Posts Supporting UK Riots

(2025-009-FB-UA, 2025-010-FB-UA, 2025-011-FB-UA)

Summary

In reviewing three different posts shared during the UK riots of summer 2024, the Board has overturned Meta's original decisions to leave them up on Facebook. Each created the risk of likely and imminent harm. They should have been taken down. The content was posted during a period of contagious anger and growing violence, fueled by misinformation and disinformation on social media. Anti-Muslim and anti-immigrant sentiment spilled onto the streets. Meta activated the Crisis Policy Protocol (CPP) in response to the riots and subsequently identified the UK as a High-Risk Location on August 6. These actions were too late. By this time, all three pieces of content had been posted. The Board is concerned about Meta being too slow to deploy crisis measures, noting this should have happened promptly to interrupt the amplification of harmful content.

Additional Note: Meta's January 7, 2025, revisions to the renamed Hateful Conduct policy 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 any adverse impacts on human rights globally are identified and prevented.

About the Cases

In the first case, a text-only post shared at the start of the riots, called for mosques to be smashed and buildings where “migrants,” “terrorists” and “scum” live to be set on fire. This post had more than 1,000 views.

The second and third cases both involve reposts of likely AI-generated images. One is of a giant man in a Union Jack T-shirt chasing smaller Muslim men in a menacing way. Text over the image gives a time and place to gather for one of the protests and includes the “EnoughIsEnough” hashtag, while the accompanying caption says: “Here we go again.” This post had fewer than 1,000 views. The other image is of four Muslim men running in front of the Houses of Parliament after a crying blond-haired toddler. One of the men waves a knife while a plane flies overhead towards Big Ben. This image includes the logo of an influential social media account known for anti-immigrant commentary in Europe, including misinformation and disinformation. This had more than 1,000 views.

All three were reported by other Facebook users for either hate speech or violence. Meta kept all three up following reviews by its automated systems only. After the users appealed to the Board and these cases were selected, the content was reviewed by humans, with Meta removing the text-only post in the first case. The company confirmed the original decisions to keep up the two likely AI-generated images.

Between July 30 and August 7, 2024, violent riots broke out in the UK after three girls were murdered in the town of Southport. Shortly after this knife attack, misinformation and disinformation spread on social media falsely suggesting the perpetrator was a Muslim and an asylum seeker.

Key Findings

The Board has found that the text-based post and giant man image both violate the Violence and Incitement policy, which does not allow threats of high-severity violence against a target, or threats of violence against individuals or groups based on protected characteristics and immigration status. The text-based post contains a general threat and incitement of violence against people and property, as well as identifying targets based on religion and immigration status. The giant man image is a

clear call for people to gather and carry out acts of discriminatory violence at a particular time and place. Meta's conclusion that this image – an aggressive man chasing fleeing Muslim men, combined with a time and place and the "EnoughIsEnough" hashtag – contains no target or a threat, strains credibility. This content was shared on August 4, well into the week-long riots. By this time, there was more than enough context to warrant removal.

The AI image of four Muslim men pursuing a crying, blond-haired toddler broke the rule under the Hateful Conduct (previously named Hate Speech) policy against attacking people based on their protected characteristics, including by making an allegation of serious criminality. Meta interpreted this post as being a qualified statement in visual form by referring to the specific "Muslim man or men who were incorrectly accused of stabbing the children in Southport." Before January 7, Meta's internal guidance stated qualified statements that avoid generalizing all members of a group as criminals were allowed. The Board disagrees with Meta's application of the rule in this case, noting the image does not represent a qualified statement as it does not depict the Southport stabbing in any form. It is set in London (not Southport), with four men (not one) running after a male toddler (not three young girls), and a plane flying towards Big Ben, the latter evoking 9/11 imagery and portraying Muslims as a threat to Britain.

When reviewing these cases, the Board noted issues of clarity around both the Violence and Incitement and Hateful Conduct policies, caused by discrepancies between public-facing language and internal guidelines. The Board also has strong concerns about Meta's ability to accurately moderate hateful and violent imagery. Given Meta's experts failed to identify violations in both of the likely AI-generated images, this would indicate that current guidance to reviewers is too formulaic, ignores how visual imagery works and is outdated.

Finally, the Board notes that Meta had third-party fact-checkers reviewing certain pieces of content containing the false name of the Southport perpetrator during the riots, labelling them as "false" and reducing their visibility. With Meta replacing its third-party fact-checking system in the United States, the Board recommends the

company examine the experience of other platforms using Community Notes and research their effectiveness.

The Oversight Board's Decision

The Board overturns Meta's original decisions to leave up the three posts.

The Board also recommends Meta:

- Specify that all high-severity threats of violence against places are prohibited, as well as against people.
- Develop clear and robust criteria for what may constitute allegations of serious criminality, based on protected characteristics, in visual form. They should align with and adapt existing standards for text-based hateful conduct.
- Revise criteria for initiating the Crisis Policy Protocol, including identifying core criteria that, when met, are sufficient for the immediate activation of the protocol.
- Under the Crisis Policy Protocol, ensure potential policy violations that could lead to likely and imminent violence are flagged for in-house human reviewers who should provide time-bound, context-informed guidance for reviewers at-scale.
- Undertake continuous assessments of the effectiveness of Community Notes, as compared to third-party fact-checking, particularly relevant to situations where the rapid dissemination of false information creates risks to public safety.

*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 three cases involving content posted by different users on Facebook during riots in the UK between July 30 and August 7, 2024.

The riots followed a knife attack at a dance workshop in Southport on July 29 in which three young girls were killed and ten others injured. Axel Rudakubana, a British 17-year-old, was immediately arrested and later convicted for the attack. Yet, misinformation and disinformation about his identity, including a false name, rapidly circulated online after the attack, wrongly asserting that he was a Muslim and an asylum seeker who had recently arrived in Britain by boat. One such post was shared more than six million times. Notwithstanding a police statement at noon on July 30 disputing the online rumors, anti-immigration and anti-Muslim protests took place across 28 cities and towns, with many turning into riots. They mobilized thousands of people, including anti-Muslim and anti-immigration groups. Refugee centers and hotels housing immigrants were among many buildings attacked or set on fire, alongside looting and other disorder. The violence led to many people, including more than 100 police officers, being injured. On August 1, a judicial order lifted the Southport attacker's anonymity as a minor to quell the disorder but it was not immediately successful.

The first post under the Board's review was shared two days after the killings. It supported the ongoing riots, calling for mosques to be smashed and buildings where "migrants," "terrorists" and "scum" are living to be set on fire. The post acknowledged the riots had damaged private property and injured police officers, but argued this violence was necessary for the authorities to listen and put a stop to "all the scum coming into Britain." The post asked those who disagreed with the riots to think about the murder of the "little girls," stating they would not be "the last victims" if the public did not do something. The post had more than 1,000 views and fewer than 50 comments.

The second post was shared six days after the attack and is a reshare of another post. It contains what looks like an AI-generated image of a giant, angry and aggressive white man wearing a Union Jack (the UK flag) T-shirt menacingly chasing several smaller, fleeing Muslim men. The image is accompanied by the caption: "Here we go again." A text overlay provides a time and place to gather for a protest in the city of Newcastle on August 10 and includes the hashtag "EnoughIsEnough." This content has had fewer than 1,000 views.

The third post, shared two days after the attack, is a repost of another likely AI-generated image. In it, four bearded Muslim men wearing white kurtas (tunics) are running in front of the Houses of Parliament in London, pursuing a crying blond-haired toddler in a Union Jack T-shirt. One of the men carries a knife. A plane flies towards Big Ben, seemingly a reference to the 9/11 terror attacks in 2001 in New York. The caption includes the words "Wake up" and the logo of an influential social media account known for anti-immigrant commentary in Europe, including misinformation and disinformation. This piece of content has had more than 1,000 views and fewer than 50 comments.

Facebook users reported all three posts for violating either the Hate Speech (renamed Hateful Conduct) or Violence and Incitement policies. Meta's automated tools assessed all three posts as non-violating and they were kept up. When the users appealed to Meta, the company's automated systems confirmed the decisions to leave up the content. The Board's selection of these cases was the first time any of the three posts were reviewed by humans. Following this, Meta reversed its decision on the text-only post, removing it for violating the Violence and Incitement policy, but confirmed its original decisions on the other two posts.

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

None of the users who posted the content in these cases responded to invitations to submit a statement to the Board.

The users who reported the posts provided statements to the Board claiming the posts were clearly encouraging people to attend racist protests, inciting violence against

immigrants and Muslims, or encouraging far-right supporters to continue rioting. One of the users said they were an immigrant and felt threatened by the post they were appealing about.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Violence and Incitement

Meta's Violence and Incitement policy rationale provides that the company removes "language that incites or facilitates violence and credible threats to public or personal safety," including "violent speech targeting a person or group of people on the basis of their protected characteristic(s) or immigration status." It also explains that Meta considers "language and context in order to distinguish casual or awareness-raising statements from content that constitutes a credible threat to public or personal safety."

The policy states that everyone is protected from "threats of violence that could lead to death (or other forms of high-severity violence)" and from "threats of violence that could lead to serious injury (mid-severity violence)." Meta's internal guidance to moderators mentions that this protection also extends to attacks on places that could lead to death or serious injury of a person. It includes calls to burn down or attack a place. The policy does not require moderators to confirm that people are inside the building.

The policy defines threats of violence as "statements or visuals representing an intention, aspiration, or call for violence against a target, and threats can be expressed in various types of statements such as statements of intent, calls for action, advocacy, expressions of hope, aspirational statements and conditional statements."

Hateful Conduct (previously named Hate Speech)

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 race, ethnicity, religious affiliation and national origin. The policy continues

to protect “refugees, migrants, immigrants and asylum seekers” under Tier 1 of the policy, which Meta considers to be the most severe attacks. However, they are not protected from attacks under Tier 2, in order to allow “commentary and criticism of immigration policies.” According to the policy rationale, this is because people sometimes “call for exclusion or use insulting language in the context of discussing political or religious topics, such as when discussing ... immigration.” Meta explicitly states that its “policies are designed to allow room for these types of speech.”

Tier 1 of the policy prohibits direct attacks that target people, based on a protected characteristic or immigration status, with “allegations of serious immorality and criminality,” providing violent criminals (“terrorists,” “murderers”) as examples.

Before January 7, Meta’s internal guidance to reviewers allowed “qualified behavioral statements,” distinguishing these from prohibited generalizations, including unqualified behavioral statements, alleging serious criminality. Qualified behavioral statements describe actions that individuals or groups have taken or their participation in events while mentioning their protected characteristic or immigration status. Prohibited generalizations attribute inherent traits to all or most members of an entire group (such as saying they are “killers” or they “kill”). Since January 7, Meta’s guidance to reviewers no longer prohibits behavioral statements, including against an entire protected characteristic group or based on immigration status. Meaning saying a protected characteristic group “kill” would be non-violating as a behavioral statement.

II. Meta’s Submissions

Text-Only Post

Meta reversed its original decision on this case, removing it for violating the Violence and Incitement policy. It did so because the calls for people to riot, “smash mosques,” and “do damage to buildings” where “migrants” and “terrorists” are living, are “statements advocating violence against a place that could result in death or serious injury.”

Giant Man Post

Meta found this post did not violate the Violence and Incitement policy. While it contains a call for people to attend a specific gathering, according to Meta it does not contain a threat of violence against people or property. Meta emphasized its policy, informed by its value of "voice," seeks to protect political speech around protests. Therefore, even with ongoing widespread disorder, a post would need to contain a threat or clear target to be violating.

Four Muslim Men Post

Meta found this post did not violate the Hateful Conduct (formerly Hate Speech) policy. While generalizations, such as attacking all or most Muslims as violent criminals would be violating, "referring to specific Muslim people as violent criminals" would not. Meta interpreted the image as referring to a specific "Muslim man or men who were incorrectly accused of stabbing the children in Southport," given the false information circulating at the time.

Crisis Measures

In response to the Board's questions, Meta explained it activated the Crisis Policy Protocol (CPP) in August and designated the entire UK as a Temporary High-Risk Location (THRL) from August 6–20, once the CPP was activated. THRL is a mechanism that enables Meta to implement additional safety measures, such as additional content restrictions or proactive monitoring to prevent incitement to violence in locations identified to be high-risk due to real-world events. During that time, Meta removed any calls to bring weapons to any location within the UK or to forcibly enter high-risk locations. The company did not set up an Integrity Product Operations Center (IPOC), which Meta describes as a "measure that brings together different teams, subject matter experts and capabilities from across the company (...) to respond in real time to potential problems or trends."

Third-Party Fact-Checking

Meta relied on third-party fact-checkers to review content during the riots and rate its accuracy. For "several pieces of content ... containing the false name of the Southport perpetrator" and rated as "false," Meta kept the content on the platform but attached

labels. It also removed the content from recommendations while demoting it in the feed of users that follow the account. Meta says it reduced such content's visibility "within hours of it appearing on the platform." Meta also established an internal working group of people from its policy, operations and law enforcement outreach teams to monitor and respond to the situation.

The Board asked Meta 13 questions about specific crisis-related measures deployed during the UK riots, including the role of third-party fact-checkers, details about the capabilities of its Hate Speech classifiers, how the context of the riots informed Meta's analysis of the content, whether any of the posts was demoted and the risks to free expression and access to information from overenforcement. Meta responded to all these questions.

4. Public Comments

The Oversight Board received nine public comments that met the terms for submission. Five of the comments were submitted from Europe, three from the United States and Canada and one from the Middle East and North Africa. Because the public comments period closed before January 7, 2025, none of the comments address the policy changes Meta announced on that date. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: social media's role in the 2024 UK riots, including in spreading misinformation and organizing and coordinating riots; the links between online anti-immigrant and anti-Muslim speech and violence; the use of imagery in hate speech and dehumanization; risks to freedom of expression from overenforcement; and, moderation measures short of removal.

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 immigrants and religious minorities in the context of a crisis. This case falls within the Board's strategic priorities of Crisis and Conflict Situations 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

I. *Content Rules*

Text-Only Post

The Board finds this post violates Meta's Violence and Incitement policy that prohibits credible threats of high-severity violence against a target and threats of violence against individuals or groups based on religion as a protected characteristic and immigration status.

While people may often post violent or threatening language online as hyperbole or in non-serious and joking ways, language and context distinguish casual statements from credible threats to public or personal safety. This post explicitly encourages people to riot, "smash mosques" and "do damage to buildings" where "migrants" and "terrorists" are living. This makes it a clear violation of Meta's Violence and Incitement policy in two ways: one, the general threat and incitement of high-severity violence against people and property; two, by targets being identified based on people's religion and immigration status. There is no way to interpret this post as a casual or non-serious statement. It was published on July 31 while violence was spreading across the UK, a day after a group threw bricks and petrol bombs at a mosque, and set a police car on fire, injuring eight officers. In the weeks following, similar violence ensued across the country.

Giant Man Post

The Board finds this post violates Meta's Violence and Incitement policy prohibiting threats of violence against individuals or groups based on religion as a protected characteristics and immigration status.

The Board notes that there were no written words in this post directly and expressly calling for people to engage in violence. However, this content demonstrates how

imagery combined with less direct written references to violence can also be an unambiguous form of incitement.

The text overlay to the image specified the date, time and location for people to gather, at a specific monument in Newcastle on August 10. It was posted after several days of violent riots across the country in which Muslims and immigrants were targets and people already had, among other things, attacked a hotel housing asylum seekers, torched a library and a community center, and pelted police officers with bottles and cans. The caption "Here we go again" is, when combined with the imagery of a giant white man aggressively pursuing smaller brown men in Islamic dress, a clear call for people to continue those ongoing acts of discriminatory violence and intimidation at a specified time and place. While the statement "Enough Is Enough" could be, alone and divorced from its context, a non-violent political statement about immigration, it had been used as a hashtag to organize prior riots and connect people for that purpose.

The Board finds that the combined elements of this post make the content policy violation clear. Meta's conclusion that the image contains no target or threat strains credulity and raises questions about why it took so long for the company to activate the Crisis Policy Protocol. By the time this post was shared, there was more than enough context about how information on the riots was spreading online to ensure violating inciting elements in this post could have been identified, if content like it had been prioritized for human review and appropriate interpretative guidance provided.

Four Muslim Men Post

The Board finds the content in the third case violates Meta's Hateful Conduct prohibition on allegations of serious criminality against a protected characteristic group. The January 7 policy changes did not change this assessment.

In this case, the visual of Muslim men pursuing a crying blond-haired toddler, alongside the terrorist imagery, generalizes that Muslims are violent criminals and terrorists, and a threat to British people and children specifically.

The image is a very clear example of a dehumanizing trope seeking to harness anti-immigrant sentiment by mobilizing anti-Muslim stereotypes. Through its elements, the

post generalizes Muslims as a collective national threat, portraying them as menacing and falsely attributing criminality and violence to them as a group defined by their religion. By visually linking Muslims to one of the most infamous terrorist events in modern history, the image falsely suggests that all Muslims are terrorists and a danger to Britain.

The Board disagrees with Meta's assessment that the image was a "qualified statement," i.e., that the depiction of a knife-yielding Muslim referred to the rumored perpetrator of the Southport attack, rather than Muslims more broadly. For the Board, while this content was posted in the context of the public disorder following the Southport stabbings and seeks to exploit the heightened emotions around them, it does not visually represent those events. At the time the image was posted, the Southport attacker was known to be a lone person and not a Muslim, the victims were three young girls and not a male toddler, and the attacks had no association with London or the 9/11 terrorist attacks. Inferring that the depiction of four Muslim men could be a reference to that lone attacker is incorrect. Moreover, even if the content depicted a lone Muslim, it would be a strange logic to invoke disinformation largely fueled by anti-Muslim prejudice to permit hate speech.

II. Enforcement Action

The two cases involving image-based violations of Meta's Violence and Incitement and Hateful Conduct policies raise concerns about how Meta moderates harmful content when it is based on imagery, rather than text. The Board has previously raised similar concerns in Posts in Polish Targeting Trans People, Planet of the Apes Racism, Hateful Memes Video Montage, Media Conspiracy Cartoon and Knin Cartoon. This concern is only heightened in these cases, as they demonstrate how the barriers to creating persuasive visual hate speech and incitement to violence are drastically lowering with the development of new AI tools. While an image being automatically generated will not change whether it is violating or not, new AI tools could significantly increase the prevalence of this content. This requires Meta to ensure its automated tools are better trained to detect violations in imagery and prioritize its human review until such a time that automated review is more reliable.

The Board is concerned about the delay in Meta activating its Crisis Policy Protocol, a mechanism the company created in response to previous Board recommendations. The company took almost a full week to designate the UK as a Temporary High-Risk Location. As part of this measure, Meta instituted temporary prohibitions on calls to bring weapons to or forcibly enter specific locations.

The Board believes that activation of the Crisis Policy Protocol would have been more effective if deployed promptly, in the critical hours and days following the attack, when false information about the attacker spread rapidly online and social media was used to organize and coordinate violence fueled by anti-immigrant, racist and anti-Muslim sentiment.

Additional interventions could have facilitated quicker and more accurate proactive moderation of content linked to the riots, interrupting amplification of harmful content and potentially reducing the risk of further harm. Operational tools could have been deployed to identify and review potentially violating content, proactively scan the platforms for specific keywords or hashtags and assign specialized regional teams. These teams could have provided additional context and guidance to at-scale reviewers moderating hate speech and incitement, including in visual forms.

The Board emphasizes that decisions to activate crisis-related measures must be made as quickly as possible. To achieve this, the company should identify core criteria that, when met in predefined combinations or individually, will trigger the immediate activation of the Crisis Policy Protocol. Additionally, this assessment should be repeated throughout the crisis to ensure that the measures in place are appropriate, effective and calibrated to the evolving risks.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that the removal of all three posts, as required by a proper interpretation of Meta's content policies, is also consistent with Meta's human rights 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). 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 the 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 (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.*). The UN Special Rapporteur on freedom of expression has stated that when applied to private actors' governance of online speech, rules should be clear and specific (A/HRC/38/35, para. 46). People using Meta's

platforms should be able to access and understand the rules and content reviewers should have clear guidance regarding their enforcement.

The Board finds it is not clear to users that Meta's Violence and Incitement policy prohibits threats against places as well as against people, noting in the context of the UK riots many places were targeted because of their association with Muslims, asylum seekers and immigrants.

The Board finds that the Hateful Conduct prohibition on allegations about "[v]iolent criminals (including but not limited to: terrorists, murderers)" is sufficiently clear as applied to the four Muslim men post. However, Meta's attempt to distinguish prohibited generalizations about an entire group's inherent qualities from permissible behavioral statements that may not apply to an entire group (i.e. referring to a group as "terrorists" or "murderers" versus saying they "murder") causes significant confusion. Both can be dehumanizing generalizations, depending on the context, and the distinction in enforcement may create perceptions of arbitrariness.

II. Legitimate Aim

Any restriction on freedom of expression should pursue one of the legitimate aims of the ICCPR, which includes the "rights of others" and the "protection of public order" (Article 19, para. 3, ICCPR). The Board has previously held that Meta's Violence and Incitement policy pursues the legitimate aim of protecting public order and the rights of others, including in particular the right to life (see Iranian Woman Confronted on Street and Tigray Communication Affairs Bureau). The Board has also previously held 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 Special Rapporteur on freedom of expression has also noted that on social media, "the scale and complexity of addressing hateful expression presents long-term challenges," (A/HRC/38/35, para. 28). However, according to the Special Rapporteur, companies should "demonstrate the necessity and proportionality of any content actions (such as removals or account suspensions)." Companies are required "to assess the same kind of questions about protecting their users' right to freedom of expression" (ibid para. 41).

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. However, when such conditions are met, it may merit removal of content (see also Criticism of EU Migration Policies and Immigrants decision).

The Board finds that all three posts should have been removed under Meta's policies, and their removal is necessary and proportionate considering the six factors outlined in the Rabat Plan of Action (The Rabat Plan of Action, OHCHR, A/HRC/22/17/Add.4, 2013). Those factors are: the social and political context; the status of the speaker; the intent to incite people to act against a target group; the content and form of the speech; the extent of dissemination; and, the likelihood and imminence of harm.

- **Context:** The ongoing riots were marked by escalating violence seeking to target specific groups. The riots were fueled by viral disinformation on social media often amplified by influential accounts (see also public comment by The Institute for Strategic Dialogue, PC-30832). Some of these accounts were linked to far-right groups and individuals. They used the spike in online activity to organize and mobilize people for anti-Muslim protests outside the mosque in Southport.

Those protests turned violent, as did many subsequent demonstrations, including the ones in Sunderland, Rotherham and Manchester.

- **Content and Form:** As outlined above, the content of all three posts, whether in writing or visual form, would be clearly understood as encouraging people to join the riots, either directly, or by deploying dehumanizing and hateful language towards Muslims and immigrants amid violent riots.
- **Speaker's status, intent and extent:** In the context of severe public disorder, the posts of even non-influential figures encouraging specific acts of violence had the potential to go viral and do great harm. The third piece of content, showing four Muslim men running after a toddler, also displayed the logo of a prominent social media account. In December 2024, Radio Free Europe published an investigation documenting the account's pattern of spreading false information targeting immigrants and the measures taken by the account to avoid its ownership being identified.
- **Likelihood and imminence of violence, discrimination and hostility**: During this period, and given that each post directly calls for or encourages violence against Muslims and immigrants, the likelihood of a single hateful post inciting additional unrest and violence was significant. Given the context in which these posts were shared, less restrictive measures would not have been sufficient to address the likely and imminent risk of violence, making removal under Meta's policies necessary and proportionate to their legitimate aim.

Enforcement

It is a concern that, even after the Board selected these cases, Meta maintained that two posts including AI-generated imagery were non-violating. It seems moderators (and even Meta's policy teams) are given a checklist that is interpreted too formulaically, depending on singular elements to be present for a violation to be found. This appears to be in the pursuit of consistent enforcement. But this guidance, mainly written with text-based posts in mind, ignores how visual imagery works, resulting in inconsistencies in enforcement. This indicates a particular challenge for Meta when it comes to its rules on content alleging inherent criminality against a protected characteristic group, as these cases demonstrate. The current guidance to reviewers

appears to be especially outdated given how much social media content is predominantly image and video-based.

While consistency can be an important measure of the quality of Meta's moderation, this should not be at the expense of accurately accounting for context, particularly in visual portrayals of hate speech and incitement. During a rapidly unfolding crisis, like the UK riots, the real threat of loss of life and property is too high a cost. Accuracy requires considering context and using judgment. As discussed above, it is particularly important that Meta's Crisis Policy Protocol is activated swiftly and that reviewers are given context-specific guidance to ensure Meta's policies are accurately enforced.

The Board notes that in contexts like the UK riots, unverified and false information left unchallenged and uncorrected can be especially dangerous. Analysis by Professor Marc Owen Jones (specializing in misinformation and disinformation) in an X thread on July 30 explained that there were at least 27 million impressions for posts on X stating or speculating that the attacker was Muslim, a migrant, a refugee or a foreigner. He also noted that there were more than 13 million impressions for posts denouncing such speculation.

Meta's policies on misinformation are important in this context, in particular, its rule on removing "misinformation or unverifiable rumors that expert partners have determined are likely to directly contribute to a risk of imminent violence or physical harm to people," (see Alleged Crimes in Raya Kobo decision). For misinformation that does not risk imminent violence or physical harm, measures less intrusive than removal may be necessary, for example, providing additional information to correct falsehoods. Meta informed the Board its third-party fact-checkers reviewed "several pieces of content" that contain "the false name of the Southport perpetrator," soon after it began to spread, categorizing them as "false." Fact-checkers should have been able to do this once UK authorities released statements about the false name on July 30. These posts were then covered with the fact-check label, their visibility reduced "within hours of appearing on the platform," and users were directed to a fact-checker's article correcting the falsity. For more on Meta's approach to fact-checking see Removal of COVID-19 Misinformation policy advisory opinion. The Board does not know what

percentage of false content posted during the UK riots was reviewed by fact-checkers. The Board recalls its concerns that the number of fact-checkers Meta relies on is limited and too often a significant volume of content queued for review by fact-checkers is never assessed.

As Meta explores the rollout of its Community Notes program – with which it intends to replace third-party fact-checking, starting in the U.S. – it should examine the experience of platforms using similar tools to respond to misinformation during the riots in the UK and broader research into the effectiveness of Community Notes. For example, research by the Center for Countering Digital Hate (CCDH) of posts on X from five high-profile accounts that pushed false information during the UK riots found these accounts amassed over 430 million views. According to its analysis, of the 1,060 posts shared by these accounts between July 29 and August 5, only one had a Community Note.

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 different 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). The Board notes the relevance of the first recommendation in the Criticism of EU Migration Policies and Immigrants cases to addressing these concerns.

6. The Oversight Board's Decision

The Oversight Board overturns Meta's original decisions to leave up all three pieces of content, requiring the second and third posts to be removed.

7. Recommendations

Content Policy

1. To improve the clarity of its Violence and Incitement Community Standard, Meta should specify that all high-severity threats of violence against places are prohibited, as well as against people.

   The Board will consider this recommendation implemented when Meta updates the Violence and Incitement Community Standard.

Enforcement

2. To improve the clarity of its Hateful Conduct Community Standard, Meta should develop clear and robust criteria for what constitutes allegations of serious criminality, based on protected characteristics, in visual form. These criteria should align with and adapt existing standards for text-based hateful conduct, ensuring consistent application across both text and imagery.

   The Board will consider this recommendation implemented when the internal implementation standards reflect the proposed change.

3. To ensure Meta responds effectively and consistently to crises, the company should revise the criteria it has established to initiate the Crisis Policy Protocol. In addition to the current approach, in which the company has a list of conditions that may or may not result in protocol activation, the company should identify core criteria that, when met, are sufficient for the immediate activation of the protocol.

The Board will consider this recommendation implemented when Meta briefs the Board on its new approach for activation of the Crisis Policy Protocol and concludes a disclosure of the procedures in its Transparency Center.

4. To ensure accurate enforcement of its Violence and Incitement and Hateful Conduct policies in future crises, Meta's Crisis Policy Protocol should ensure potential policy violations that could lead to likely and imminent violence are flagged for in-house human reviewers. These reviewers should provide time-bound, context-informed guidance for at-scale reviewers, including for image-based violations.

   The Board will consider this implemented when Meta shares documentation on this new Crisis Policy Protocol lever, outlining how (1) potential violations are flagged for in-house review; (2) context-informed guidance is cascaded down; and (3) implemented for at-scale reviewers.

5. As the company rolls out Community Notes, it should undertake continuous assessments of the effectiveness of Community Notes as compared to third-party fact-checking. These assessments should focus on the speed, accuracy and volume of notes or labels being affixed in situations where the rapid dissemination of false information creates risks to public safety.

   The Board will consider this recommendation implemented when Meta updates the Board every six months until implementation is completed and shares the results of this evaluation publicly.

***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). When 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.