



OVERTURNED

2022-007-IG-MR



# UK drill music

The Oversight Board has overturned Meta's decision to remove a UK drill music video clip from Instagram.

## Policies and topics

-  Freedom of expression, Art / Writing / Poetry, Governments
-  Violence and incitement

## Region and countries

-  Europe
-  United Kingdom

## Platform

-  Instagram

## Attachments

[UK drill music - public comments](#)

[Metropolitan Police Freedom of Information request response](#)

## Case summary

The Oversight Board has overturned Meta's decision to remove a UK drill music video clip

from Instagram. Meta originally removed the content following a request from the Metropolitan Police. This case raises concerns about Meta's relationships with law enforcement, which has the potential to amplify bias. The Board makes recommendations to improve respect for due process and transparency in these relationships.

## About the case

In January 2022, an Instagram account that describes itself as publicizing British music posted content highlighting the release of the UK drill music track, "Secrets Not Safe" by Chinx (OS), including a clip of the track's music video.

Shortly after, the Metropolitan Police, which is responsible for law enforcement in Greater London, emailed Meta requesting that the company review all content containing "Secrets Not Safe." Meta also received additional context from the Metropolitan Police. According to Meta, this covered information on gang violence, including murders, in London, and the Police's concern that the track could lead to further retaliatory violence.

Meta's specialist teams reviewed the content. Relying on the context provided by the Metropolitan Police, they found that it contained a "veiled threat," by referencing a shooting in 2017, which could potentially lead to further violence. The company removed the content from the account under review for violating its Violence and Incitement policy. It also removed 52 pieces of content containing the track "Secrets Not Safe" from other accounts, including Chinx (OS)'s. Meta's automated systems later removed the content another 112 times.

Meta referred this case to the Board. The Board requested that Meta also refer Chinx (OS)'s post of the content. However, Meta said that this was impossible as removing the "Secrets Not Safe" video from Chinx (OS)'s account ultimately led to the account being deleted, and its content was not preserved.

## Key findings

The Board finds that removing this content does not align with Meta's Community Standards, its values, or its human rights responsibilities.

Meta lacked sufficient evidence to conclude that the content contained a credible threat, and the Board's own review did not uncover evidence to support such a finding. In the absence of such evidence, Meta should have given more weight to the content's artistic nature

such evidence, Meta should have given more weight to the content's artistic nature.

This case raises concerns about Meta's relationships with governments, particularly where law enforcement requests lead to lawful content being reviewed against the Community Standards and removed. While law enforcement can sometimes provide context and expertise, not every piece of content that law enforcement would prefer to have taken down should be taken down. It is therefore critical that Meta evaluates these requests independently, particularly when they relate to artistic expression from individuals in minority or marginalized groups for whom the risk of cultural bias against their content is acute.

The channels through which law enforcement makes requests to Meta are haphazard and opaque. Law enforcement agencies are not asked to meet minimum criteria to justify their requests, and interactions therefore lack consistency. The data Meta publishes on government requests is also incomplete.

The lack of transparency around Meta's relationship with law enforcement creates the potential for the company to amplify bias. A freedom of information request made by the Board revealed that all of the 286 requests the Metropolitan Police made to social media companies and streaming services to review or remove musical content from June 2021 to May 2022 involved drill music, which is particularly popular among young Black British people. 255 of these requests resulted in platforms removing content. 21 requests related to Meta platforms, resulting in 14 content removals. The Board finds that, to honor its values and human rights responsibilities, Meta's response to law enforcement requests must respect due process and be more transparent.

This case also raises concerns around access to remedy. As part of this case, Meta told the Board that when the company takes content decisions "at escalation," users cannot appeal to the Board. A decision taken "at escalation" is made by Meta's internal specialist teams. According to Meta, all decisions on law enforcement requests are made "at escalation" (unless the request is made through a publicly available "in-product reporting tool"), as are decisions on certain policies that can only be applied by Meta's internal teams. This situation adds to concerns raised in preparing the Board's policy advisory opinion on cross-check where Meta revealed that, between May and June 2022, around a third of content in the cross-check system could not be escalated to the Board.

Meta has referred escalated content to the Board on several occasions, including this one. However, the Board is concerned that users have been denied access to remedy when Meta

makes some of its most consequential content decisions. The company must address this problem urgently.

## **The Oversight Board's decision**

The Oversight Board overturns Meta's decision to remove the content.

The Board recommends that Meta:

- Create a standardized system for receiving content removal requests from state actors. This should include asking for criteria such as how a policy has been violated, and the evidence for this.
- Publish data on state actor content review and removal requests for Community Standard violations.
- Regularly review its data on content moderation decisions prompted by state actor requests to assess for any systemic biases, and create a mechanism to address any bias it identifies.
- Provide users with the opportunity to appeal to the Oversight Board when it makes content decisions “at escalation.”
- Preserve accounts and content penalized or disabled for posting content that is subject to an open investigation by the Board.
- Update its value of “Voice” to reflect the importance of artistic expression.
- Clarify in the Community Standards that for content to be removed as a “veiled threat,” one primary and one secondary signal is required, and make clear which is which.

\*Case summaries provide an overview of the case and do not have precedential value.

## Full case decision

### **1. Decision summary**

The Oversight Board overturns Meta's decision to remove a clip from Instagram announcing the release of a UK drill music track by the artist Chinx (OS).

Meta referred this case to the Board because it raises recurring questions about the

appropriate treatment of artistic expression that references violence. It involves a balance between Meta's values of "Voice," in the form of artistic expression, and "Safety."

The Board finds that Meta lacked sufficient evidence to independently conclude that the content contained a credible veiled threat. In the Board's assessment the content should not have been removed in the absence of stronger evidence that the content could lead to imminent harm. Meta should have more fully taken into account the artistic context of the content when assessing the credibility of the supposed threat. The Board finds that the content did not violate Meta's Community Standard on Violence and Incitement, and its removal did not sufficiently protect Meta's value of "Voice," or meet Meta's human rights responsibilities as a business.

This case raises broader concerns about Meta's relationship with governments, including where law enforcement requests Meta to assess whether lawful content complies with its Community Standards. The Board finds the channels through which governments can request such assessments to be opaque and haphazard. The absence of transparency and adequate safeguards around Meta's relationship with law enforcement creates the potential for the company to exacerbate abusive or discriminatory government practices.

This case also reveals that, for content moderation decisions Meta takes at escalation, users have been wrongly denied the opportunity to appeal to the Oversight Board. Decisions "at escalation" are those made by Meta's internal, specialist teams rather than through the "at scale" content review process. This lack of appeal availability adds to concerns about access to the Board that will be addressed in the Board's upcoming policy advisory opinion on cross-check. Combined, these concerns raise serious questions about users' right of access to remedy when Meta makes some of its most consequential content decisions at escalation. The company must address this problem urgently.

## **2. Case description and background**

In January 2022, an Instagram account that describes itself as promoting British music posted a video with a short caption on its public account. The video was a 21-second clip of the music video for a UK drill music track called "Secrets Not Safe" by the rapper Chinx (OS). The caption tagged Chinx (OS) as well as an affiliated artist and highlighted that the track had just been released. The video clip is of the second verse of the track and ends by fading out to a black screen with the text "OUT NOW."

Drill is a subgenre of rap music popular in the UK, in particular among young Black people, with many drill artists and fans in London. This music genre is hyper-local, where drill collectives can be associated with areas as small as single housing estates. It is a grassroots genre, widely performed in English in an urban context with a thin line separating professional and amateur artists. Artists often speak in granular detail about ongoing violent street conflicts, using a first-person narrative with imagery and lyrics that depict or describe violent acts. Potential claims of violence and performative bravado are considered to be part of the genre – a form of artistic expression where fact and fiction can blur. Through these claims, artists compete for relevance and popularity. Whether drill music causes real-world violence or not is disputed, particularly the reliability of evidential claims made in the debate.

In recent years, recorded incidents of gun and knife violence in London have been high, with disproportionate effects on Black communities.

The lyrics of the track excerpt are quoted below. The Board has added meanings for non-standard English terms in square brackets and redacted the names of individuals:

*Ay, broski [a close friend], wait there one sec (wait). You know the same mash [gun] that I showed [name redacted] was the same mash that [name redacted] got bun [shot] with. Hold up, I'm gonna leave somebody upset (Ah, fuck). I'm gonna have man fuming. He was with me putting loud on a Blue Slim [smoking cannabis] after he heard that [name redacted] got wounded. [Name redacted] got bun, he was loosing (bow, bow) [he was beaten]. Reverse that whip [car], confused him. They ain't ever wheeled up a booting [a drive-by shooting] (Boom). Don't hit man clean, he was moving. Beat [shoot] at the crowd, I ain't picking and choosing (No, no). Leave man red [bleeding], but you know [track fades out].*

Shortly after the video was posted, Meta received a request via email from the UK Metropolitan Police to review all content that included this Chinx (OS) track. Meta says law enforcement provided context of gang violence and related murders in London, and flagged that elements of the full track might increase the risk of retaliatory gang violence.

Upon receiving the request from the Metropolitan Police, Meta escalated the content review to its internal Global Operations team and then to its Content Policy team. Meta's Content Policy team makes removal decisions following input from subject matter experts and after specialized contextual reviews. Based on the additional context the Metropolitan Police provided, Meta took the view that the track excerpt referenced a shooting in 2017. It

determined that the content violated the Violence and Incitement policy, specifically the prohibition on "coded statements where the method of violence or harm is not clearly articulated, but the threat is veiled or implicit." Meta believed that the track's lyrics, "acted as a threatening call to action that could contribute to a risk of imminent violence or physical harm, including retaliatory gang violence." Meta therefore removed the content.

Hours later, the content creator appealed the decision to Meta. Usually, users cannot appeal to Meta content decisions the company takes through its escalation process. This is because user appeals to Meta are not routed to escalation teams but to at-scale reviewers. Without access to the additional context available at escalation, those reviewers would be at increased risk of making errors, and incorrectly reversing decisions made at escalation. In this case however, due to human error, the user was able to appeal the escalated decision to Meta's at-scale reviewers. An at-scale reviewer assessed the content as non-violating and restored it to Instagram.

Eight days later, following a second request from the UK Metropolitan Police, Meta removed the content through its escalations process again.

The account in this case has fewer than 1,000 followers, the majority of whom live in the UK. The user received notifications from Meta both times their content was removed but was not informed that the removals were initiated following a request from UK law enforcement.

Alongside removing the content under review, Meta identified and removed 52 pieces of content featuring the "Secrets Not Safe" track from other accounts, including Chinx (OS)'s account. Meta added the content at issue in this case to the Violence and Incitement Media Matching Service bank, marking it as violating. These banks automatically identify matching content and can remove it or prevent it from being uploaded to Facebook and Instagram. Adding the video to the Media Matching Service bank resulted in an additional 112 automated removals of matching content from other users.

Meta referred this case to the Board. The content it referred was posted from an account that was not directly associated with Chinx (OS). Because Chinx (OS)'s music is at the center of this case, the Board requested Meta to additionally refer its decision to remove Chinx (OS)'s own post featuring the same track. Meta explained this was not possible, because the removal of the video from the artist's account had resulted in a strike. This caused the account to exceed the threshold for being permanently disabled. After six months, and prior

to the Board's request for the additional referral, Meta permanently deleted Chinx (OS)'s disabled account, as part of a regular, automated process, despite the Oversight Board's pending decision on this case. The action to delete Chinx (OS)'s account, and the content on it, was irreversible, making it impossible to refer the case to the Board.

### **3. Oversight Board authority and scope**

The Board has authority to review decisions that Meta refers for review (Charter Article 2, Section 1; Bylaws Article 2, Section 2.1.1). The Board may uphold or overturn Meta's decision (Charter Article 3, Section 5), and this decision is binding on the company (Charter Article 4). Meta must also assess the feasibility of applying its decision in respect of identical content with parallel context (Charter Article 4). The Board's decisions may include policy advisory statements with non-binding recommendations that Meta must respond to (Charter Article 3, Section 4; Article 4).

When the Board identifies cases that raise similar issues, they may be assigned to a panel together. In this case, the Board requested that Meta additionally refer content featuring the same track posted by the artist Chinx (OS). In the Board's view, the difficulty of balancing safety and artistic expression could have been better addressed by Meta referring Chinx (OS)'s post of his music from his own account. This would also have allowed the Board to issue a binding decision in respect of the artist's post. Meta's actions in this case have effectively excluded the artist from formally participating in the Board's processes, and have removed Chinx (OS)'s account from the platform without access to remedy.

On several occasions, including this one, Meta has referred content that was escalated within Meta to the Board (see, for example, the "Tigray Communication Affairs Bureau" case, and the "Former President Trump's suspension" case). When Meta takes a content decision "at escalation," users are unable to appeal the decision to the company or to the Board. As Meta is able to refer cases decided at escalation to the Board, users who authored or reported the content should equally be entitled to appeal to the Board. Decisions made at escalation are likely to be among the most significant and difficult, where independent oversight is at its most important. The Board's governing documents provide that all content moderation decisions that are within scope and not excluded by the Bylaws (Bylaws Article 2, Sections 1.2, 1.2.1) and that have exhausted Meta's internal appeal process (Charter Article 2, Section 1) be eligible for people to appeal to the Board.

### **4. Sources of authority**



## 4. Sources of authority

The Oversight Board considered the following authorities and standards:

### *I. Oversight Board decisions:*

- “Colombian police cartoon” (case decision [2022-004-FB-UA](#)): responding to concerns that Media Matching Service banks can amplify the impact of incorrect decisions, the Board recommended Meta urgently improve procedures to quickly remove non-violating content incorrectly added to these banks.
- “Knin cartoon” (case decision [2022-001-FB-UA](#)): the Board overturned Meta’s decision to leave content up on the platform, finding that an implicit comparison between ethnic Serbs and rats violated the Hate Speech Community Standard, and the reference within this comparison to a past historical event also violated the Violence and Incitement Community Standard.
- “Wampum belt” (case decision [2021-012-FB-UA](#)): the Board emphasized the importance of Meta respecting artistic expression.
- “Shared Al Jazeera post ” (case decision [2021-009-FB-UA](#)): the Board recommended that Meta formalize a transparent process for how it handles and publicly reports on government requests. Transparency reporting should distinguish between government requests resulting in removals for violations of the Community Standards, requests that led to removal or geo-blocking for violating local law, and requests that led to no action.
- “Öcalan's isolation” (case decision [2021-006-IG-UA](#)): the Board recommended that Meta notify users when their content is removed following a government request. The decision also made recommendations on transparency reporting on these requests.
- “Protest in India against France” (case decision [2020-007-FB-FBR](#)): the Board noted the challenges of addressing veiled threats of violence at scale due to the importance of contextual analysis.
- “Breast cancer symptoms and nudity” (case decision [2020-004-IG-UA](#)): the Board recommended that Meta clarify to Instagram users that Facebook's Community Standards apply to Instagram in the same way they apply to Facebook. This recommendation was repeated in the “Reclaiming Arabic words” case decision ( [2022-003-IG-UA](#)).

## II. Meta's content policies:

This case involves [Instagram's Community Guidelines](#) and [Facebook's Community Standards](#).

The [Instagram Community Guidelines](#) say, under the heading “Respect other members of the Instagram Community,” that the company wants to “foster a positive, diverse community.” The company removes content that contains “credible threats,” with those words linked to the Facebook Violence and Incitement Community Standard. The Guidelines further set out that:

*Serious threats of harm to public and personal safety aren't allowed. This includes specific threats of physical harm as well as threats of theft, vandalism and other financial harm. We carefully review reports of threats and consider many things when determining whether a threat is credible.*

The policy rationale for the Facebook [Community Standard on Violence and Incitement](#), states it “aim[s] to prevent potential offline harm that may be related to content on Facebook” and that while Meta “understand[s] that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence.” It further provides that Meta removes content, disables accounts and works with law enforcement “when [it] believe[s] there is a genuine risk of physical harm or direct threats to public safety.” Meta states it tries “to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat.”

Under a subheading stating that Meta requires “additional information and/or context to enforce,” the Community Standard provides that users should not post coded statements “where the method of violence or harm is not clearly articulated, but the threat is veiled or implicit.” Those include “[r]eferences [to] historical or fictional incidents of violence” and where “[l]ocal context or subject matter expertise confirm that the statement in question could be threatening and/or could lead to imminent violence or physical harm.”

## III. Meta's values:

The value of “Voice” is described as “paramount”:

*The goal of our Community Standards is to create a place for expression and give people a*

voice. Meta wants people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable.

Meta limits "Voice" in the service of four values. "Safety" and "Dignity" are the most relevant in this case:

Safety: We're committed to making Facebook a safe place. We remove content that could contribute to a risk of harm to the physical security of persons. Content that threatens people has the potential to intimidate, exclude or silence others and isn't allowed on Facebook.

Dignity: We believe that all people are equal in dignity and rights. We expect that people will respect the dignity of others and not harass or degrade others.

#### IV. International human rights standards:

The UN Guiding Principles on Business and Human Rights (UNGPs), endorsed by the UN Human Rights Council in 2011, establish a voluntary framework for the human rights responsibilities of private businesses. In 2021, Meta announced its Corporate Human Rights Policy, where it reaffirmed its commitment to respecting human rights in accordance with the UNGPs. The Board's analysis of Meta's human rights responsibilities in this case was informed by the following human rights standards:

- The right to freedom of opinion and expression: Article 19, International Covenant on Civil and Political Rights ( ICCPR), General Comment No. 34, Human Rights Committee, 2011; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018), A/74/486 (2019), A/HRC/44/49/Add.2 (2020).
- The right to life: Article 6, ICCPR.
- The right to security of person: Article 9, para. 1, ICCPR.
- The right to effective remedy: Article 2, ICCPR; General Comment No. 31, Human Rights Committee, (2004).
- Equality and non-discrimination: Article 2, para. 1 and Article 26 (ICCPR); Article 2, ICERD.
- Cultural rights: Article 27, ICCPR; Article 15, International Covenant on Economic, Social and Cultural Rights (ICESCR); UN Special Rapporteur in the field of cultural rights, report on artistic freedom and creativity, A/HRC/23/34, 2013.

## 5. User submissions

Meta referred this case to the Board, and the Board selected it in mid-June 2022. Due to a technical error that has since been fixed, Meta did not successfully notify the user that the Board selected a case concerning content they had posted and did not invite them to submit a statement to the Board. At the end of August, Meta manually notified the user, but the user did not provide a user statement within the deadline of 15 days.

## 6. Meta's submissions

Meta explained to the Board that it removed the content because the Instagram post violated its Violence and Incitement policy by containing a veiled threat of violence. Meta argued that its decision comports with international human rights principles because the Community Standards explain that users may not post veiled threats of violence; because the application of this policy serves the legitimate aim of protecting the rights of others and public order; and because the removal of the content at issue was necessary and proportionate to accomplish those aims.

Meta deems the decision particularly challenging because its Violence and Incitement policy does not have explicit exceptions for humor, satire, or artistic expression. The policy requires Meta to assess whether a threat is credible or merely a show of bravado or provocative, but ultimately nonviolent, expression. Meta considers this case to be significant because it raises recurring questions about the appropriate treatment of artistic expression that references violence. This assessment involves balancing its values of "Voice" and "Safety." Meta told the Board that, when a creator's work includes threats of violence or statements that could contribute to a risk of violence it "err[s] on the side of removing it from our platforms."

Meta's internal guidance for moderators, the internal Implementation Standards, sets out the "veiled threats analysis" which Meta uses to determine the existence of veiled threats under the Violence and Incitement Community Standard. This explains that for content to qualify as a veiled threat there must be both a primary signal (such as a reference to a past act of violence), and a secondary signal. Secondary signals include local context or subject matter expertise indicating that the content is potentially threatening, or confirmation by the target that they view the content as threatening. According to Meta, local NGOs, law enforcement agencies, Meta's Public Policy team, or other local experts provide secondary signals. The

“veiled threats analysis” is only performed “at escalation,” meaning it cannot be performed by “at-scale” reviewers. It can only be conducted by Meta’s internal teams.

In this case, Meta found the content contained a primary signal in referring to the rapper’s participation in an earlier shooting and indicating an intent to respond further. According to Meta, the secondary signal was the UK Metropolitan Police’s confirmation that it viewed the content as potentially threatening or likely to contribute to imminent violence or physical harm. Law enforcement did not allege the content violated local law. Meta says it assesses law enforcement reports alongside political, cultural and linguistic expertise from Meta’s internal teams.

Meta argued that drill music has often been connected to violence, citing a [Policy Exchange report](#) claiming that approximately one quarter of London’s gang murders have been linked to drill music. However, Meta later acknowledged this report faced “some criticism from criminologists.” Meta quoted an [open letter](#), signed by 49 criminologists, social scientists and professional organizations, which “dismiss[ed] the report as factually inaccurate, misleading and politically dangerous” and for “committing grave causation-correlation errors.”

Certain Meta policies can only be enforced by Meta’s internal teams, through its internal escalation process. This is known as being decided “at escalation.” Meta provided a list of around 40 “escalation-only” rules in nine policy areas. The Board asked Meta how, in this case, an at-scale reviewer was able to restore a post that had been removed at escalation. Meta responded that “exceptionally in this case and due to a human error, the content creator was able to appeal the initial removal decision.” Meta disclosed that where content is actioned through Meta’s internal escalation process, it is not usually possible to appeal for a second examination by the company. This is to prevent “at-scale” reviewers from reversing decisions made “at escalation” without access to the context available in escalated review.

At the Board’s request, Meta provided a briefing on “Government requests to review content for Community Standard violations.” Meta explained to the Board that governments can make requests to the company to remove content by email, post, or the help center, as well as through in-product reporting tools, which send the content to automated or “at scale” review.

Meta explained that when it receives a request from law enforcement made outside the in-product tools, the content goes through an internal escalation process. Regardless of how an

product tools, the content goes through an internal escalation process. Regardless of how an escalation is received, there is a standard scope and prioritization process to assess the urgency, sensitivity and complexity of the escalation. This process determines which of Meta's teams will handle the request and the position of the request in the queue. Following a review request from a third party, including governments, Meta's Global Operations Team manually completes a form with signals. The prioritization model and pillars take into consideration signals related to legal, reputational, and regulatory risks, impact to the physical safety of Meta's community, the scope and viewership of the issue at hand, and the time sensitivity of the issue.

These prioritization pillars are ranked in order of importance and a priority score is automatically calculated based on the signals entered. The model is dynamic and responsive to changes in the environment (e.g., offline events) and refinements Meta introduces. High priority escalations, such as in this case, are sent to a specialist team within Meta's Global Operations team. The team reviews content against the Community Standards, investigates, reaches out to stakeholders for additional assessment where necessary, and makes an enforcement decision. In some cases, including this one, content is then escalated to the Content Policy team for input. The action taken is communicated to the external person(s) who submitted the original request. Meta shared that their human rights team does not typically weigh-in on individual applications of the veiled threats framework.

Meta states that when it receives content removal requests from law enforcement agencies, it evaluates the content against the Community Standards in the same way it would for any other piece of content, regardless of how it was detected or escalated. Meta claims this means that requests are treated the same way in all countries. Meta explained that the priority score is affected by the context that is provided in relation to the pillars. While the identity of the requester is not a factor that directly affects the prioritization, these specific types of context do impact the prioritization.

To appeal to the Oversight Board, a user must have an appeal reference ID. Meta issues these IDs as part of its appeals process. In response to the Board's questions, Meta confirmed it does this for content decisions that are eligible for internal appeal and after a second review has been exhausted. Therefore, in situations where Meta acts on content without allowing for an appeal, there is no opportunity to appeal to the Board. This is usually the case for decisions taken "at escalation," such as those reviewed following government requests (made outside of the in-product reporting tools), and content reviewed under "escalation-only" policies.

The Board formally submitted to Meta a total of 26 written questions, including three rounds of follow-up questions. These numbers exclude questions the Board asked during the in-person briefing Meta provided to the Board on how it handles government requests. Twenty-three of the written questions were answered fully and three requests were not fulfilled by Meta. Meta declined to provide data on law enforcement requests globally and in the UK focusing on “veiled threats,” drill music, or the proportion of requests resulting in removal for Community Standard violations. Further, Meta declined to provide a copy of the content review requests received from the Metropolitan Police in this case. However, the Metropolitan Police provided the Board with a copy of the first request sent to Meta, on condition that the content of the request remain confidential.

## 7. Public comments

The Oversight Board considered ten public comments related to this case. One comment was submitted from each of the following regions: Europe; Middle East and North Africa; and Latin America and Caribbean. Two comments were submitted from Central and South Asia, and five from the United States and Canada. The Metropolitan Police provided a comment; it understood that the Board would disclose this fact, but did not give permission for the comment to be published. The Board requested the Metropolitan Police revisit that decision in the interests of transparency, but it declined. The Metropolitan Police indicated it may be able to provide consent at a later point in time. If that happens, the Board will share the public comment.

The submissions covered the following themes: racial bias and disproportionate over-targeting of Black communities by law enforcement; the importance of socio-cultural context in assessing artistic expression; causal links between drill music and violence; and handling government-issued takedown requests.

The Board filed a Freedom of Information Act request (Reference No: 01/FOI/22/025946) to the Metropolitan Police to provide information about its policies and practice on making requests to social media and streaming companies to review and/or remove content. Its responses to that request inform the Board’s analysis below.

To read public comments submitted for this case, please click [here](#). To read the Metropolitan Police’s response to the Board’s Freedom of Information request, please click [here](#).

## 8. Oversight Board analysis

The Board looked at the question of whether this content should be restored through three lenses: Meta’s content policies, the company’s values and its human rights responsibilities.

### 8.1 Compliance with Meta’s content policies

#### *1. Content rules*

The Board finds that Meta’s removal of the content in this case did not comply with the Violence and Incitement Community Standard.

Detecting and assessing threats at scale is challenging, in particular where they are veiled, and when specific cultural or linguistic expertise may be required to assess context (see the Oversight Board decisions “Protest in India against France” and “Knin cartoon”). Artistic expression can contain veiled threats, as can any other medium. The challenge of assessing the credibility of veiled threats in art is especially acute. Messages in art can be intentionally obscure in their intent and deliberately subject to interpretation. Statements referencing violence can be coded, but also can be of a performative or satirical nature. They may even characterize certain art forms, such as drill music. Meta acknowledged these challenges when referring this case to the Board.

The Board agrees with Meta that the lyrics in the video clip did not contain an overt threat of violence under the Violence and Incitement Community Standard.

It is a more challenging question whether the video contains a veiled threat under the same policy. The policy rationale outlines that Meta seeks to “prevent potential offline harm,” and that language “that incites or facilitates serious violence” and that “poses a genuine risk of physical harm or direct threats to public safety” will be removed. An emphasis is placed on distinguishing credible threats from non-credible threats. Establishing a causal relationship between language and risk of harm requires a resource-intensive analysis.

For content to constitute a “veiled or implicit” threat under the Violence and Incitement Community Standard, the method of violence or harm need not be clearly articulated. Meta uses its “veiled threats analysis,” set out in its non-public guidance to moderators, the internal Implementation Guidance, to assess whether a veiled threat is present. This requires



internal implementation guidance, to assess whether a veiled threat is present. This requires that both a primary and secondary signal are identified for content to qualify as a veiled threat.

The Board agrees with Meta that a primary signal is present in this case. The lyrics contain reference to a “historical” incident of violence. Additional context is required to understand that this is a 2017 shooting between two rival gangs in London. For Meta, the excerpt in its entirety referred to these events. The Board’s conclusion, agreeing with Meta’s that a primary signal is present, is based on two independent third-party analyses of the lyrics sought by the Board. The Board notes that these analyses differed in substantial ways from Meta’s interpretation. For example, Meta interpreted the term “mash” to mean “cannabis,” whereas experts the Board consulted interpreted this term to mean “gun.” Meta interpreted the term “bun” to mean “high,” whereas the Board’s experts interpreted this to mean “shot.”

Identifying a veiled or implicit threat also requires a secondary signal showing that the reference “could be threatening and/or *could* lead to imminent violence or physical harm” [emphasis added]. That signal depends on local context, often provided by third parties such as law enforcement, confirming the content, “is considered potentially threatening, or *likely* to contribute to imminent violence or physical harm” [emphasis added]. In this case, the UK Metropolitan Police provided this confirmation. Meta determined that Chinx (OS)’s reference to the 2017 shooting was potentially threatening, or likely to contribute to imminent violence or physical harm and qualified as a veiled threat. Meta’s contextual assessment included the specific rivalry between gangs associated with the 2017 shooting, as well as the broader context of inter-gang violence and murders in London.

It is appropriate that Meta draws upon local subject matter expertise to evaluate the relevant context and credibility of veiled threats. The Board notes that there is understandable anxiety around high levels of gun and knife violence in recent years in London, with disproportionate effects on Black communities. Law enforcement can sometimes provide such context and expertise. But not every piece of content that law enforcement would prefer to have taken down – and not even every piece of content that has the *potential* to lead to escalating violence – should be taken down. It is therefore critical that Meta evaluate these requests itself and reach an independent conclusion. The company says they do this. Independence is crucial, and the evaluation should require specific evidence of how the content cause harm. This is particularly important in counteracting the potential for law enforcement to share information selectively, and the limited opportunity to gain counter-perspectives from other stakeholders. For artistic expression from individuals in minority or marginalized groups, the

risk of cultural bias against their content is especially acute.

In this case, Meta has not demonstrated that the lyrics in the content under review constituted a credible threat or risk of imminent harm, nor has the Board's own review uncovered evidence to support such a finding. To establish that the reference to a shooting five years ago presents a risk of harm today requires additional probative evidence beyond the reference itself. The fact that the track references events that involve gangs engaged in a violent rivalry does not mean artistic references to that rivalry necessarily constitute a threat. In the absence of either sufficient detail to make that causal relationship clearer, such as evidence of past lyrics materializing into violence or a report from the target of the purported threat that they were endangered, greater weight should have been afforded to the artistic nature of the alleged threat when evaluating its credibility. The fact that performative bravado is common within this musical genre was relevant context that should have informed Meta's analysis of the likelihood that the track's reference to past violence constituted a credible present threat. Third party experts informed the Board that a line-by-line lyrical analysis to determine evidence of past wrongdoing or risk of future harm is notoriously inaccurate and that verifying supposedly factual statements within drill lyrics is challenging (Digital Rights Foundation, PC-10618). Public comments (e.g., Electronic Frontier Foundation, PC-10971) criticized law enforcement's policing of lawful drill music, and research Meta cited in its submissions has been widely criticized by criminologists, as the company has acknowledged. In the Board's view, this criticism should also have factored into Meta's analysis, and prompted it to request additional information from law enforcement and/or from additional parties in relation to causation before removing the content.

The Board further notes that the full Chinx (OS) track, which was excerpted in this case, remains available on music streaming platforms accessible in the UK and the Board has seen no evidence this has led to any act of violence. This context was not available to Meta at the time of its initial decision in this case, but it is nonetheless relevant to the Board's independent review of the content.

The Board acknowledges the deeply contextual nature of this kind of content decision that Meta has to make, and the associated time pressure when there may be risks of serious harm. Reasonable people might differ, as in this case, on whether a given piece of content constitutes a veiled threat. Still, the lack of transparency on Meta's decisions to remove content that result from government requests makes it difficult to evaluate whether Meta's mistake in an individual case reflects reasonable disagreement or is indicative of potential

systemic bias that requires additional data and further investigation. Meta insists that its “veiled threats analysis” is independent — and the Board agrees that it should be — but in this context, Meta’s say-so is not enough. Meta states that it evaluates the content against the Community Standards in the same way it would for any other piece of content, regardless of how it is detected. The “veiled threats analysis” places law enforcement in the position of both reporting the content (i.e., flagging a primary signal), and providing all the contextual information Meta needs to assess potential harm (i.e., providing the local knowledge needed for the secondary signal). While there may be good reasons to adopt a prioritization framework that ensures reports from law enforcement are assessed swiftly, that process needs to be designed to ensure that such reports include sufficient information to make independent assessment possible, including seeking further input from the requesting entity or other parties where necessary.

The Board distinguishes this decision from its “Knin cartoon” decision. In that case, the additional context to enforce the “veiled threat” rule was the presence, in the cartoon, of hate speech against the group targeted in the prior violent incident. The Board primarily justified the removal based on the Hate Speech Community Standard. The Board’s finding that the content additionally violated the Violence and Incitement Community Standard relied on contextual knowledge to understand the historical incident (Operation Storm) referenced in the post. This contextual knowledge is well known within the region, both to Croatian language speakers and ethnic Serbs. This was evidenced by the sheer number of reports the content in that case received (almost 400), compared to the content in this case which received no reports. Incitement of hatred against an ethnic group was immediately apparent to a casual observer. While understanding many of the references within the content relied on specific contextual knowledge, that knowledge could be gained without relying on external third parties.

## *II. Enforcement action and transparency*

In this case, one request from law enforcement resulted in 52 manual removals and 112 automated removals of matching pieces of content (between January 28, 2022, and August 28, 2022). It is important to recognize that the actions Meta took in response to the request from the Metropolitan Police impacted not only the owner of the Instagram account in this case, but also Chinx (OS) and many others (see also: “Colombia police cartoon” case).

The scale of these removals underscores the importance of due process and transparency

around Meta's relationship with law enforcement and the consequences of actions taken pursuant to that relationship (see also the Oversight Board decisions "Öcalan isolation" and "Shared Al Jazeera post"). To address these concerns, there needs to be a clear and uniform process with safeguards against abuse, including auditing; adequate notice to users of government involvement in the action taken against them; and transparency reporting on these interactions to the public. These three aspects are interconnected, and all must be addressed.

#### a. Transparency to the public

Meta publishes reports in its [transparency center](#) on government requests to remove content based on local law. It also publishes separate reports on governmental requests for user data. There is separate reporting on enforcement against the Community Standards. However, none of these reports differentiate data on content removed for violating *content policies* following a government request for review. Current transparency data on government removal requests underrepresent the full extent of interactions between Meta and law enforcement on content removals. By focusing on the action Meta takes (removal for violating local law), reporting on government requests excludes all reports received from law enforcement that result in removal for violating content policies. Content reported by law enforcement that violates both local law and the content policies is not included. For this reason, the Board submitted a freedom of information request to the Metropolitan Police to understand more fully the issues in this case.

Meta has claimed that transparency around government removal requests based on content policies is of limited use, since governments can (and do) also use in-product reporting tools. These tools do not distinguish between government requests and those made by other users.

This case demonstrates the level of privileged access law enforcement has to Meta's internal enforcement teams, evidenced by correspondence the Board has seen, and how certain policies rely on interaction with third parties, such as law enforcement, to be enforced. The way this relationship works for escalation-only policies, as in this case, brings into question Meta's ability to independently assess government actors' conclusions that lack detailed evidence.

The Board acknowledges Meta has made progress in relation to transparency reporting since the Board's first decisions addressing this topic. This includes conducting a scoping exercise on measuring content removed under the Community Standards following government

on measuring content removed under the Community Standards following government requests, and contributing to Lumen, a Berkman Klein Center for Internet & Society research project on government removal requests. Further transparency efforts in this area will be immensely valuable to public discussion on the implications of the interactions between governments and social-media companies.

#### b. Intake process for law enforcement requests

Although Meta has disclosed publicly [how it responds to government requests for takedowns based on local law violations](#), the channels through which governments can request review for violations of Meta's content policies remain opaque.

This case demonstrates that there are significant flaws in Meta's system governing law enforcement requests, where these requests are not based on local law and are made outside of its in-product reporting tools (i.e., the functions all regular users have access to for flagging or reporting content). In the "Shared Al Jazeera post" decision, the Board recommended that Meta formalize a transparent process on how it receives and responds to all government requests for content removals. Law enforcement agencies make requests by various communications channels, making the standardization and centralization of requests, and collecting data about them, challenging. The current intake system, where Meta fills in the intake form, focuses solely on prioritizing incoming requests. The system does not adequately ensure that third party requests meet minimum standards and does not allow for the accurate collection of data to enable the effects of this system to be properly monitored and audited. Some requests may refer to violations of Meta's Community Standards, others to violations of national law, and others to generally stated concerns about potential harms without connecting this to allegations of unlawful activity or violations of platform policies. Law enforcement is not asked to meet minimum criteria to fully contextualize and justify their requests, leading to unstructured, ad hoc, and inconsistent interactions with Meta. Minimum criteria might include, for example, an indication of which Meta policy law enforcement believes has been violated, why it has been violated, and a sufficiently detailed evidential basis for that conclusion.

#### c. Notification to users

In its [Q2 2022 Quarterly Update on the Oversight Board](#), Meta disclosed that it is improving notifications to users by specifically indicating when content was removed for violating the Community Standards after being reported by a government entity (implementing the

Board’s recommendation in the “Öcalan’s isolation” case decision). In this case, if those changes had been rolled out, all users who were subject to the 164 additional removals of content should have received notifications of this kind. Meta has acknowledged that only once it has set up the infrastructure required to collect more granular data from government requests, will it be able to design and test sending more detailed user notifications. The Board therefore agrees with Meta that this work is dependent on tracking and providing more information on government requests, which can then be published in Meta’s public transparency reporting.

## **8.2 Compliance with Meta’s values**

The Board finds that removing the content did not comply with Meta’s values. This case demonstrates the challenges that Meta faces in balancing the values of “Voice” and “Safety,” when seeking to address a high number of potential veiled threats in art, at a global scale and in a timely manner. However, Meta claims that “Voice” is its paramount value. As the Board stated in its “Wampum belt” decision, art is a particularly important and powerful expression of “Voice,” especially for people from marginalized groups creating art informed by their experiences. Meta did not have sufficient information to conclude that this content posed a risk to “Safety” that justified displacing “Voice.”

The Board is concerned that in light of doubts as to whether the content credibly risked harm, Meta describes that it errs on the side of “Safety” rather than “Voice.” Where doubt arises, as in this case, from lack of specificity in the information law enforcement has provided about a piece of artistic expression, such an approach is inconsistent with Meta’s self-described values. The Board recognizes the importance of keeping people safe from violence, and that this is especially important for communities disproportionately impacted by such violence. The Board is also mindful that decisions about alleged threats often must be made quickly, without the benefit of extended reflection. However, a presumption against “Voice” may have a disproportionate impact on the voices of marginalized people. In practice, it may also significantly increase the power and leverage of law enforcement, who may claim knowledge that is difficult to verify through other sources.

## **8.3 Compliance with Meta’s human rights responsibilities**

The Board concludes that Meta did not meet its human rights responsibilities as a business in deciding to remove this post.

The right to freedom of expression is guaranteed to all people without discrimination (Article 19, para. 2, ICCPR; Article 2, para. 1, ICCPR). This case further engages the rights of persons belonging to ethnic minorities to enjoy, in community with other members of their group, their own culture (Article 27, ICCPR) and the right to participate in cultural life (Article 15, ICESCR). The right of access to remedy is a key component of international human rights law (Article 2, para. 3, ICCPR; General Comment No. 31), and remedy is a third pillar of the UNGPs and a focus area in Meta's corporate human rights policy.

### *Freedom of expression (Article 19 ICCPR; Article 5 ICERD)*

Article 19 of the ICCPR gives specific mention to protecting expression "in the form of art." International human rights standards reinforce the importance of artistic expression ( General Comment 34, para. 11; *Shin v. Republic of Korea*, Human Rights Committee, communication No. 926/2000). The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) protects the exercise of the right to freedom of expression without discrimination based on race (Article 5). The Committee on the Elimination of Racial Discrimination has emphasized the importance of freedom of expression to assist "vulnerable groups in redressing the balance of power among the components of society" and to offer "alternative views and counterpoints" in discussions (CERD Committee, General Recommendation 35, para. 29). Drill music offers young people, and particularly young Black people, a means of creative expression.

Art is often political, and international standards recognize its unique and powerful role in challenging the status quo (UN Special Rapporteur in the field of cultural rights, A/HRC/23/34, at paras 3-4). The internet, and social media platforms such as Facebook and Instagram in particular, have special value to artists in helping them to reach new and larger audiences. Artists' livelihoods, and their social and economic rights, may depend on access to social platforms that dominate the internet. Drill music relies on boastful claims to violence to drive the commercial success of artists on social media. Such claims and performances are expected as part of the genre. As a result of Meta's actions in this case, Chinx (OS) has been removed from Instagram permanently, which is likely to have a significant impact on his ability to reach his audience and find commercial success.

ICCPR Article 19 requires that where 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). The UN Special Rapporteur on freedom of expression has

Article 19, para. 3, ICCPR). The UN Special Rapporteur on Freedom of Expression has encouraged social media companies to be guided by these principles when moderating online expression, mindful that regulation of expression at scale by private companies may give rise to concerns particular to that context (A/HRC/38/35, paras. 45 and 70). The Board has employed the three-part test based on Article 19 of the ICCPR in all its decisions to date.

### *I. Legality (clarity and accessibility of the rules)*

The principle of legality requires laws limiting expression to be clear and accessible, so people understand what is permitted and what is not. Furthermore, it requires those laws to be specific, to ensure that those charged with their enforcement are not given excessive discretion (General Comment 34, para. 25). The Board applies these principles to assess the clarity and accessibility of Meta's content rules, and the guidance reviewers have to make fair decisions based on those rules.

The Board reiterates its previously stated concerns that the relationship between the Instagram Community Guidelines and Facebook Community Standards is unclear. In August 2022, Meta committed to implement the Board's prior recommendations in this area ([Q2 2022 Quarterly Update on the Oversight Board](#)) and align the Community Standards and Guidelines in the long term.

The differences between the publicly facing Violence and Incitement Community Standard and Meta's internal Implementation Standards is also a concern. Meta uses "signals" to determine whether content contains a veiled threat. The "signals" were added to the public-facing Community Standards as a result of the Board's prior recommendations. However, the Community Standards do not explain that Meta divides these into primary and secondary signals, or that both a primary and secondary signal is required to find a policy violation. Making this clear will be useful to those raising complaints about content on the platform, including trusted third parties and law enforcement. Clarity about signals is especially important, as the secondary signal validates the risk of harm resulting from the content and leads to the removal decision. Third parties who provide a primary signal without a secondary signal may be confused if the content they report is not actioned.

### *II. Legitimate aim*

Restrictions on freedom of expression must pursue a legitimate aim. The Violence and Incitement Community Standard exists in part to prevent offline harm. This policy therefore



serves the legitimate aim of the protection of the rights of others (the rights to life and security of person of those targeted by the post).

### *III. Necessity and proportionality*

The Board finds the content removal was not necessary to achieve the aim of the policy.

The principle of 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 that might achieve their protective function; they must be proportionate to the interest to be protected" (General Comment 34, para. 34). The form of expression at issue, such as expression in the form of art, must be taken into consideration (General Comment 34, para. 34).

The UN Special Rapporteur on freedom of expression has observed the difficulties artists face in their use of social media, as such expression tends to have complex characteristics and can easily fall foul of platforms' rules, with inadequate remedial mechanisms (A/HRC/44/49/Add.2, at para. 44 - 46). This affirms broader observations the Special Rapporteur has made on deficiencies in contextual content moderation by platforms, including on issues requiring historical or cultural nuance (A/HRC/38/35, at para. 29).

The complexity of artistic expression was emphasized by the UN Special Rapporteur in the field of cultural rights (A/HRC/23/34, at para. 37):

*An artwork differs from non-fictional statements, as it provides a far wider scope for assigning multiple meanings: assumptions about the message carried by an artwork are therefore extremely difficult to prove, and interpretations given to an artwork do not necessarily coincide with the author's intended meaning. Artistic expressions and creations do not always carry, and should not be reduced to carrying, a specific message or information. In addition, the resort to fiction and the imaginary must be understood and respected as a crucial element of the freedom indispensable for creative activities and artistic expressions: representations of the real must not be confused with the real... Hence, artists should be able to explore the darker side of humanity, and to represent crimes... without being accused of promoting these.*

The Special Rapporteur's observations do not exclude the possibility that art can be intended

to cause harm and may achieve that objective. For a company in Meta's position, making these assessments quickly, at scale, and globally is challenging. Meta's creation of "escalation only" policies that require a fuller contextual analysis to remove content shows respect for the principle of necessity. Meta's human rights responsibilities require Meta to prevent and mitigate risks to the right to life, and the right to security of person, for those who may be put in danger by posts that contain veiled threats. However, for reasons stated in Section 8.1 of this decision, that analysis requires a closer examination of causality and must be more nuanced in its assessment of art in order to meet the requirements of necessity. As the Board has not seen sufficient evidence to show a credible threat in this case, removal was not necessary.

In this respect, the Violence and Incitement policy uses terminology that may be read to permit excessive removals of content. A "potential" threat, or content that "could" result in violence somewhere down the line, such as a taunt, was too broad in this decision to satisfy the requirements of necessity. In prior cases, the Board has not required that the risk of future violence be imminent for removal to be allowed (see, for example, the "Knin cartoon" case), since Meta's human rights responsibilities may differ from those of a state imposing criminal or civil penalties (see, for example, the "South Africa slurs" case). The Board has, however, required a more substantial evidential basis that a threat was present and credible than appears in this case (see, for example, the "Protest in India against France" case).

#### *Non-discrimination and access to remedy (Article 2(1), ICCPR)*

The Human Rights Committee has made clear that any restrictions on expression must respect the principle of non-discrimination (General Comment No. 34, at para. 32). This principle informs the Board's interpretation of Meta's human rights responsibilities (UN Special Rapporteur on freedom of expression, A/HRC/38/35, para. 48).

In its public comment, the Digital Rights Foundation argued that while some have portrayed drill music as a rallying call for gang-violence, it serves as a medium for youth, in particular Black and Brown youth, to express their discontent with a system that perpetuates discrimination and exclusion (PC-10618). JUSTICE, in its report "Tackling Racial Injustice: Children and the Youth Justice System," cites law enforcement's misuse of drill music to secure convictions as an example of systemic racism. As the Board learned through its freedom of information request, all 286 requests the Metropolitan Police made to social media companies and streaming services to review music content from June 1, 2021 to May

31 2022 involved drill music. 255 of those requests resulted in content being removed from

in 2022 involved anti-music. 200 of these requests resulted in content being removed from the platform. 21 of those 286 requests related to Meta's platforms, and 14 of those requests were actioned through removals. As outlined above, one request can result in multiple content removals. This intensive focus on one music genre among many that include reference to violence raises serious concerns of potential over-policing of certain communities. It is beyond the Board's purview to say whether these requests represent sound police work, but it does fall to the Board to assess how Meta can honor its values and human rights responsibilities when it responds to such requests. Accordingly, and as described below, Meta's response to law enforcement requests must, in addition to meeting minimal evidential requirements, be sufficiently systematized, audited, and transparent, to affected users and the broader public, to enable the company, the Board and others to assess the degree to which Meta is living up to its values and meeting its human rights responsibilities.

Where a government actor is implicated in interference with an individual's expression, as in this case, due process and transparency are key to empower the affected users to assert their rights and even challenge that government actor. Meta should consider whether its processes as they currently stand enable or obstruct this. The company cannot allow its cooperation with law enforcement to be opaque to the point that it creates a barrier to users accessing remedies for potential human rights violations.

It is also important that Meta provides its users with adequate access to remedy for the content decisions it takes that impact users' rights. The UN Special Rapporteur on freedom of expression has addressed the responsibilities of social media companies in relation to artistic expression (A/HRC/44/49/Add.2, at para. 41 onwards). Their observations on access to remedy of female artists is relevant to the situation of Black artists in the United Kingdom:

*Artists reportedly have experienced shutdowns of personal and professional Facebook and Twitter pages... Violations of vague community guidelines can leave artists without "counter-notice" procedures allowing challenges to removals of their art. The lack of procedural safeguards and access to remedies for users leaves artists without access to a platform to display their art, and without viewership to enjoy their art. In some cases, States work with companies to control what kinds of content is available online. This dangerous collaboration has the effect of silencing artists and preventing individuals... from receiving art as expression.*

The UN Special Rapporteur on freedom of expression has stated that the process of

remediation for social media companies "should include a transparent and accessible process for appealing platform decisions, with companies providing a reasoned response that should also be publicly accessible" (A/74/486, para 53).

Even though the content under review in this case was posted by an Instagram account not belonging to Chinx (OS), the artist had posted the same video to his own account. This was removed at the same time as the content in this case, resulting in his account being first disabled, and then deleted. This shows how collaboration between law enforcement and Meta can result in significantly limiting the expression of artists, denying their audience access to art on the platform. As the Board's freedom of information request confirms, this collaboration specifically and exclusively targets drill artists, who are mostly young Black men.

The Board requested that Meta refer the removal of content from Chinx (OS)'s account for review, so that it could be examined alongside the content in this case. That was not technically possible due to Meta deleting the account. This raises significant concerns about the right to remedy, as does the fact that users cannot appeal decisions taken "at escalation" to the Oversight Board. This includes significant and difficult decisions concerning "additional context to enforce" policies, which are only decided "at escalation." It also includes all government requests for removals (besides "in-product tool" usage), including lawful content, that are eligible for review and within scope under the Oversight Board Charter. The latter is especially concerning for individuals who belong to discriminated-against groups, who are likely to experience further barriers to accessing justice as a result of Meta's product design choices.

These concerns about the right to remedy add to those raised during the Board's work in the upcoming policy advisory opinion on cross-check. Cross-check is the system Meta uses to reduce enforcement errors by providing additional layers of human review for certain posts initially identified as breaking its rules, before removing content. Meta has told the Board that, between May and June 2022, around a third of content decisions in the cross-check system could not be appealed by users to the Board. The Board will address this further in the cross-check policy advisory opinion.

#### **8.4 Identical content with parallel context**

The Board notes that this content was added to the Violence and Incitement Media Matching

Service bank, which resulted in automated removals of matching content and potentially additional account-level actions on other accounts. Following this decision, Meta should ensure the content is removed from this bank, restore identical content it has wrongly removed where possible, and reverse any strikes or account-level penalties. It should remove any bar on Chinx (OS) re-establishing an account on Instagram or Facebook.

## **9. Oversight Board decision**

The Oversight Board overturns Meta's decision to take down the content, requiring the post to be restored.

## **10. Policy advisory statement**

### A. Content Policy

1. Meta's description of its value of "Voice" should be updated to reflect the importance of artistic and creative expression. The Board will consider this recommendation implemented when Meta's values have been updated.
2. Meta should clarify that for content to be removed as a "veiled threat" under the Violence and Incitement Community Standard, one primary and one secondary signal is required. The list of signals should be divided between primary and secondary signals, in line with the internal Implementation Standards. This will make Meta's content policy in this area easier to understand, particularly for those reporting content as potentially violating. The Board will consider this recommendation implemented when the language in the Violence and Incitement Community Standard has been updated.

### B. Enforcement

3. Meta should provide users with the opportunity to appeal to the Oversight Board for any decisions made through Meta's internal escalation process, including decisions to remove content and to leave content up. This is necessary to provide the possibility of access to remedy to the Board and to enable the Board to receive appeals for "escalation-only" enforcement decisions. This should also include appeals against removals made for Community Standard violations as a result of "trusted flagger" or government actor reports made outside of in-product tools. The Board will consider this implemented when it sees user appeals coming from decisions made on escalation and when Meta shares data with the

appeals coming from decisions made on escalation and which Meta shares data with the

Board showing that for 100% of eligible escalation decisions, users are receiving reference IDs to initiate appeals.

4. Meta should implement and ensure a globally consistent approach to receive requests for content removals (outside of in-product reporting tools) from state actors by creating a standardized intake form asking for minimum criteria, for example, the violated policy line, why it has been violated, and a detailed evidential basis for that conclusion, before any such requests are actioned by Meta internally. This contributes to ensuring more organized information collection for transparency reporting purposes. The Board will consider this implemented when Meta discloses the internal guidelines that outline the standardized intake system to the Board and in the transparency center.

5. Meta should mark and preserve any accounts and content that were penalized or disabled for posting content that is subject to an open investigation by the Board. This prevents those accounts from being permanently deleted when the Board may wish to request content is referred for decision or to ensure its decisions can apply to all identical content with parallel context that may have been wrongfully removed. The Board will consider this implemented when Board decisions are applicable to the aforementioned entities and Meta discloses the number of said entities affected for each Board decision.

### C. Transparency

6. Meta should create a section in its Transparency Center, alongside its “Community Standards Enforcement Report” and “Legal Requests for Content Restrictions Report,” to report on state actor requests to review content for Community Standard violations. It should include details on the number of review and removal requests by country and government agency, and the numbers of rejections by Meta. This is necessary to improve transparency. The Board will consider this implemented when Meta publishes a separate section in its “Community Standards Enforcement Report” on requests from state actors that led to removal for content policy violations.

7. Meta should regularly review the data on its content moderation decisions prompted by state actor content review requests to assess for any systemic biases. Meta should create a formal feedback loop to fix any biases and/or outsized impacts stemming from its decisions on government content takedowns. The Board will consider this recommendation implemented when Meta regularly publishes the general insights derived from these audits

and the actions taken to mitigate systemic biases.

**\*Procedural note:**

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

For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by an independent research institute headquartered at the University of Gothenburg, which draws on a team of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world. The Board was also 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.