologies that promote hate. It designates white nationalism and white supremacy as hateful ideologies.

The quotation post does not violate the same policy. It describes the Odal rune in a seemingly neutral manner. There is no glorification of any hateful ideology in the quotation. The post does not reference Nazism or any other designated hateful ideology specifically.

Meta's decisions were consistent with its human rights responsibilities. For the majority, the Kolovrat symbol post's contextual cues, including clear references to Slavic nationalism and Slavic army, may be read as urging followers to take potentially violent action, and it should be removed. A minority disagrees, considering that the post did not pose a direct risk of inciting imminent or likely harm.

The Board considers the removal of the defend Europe post necessary and proportionate to prevent risk of immediate discrimination. It contains multiple

contextual cues glorifying designated hateful ideologies. Its removal is necessary and proportionate to the legitimate aim of preventing Meta's platforms from being abused to organize and incite violence or exclusion.

Leaving up the quotation post was justified, as the content does not reference a designated hateful ideology and provides more context around the user's artwork.

The Board reiterates concerns about the lack of transparency around designation processes under Tier 1 of the Dangerous Organizations and Individuals policy, as this makes it challenging for users to understand which entities, ideologies and related symbols they can share. Meta should provide more transparency around designated symbols, especially those associated with designated hate entities or ideologies, establishing an evidence-based, global and iterative process. It should publish a clear explanation of its processes and criteria for designating the symbols and enforcement against them.

The Board is concerned about potential overenforcement of references to designated symbols. Meta does not collect sufficiently granular data on its enforcement practices in this area. Meta told the Board that its internal definition of a "reference" is broader than the definition of "unclear reference" in its public-facing policy. Meta should publicly provide the internal definition of "references" and define its subcategories, for clarity and accessibility for users.

The Oversight Board's Decision

The Oversight Board upholds Meta's decisions to take down the content in the first and second cases and to leave up the content in the third case.

The Board recommends that Meta:

- Make public the internal definition of "references" and define its subcategories under the Dangerous Organizations and Individuals Community Standard.

- Introduce a process to determine how designated symbols are added to the groups and which group each designated symbol is added to, and periodically audit all designated symbols, ensuring the list covers all relevant symbols globally and removing those no longer satisfying published criteria.
- Develop a system to automatically identify and flag instances where designated symbols lead to "spikes" that suggest a large volume of non-violating content is being removed.
- Publish a clear explanation of how it creates and enforces its designated symbols list under the Dangerous Organizations and Individuals Community Standard.

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

Full Case Decision

1. Case Description and Background

In November 2024, Meta referred three Instagram posts to the Board all involving symbols often used by hate groups, but which can also have other uses.

The first post, from April 2016, involved an image showing a blonde woman with the bottom half of her face covered by a scarf. The words "Slavic Army" and a Kolovrat symbol were superimposed over the face covering. The Kolovrat is a type of swastika symbol that is used by neo-Nazis and some neo-pagans. In the caption, the user expressed pride in being Slavic, stating the Kolovrat is a symbol of faith, war, peace, hate and love. The user hoped their "people will wake up" and said they would follow "their dreams to the death." The post was viewed under 100 times and received under 500 reactions and under 50 comments. When Meta selected this content for referral to the Board, the company's policy subject matter experts determined the post violated the Dangerous Organizations and Individuals policy and removed it.

The second post, from October 2024, was a carousel of selfie photographs. The photos showed a blonde woman in various poses. She is wearing an iron cross necklace with a swastika on it and a T-shirt with an AK-47 assault rifle and "Defend Europe," written in Fraktur font, printed on it. The caption contained the Odal (or Othala) rune, the hashtag #DefendEurope, symbols outlining an M8 rifle, a flexed bicep emoji and heart emojis. The Odal rune is a symbol of the runic alphabet used across many parts of Europe until it was replaced by the Latin alphabet in the seventh century. It was appropriated by the Nazis and is now used by neo-Nazis and other white supremacists to represent ideas connected to what they describe as the "Aryan race." The post was viewed about 3,000 times and received under 500 reactions and under 50 comments. After the company identified this case to refer to the Board, Meta's policy subject matter experts also determined that the post violated the Dangerous Organizations and Individuals policy and removed it from the platform.

The third post also concerns a carousel of images. Posted in February 2024, the images are drawings of an Odal rune wrapped around a sword with a quotation about the relationship between blood and fate by Ernst Jünger, a German author, philosopher and soldier who fought in the First and Second World Wars. Jünger was a German nationalist and a critic of the Nazi party. The caption repeats the quotation before sharing a selective early history of the rune, without mentioning its Nazi and neo-Nazi appropriation. The caption concludes by describing the rune as a symbol of "heritage, homeland, and family" and stating that prints of the image are for sale. The post has been viewed about 25,000 times and has received under 1,000 reactions and under 50 comments. After Meta selected this content to be referred to the Board, the company's policy subject matter experts concluded that this third post does not breach any of its rules.

2. User Submissions

The authors of the posts were notified of the Board's review and provided with an opportunity to submit a statement. None of the users submitted a statement.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

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 ideologies that promote hate," also described in the policy as hateful ideologies. Hateful ideologies are considered part of Tier 1 of Meta's policy, which result in the most extensive enforcement as Meta believes Tier 1 entities and individuals have "the most direct ties to offline harm."

Meta explains it designates prohibited ideologies, which the policy lists as "includ[ing] 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."

Meta removes glorification for Tier 1 entities and ideologies. Among other things, glorification includes, "legitimizing or defending the violent or hateful acts of a designated entity by claiming that those acts have a moral, political, logical or other justification that makes them acceptable or reasonable," or "characterizing or celebrating the violence or hate of a designated entity as an achievement or accomplishment."

Under Tier 1 of the policy, Meta also removes "unclear references," that "include unclear humor, captionless or positive references, that do not glorify the designated entity's violence or hate." Meta twice states it removes "unclear references" to hateful ideologies - once in the policy rationale and again under the description of Tier 1 organizations.

The policy allows users to report on, neutrally discuss or condemn designated organizations or individuals or their activities in the context of "social and political discourse." According to the policy rationale, Meta requires users to "clearly indicate their intent" when creating or sharing such content. If the user intent is "ambiguous or unclear," Meta defaults to removing content.

II. Meta's Submissions

Meta determined that the first (Kolovrat symbol post) and second (defend Europe post) cases violated the Dangerous Organizations and Individuals policy, while the third case (quotation post) does not violate this policy.

In response to questions from the Board, Meta explained that symbols associated with Tier 1 entities or ideologies are sorted into three groups. These groups determine how the company reviews the posts and the enforcement action that can be taken. The first group consists of a very short list of symbols that have gained "public notoriety" as well-known aliases of designated entities and are "routinely and heavily" used by them. Symbols in the first group are treated as inherently violating the Dangerous Organizations and Individuals policy when moderated at scale.

The second, much larger group is made up of symbols which are mostly used in the context of a designated entity or ideology. Meta treats symbols in this second group as violating at scale. The Kolovrat symbol in the first case falls into this second group.

The third group consists of a very short list of symbols associated with designated entities but also commonly used in benign contexts. Meta assesses these symbols on escalation to avoid risks of overenforcement and considers them violating only when context suggests the symbol is being used to refer to a designated ideology. This group includes the Odal rune at issue in the second and third cases.

With respect to the enforcement against the symbols in all three groups, if they are shared to glorify, support or represent a designated entity or ideology Meta removes

the post and applies a strike to the user's account. Meta also removes posts and applies a strike where a symbol from the first group is posted as the primary focus of an image without context or a caption. When an image from the first group is posted without context or a caption but is not the primary focus of an image the content is removed as an unclear reference but without a strike being given. When an image from the second group is shared without context or caption, regardless of whether it is the primary or secondary focus of the content, it will be removed without a strike as an unclear reference. When an image from the third group is posted without any context or caption the content is also removed as an unclear reference without a strike.

The company noted that the Kolovrat symbol is removed as violating at scale unless used in an allowable context such as social and political discussion. Meta found that the Kolovrat symbol post contained a reference to white nationalism, a designated hateful ideology. Meta viewed the phrase "Slavic pride" as "intended to convey the racist chauvinism of white nationalism." The company concluded that given the content's clear references to Slavic nationalism and the Slavic army, and its militaristic mentions, the symbol was used to reference neo-Nazi ideology stemming from beliefs in Slavic supremacy.

Meta allows the Odal rune unless, when reviewed on escalation, the company concludes that the symbol is used in a hateful context. The company concluded that in the defend Europe post, the Odal rune was shared with context that established that the rune was being used to glorify white supremacy. Meta said that several signals helped its teams determine that the symbol was used to celebrate the violence of a hateful ideology. This included the use of "#DefendEurope," a hashtag often used by European white supremacist groups; the phrase "Defend Europe" written in Fraktur font and the iron cross around the woman's neck are associated with Nazi materials and propaganda; and the symbols constructing an M8 rifle in the caption with heart and gun emojis surrounding the Odal rune.

Meta determined that the quotation post does not violate the Dangerous Organizations and Individuals policy. The company noted that the Odal rune was accompanied by a

neutral description of the historical origins and linguistic significance of the rune. While the assessment of the Jünger quotation was more challenging, Meta ultimately concluded that there was no clear indication that the Odal rune was used to refer to Nazism or white supremacy.

In response to questions from the Board, Meta explained that it has not completed a systemic audit of either its first or third list of symbols. The company continuously updates the second list. Meta added that the company is exploring a further policy development in this area that will help inform how it may approach future audits.

The Board asked questions on the designation and enforcement of designated symbols. Meta responded to all questions.

4. Public Comments

The Oversight Board received 100 public comments that met the terms for submission. Of these comments, 95 were submitted from the United States and Canada, two from Central and South Asia, two from the Middle East and North Africa, and one from Europe. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: approaches to moderating potentially hateful symbols; the importance of allowing users to provide further context on content that is potentially violating; the need to work with external subject experts and trusted partners; and proposals to focus AI training on analysis of signal combinations.

5. Oversight Board Analysis

The Board selected these cases to consider how Meta should moderate symbols that may promote dangerous organizations while respecting users' freedom of expression. These cases fall within the Board's strategic priority of Hate Speech Against Marginalized Groups.

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

5.1 Compliance With Meta's Content Policies

The Board agrees with Meta that the Kolovrat symbol and defend Europe posts violated Meta's Dangerous Organizations and Individuals policy, while the quotation post does not.

While the Board agrees with Meta that the Kolovrat symbol post should be taken down, it takes that position for a different reason. Meta finds the Kolovrat symbol post violating as a "reference" to a hateful ideology. In contrast, a majority of the Board considers that it glorified white nationalism, a designated hateful ideology. The references to Slavic pride and the Slavic army, together with expressions of hope that their "people will wake up" and stating that they would follow "their dreams to the death" illustrate the user's intent to glorify under Tier 1 of the policy, particularly to, as the policy delineates, "legitimize or defend the violent or hateful acts ... by claiming that those acts have a moral, or political justification that makes them acceptable." Since the "social and political discourse" policy exception requires that the post does not contain glorification, this post did not qualify for the policy exception.

A minority of the Board does not find the first post violating as either an unclear reference or glorification of a hateful ideology. This minority disputes any automatic link between Slavic pride, a term that also has cultural and historical connotations, and white nationalism. The post could have also benefitted from the "social and political discourse" policy exception. Its removal is indicative of how relying on the assumption of "unclear references" in the policy could lead to unnecessary overenforcement. The minority also urges Meta to periodically review the accuracy and precision of the impact of this policy line on overall removals and whether it needs narrowing and adjusting.

The Board finds that the defend Europe post celebrated the violence of white supremacy, constituting glorification of a designated hateful ideology. The Board's conclusion is supported by the following contextual signals in the post. While each of these signals may not individually constitute a violation, together they are explicit glorification:

- The iron cross with a swastika on the woman's neck. The iron cross is a German military medal, that the Nazis appropriated by adding a swastika on it. While the medal was discontinued after the Second World War, neo-Nazis and white supremacy groups have adopted it as a hate symbol.
- The use of the phrase "Defend Europe" on the t-shirt – written in Fraktur font – and in the caption. "Defend Europe" is a slogan used by white supremacists and other anti-migrant organizations connected to acts of violence. The Fraktur font is a typeface sometimes associated with Nazis and neo-Nazis.
- The Odal rune being accompanied by symbols constructing an M8 rifle, a flexed bicep and heart emojis in the caption.

Since the post contained glorification, the Board concluded that the "social and political discourse" policy exception was not applicable to this post.

The Board agrees with Meta that the quotation post does not violate the Dangerous Organizations and Individuals policy. The post describes the Odal rune in a seemingly neutral manner, without any mention of its appropriation by Nazis, and intends to sell artwork involving the Odal rune. Although the post contains a quote by Ernst Jünger, a German nationalist, the references to fate and blood in the quote itself do not constitute glorification, support or representation of any designated hateful ideology. More generally, the post does not reference Nazism or any other designated hateful ideology specifically. It appears to provide context around the artwork in the post, which includes the Odal rune. While the sword may be viewed as a symbol of violence, a single indicator is not sufficient to constitute a violation.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that Meta's decision to remove the content in the first and second cases but keep the third post on the platform was consistent with Meta's human rights responsibilities. A minority of Board Members disagree with the removal of the first post.

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 "political discourse," "commentary on public affairs" and expression that may be considered "deeply offensive" (General Comment No. 34, 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 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 reiterates its concerns about the lack of transparency around designation processes under Tier 1 of the Dangerous Organizations and Individuals policy. This makes it challenging for users to understand which entities, ideologies and related symbols they can share (see Greek 2023 Election Campaign). The Board has urged Meta to publish the Tier 1 list (see Nazi Quote). Meta declined to publish the list, arguing that such a publication could allow "bad actors to circumvent the enforcement mechanisms" and affect the "safety of [Meta's] employees." Meta has committed to hyperlink the U.S. Foreign Terrorist Organizations and Specially Designated Global Terrorists lists in its Community Standards, where these lists are mentioned, in response to recommendation no 3 in Sudan's Rapid Support Forces Video Captive. However, the list of hate entities is not based on an equivalent public list, making it challenging for people to deduce which related hate symbols are prohibited.

The Board urges Meta to provide more transparency around designated symbols, especially those associated with designated hate entities or ideologies. The Board calls on Meta to introduce a clear process to determine how symbols are added to the three groups and which group each symbol is added to. The Board considers that this process, at a minimum, should be evidence-based, global in scope and iterative in nature. First, this process will help ensure the symbols list is up-to-date, and the symbols that do not meet the inclusion criteria are removed, while the remaining symbols are enforced through the most applicable assessment process. For example, all designated symbols with multiple non-violative purposes are assessed through an escalated review from Meta's internal subject matter teams or specific guidance is issued to human reviewers. In determining the dominant use cases for specific symbols, Meta should rely on relevant research findings, that may include research into symbol

usage trends on the company's platforms across languages and regions. Next, in line with Meta's commitments to develop and enforce its global rules in a non-discriminatory manner (Article 2 and 26, ICCPR; General Comment No. 34, para. 26), Meta should review the list of designated symbols. In doing so, Meta should ensure that the list covers, for example, hateful ideologies not only in Global Minority but also Global Majority regions, as the Board noted in the Posts Displaying South Africa's Apartheid-Era Flag decision. The Board also notes that the use of symbols may change over time, and, therefore, the list should be audited periodically, to address the risks of potential under or overenforcement, based on evolving uses.

To provide more transparency to users, Meta should also publish a clear explanation of how it creates and enforces its designated symbols list. This explanation should include the processes and criteria for designating the symbols and how the company enforces against different symbols, including the application of strikes. Meta should publish this information in the Transparency Center and hyperlink to it in the public-facing language of the Dangerous Organizations and Individuals policy.

The Board is also concerned about the lack of clarity and potential overbreadth of the "references" policy line under which Meta removed the first post. In response to the Board's questions, the company disclosed that its internal definition of a "reference" is broader than the definition of "unclear reference" in its public-facing Dangerous Organizations and Individuals policy. A reference "includes but is not limited to positive references, incidental depictions, captionless photos, unclear satire or humor, and symbols." Under the "we remove" section of the policy, Meta only states that it removes "unclear or contextless references if the user's intent was not clearly indicated." This "includes unclear humor, captionless or positive references that do not glorify the designated entity's violence or hate." Therefore, the Board calls on Meta to publicly provide the internal definition of "references" and to define its subcategories, such as "positive references" and "incidental depiction," to ensure this policy line is sufficiently clear and accessible to users.

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 considered that 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 (Article 6, ICCPR) and the right to non-discrimination and equality (Articles 2 and 26, ICCPR) because it covers organizations that promote hate, violence and discrimination as well as designated violent events motivated by hate (see Sudan's Rapid Support Forces Video Captive, Greek 2023 Elections Campaign).

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).

With regards to state obligations, 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).

A majority of the Board finds the removal of the Kolovrat symbol post to be the necessary and proportionate response to prevent Meta's platforms from being exploited to incite or organize discrimination and violence. Given the contextual cues in the post, including clear references to Slavic nationalism and Slavic army, the post

may be read to urge followers to take potentially violent action. For this majority, the post shows how symbols can be part of online efforts to create and cultivate support for ideologies that further violence or exclusion. A minority disagrees, considering that the post did not pose a direct risk of inciting imminent or likely violence or discrimination. Therefore, Meta could have resorted to other tools, including not recommending the post or de-amplifying it in general. Removal of this post, according to this minority, is neither necessary nor a proportionate measure.

The Board considers that the removal of defend Europe post satisfies the principles of necessity and proportionality, to prevent likely and imminent discrimination and violence. Unlike the content in the Posts Displaying South Africa's Apartheid-Era Flag decision, the post contains multiple contextual cues that are more directly connected to glorification of violence, as described in section 5.1. This post illustrates how hate symbols can become rallying points for networked actors seeking to build connections and recruit like-minded individuals while evading content moderation. Therefore, the removal of the defend Europe post is also necessary and proportionate to the legitimate aim of preventing Meta's platforms from being abused to organize and incite violence or exclusion.

The Board finds Meta's decision to leave up the quotation post was justified, as the content does not reference a designated hateful ideology and provides more context around the user's artwork. Therefore, removal was not warranted.

The Board is also concerned about potential overenforcement of references to designated symbols under the Dangerous Organizations and Individuals Community Standard. Meta does not collect sufficiently granular data on its enforcement practices in this area. Meta should develop a system to automatically identify and flag instances where designated symbols lead to "spikes" that suggest a large volume of non-violating content is being removed, similar to the system the company created in response to the Board's recommendation no. 2 in Colombian Police Cartoon. This system will allow Meta to analyze "spikes" involving designated symbols and inform the company's future actions, including amending their practices to be more accurate and precise. For

example, Meta may consider policy changes or adjustment of enforcement measures if there is a large volume of false positives that result in incorrect removals of references to designated symbols, where the content is in fact non-violating or should fall under the "social and political discourse" policy exception. The Board expects Meta to develop this system and inform the Board of the actions taken to avoid potential overenforcement detected by the system. The Board also anticipates reviewing the actions taken in a future case.

6. The Oversight Board's Decision

The Oversight Board upholds Meta's decisions to take down the content in the first and second cases and to leave up the content in the third case.

7. Recommendations

Content Policy

1. To provide more clarity to users, Meta should make public the internal definition of "references" and define its subcategories under the Dangerous Organizations and Individuals Community Standard.

   The Board will consider this recommendation implemented when Meta updates the public-facing Dangerous Organizations and Individuals Community Standard.

Enforcement

2. To ensure that the list of designated symbols under the Dangerous Organizations and Individuals policy does not include symbols that no longer meet Meta's criteria for inclusion, Meta should introduce a clear and evidence-based process to determine how symbols are added to the groups and which group each

designated symbol is added to, and periodically audit all designated symbols, ensuring the list covers all relevant symbols globally and removing those no longer satisfying published criteria, as outlined in section 5.2 of this decision.

The Board will consider this recommendation implemented when Meta has established this process and provides the Board with the documentation and the results of its first audit based on these new rules.

3. To address potential false positives involving designated symbols under the Dangerous Organizations and Individuals Community, Meta should develop a system to automatically identify and flag instances where designated symbols lead to "spikes" that suggest a large volume of non-violating content is being removed, similar to the system the company created in response to the Board's recommendation no. 2 in Colombian Police Cartoon. This system will allow Meta to analyze "spikes" involving designated symbols and inform the company's future actions, including amending their practices to be more accurate and precise.

   The Board will consider this recommendation implemented when Meta develops this system and informs the Board of the actions taken to avoid potential overenforcement detected by the system.

Transparency

4. To provide more transparency to users, Meta should publish a clear explanation on how it creates and enforces its designated symbols list under the Dangerous Organizations and Individuals Community Standard. This explanation should include the processes and criteria for designating the symbols and how the company enforces against different symbols, including information on strikes and any other enforcement actions taken against designated symbols.

The Board will consider this recommendation implemented when the information is published in the Transparency Center and is hyperlinked in the public-facing Dangerous Organizations and Individuals Community Standard.

***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.