HOMOPHOBIC VIOLENCE IN WEST AFRICA

2024-041-FB-UA

Summary

The Oversight Board is seriously concerned about Meta's failure to take down a video showing two bleeding men who appear to have been beaten for allegedly being gay. The content was posted in Nigeria, which criminalizes same-sex relationships. In overturning the company's original decision, the Board notes that by leaving the video on Facebook for five months, there was a risk of immediate harm to the men by exposing their identities, given the hostile environment for LGBTQIA+ people in Nigeria. Such damage is immediate and impossible to undo. The content, which shared and mocked violence and discrimination, violated four different Community Standards, was reported multiple times and reviewed by three human moderators. This case reveals systemic failings around enforcement. The Board's recommendations include a call for Meta to assess enforcement of the relevant rule under the Coordinating Harm and Promoting Crime Community Standard. They also address the failings likely to have arisen from Meta identifying the wrong language being spoken in the video and how the company handles languages it does not support for at-scale content review.

About the Case

A Facebook user in Nigeria posted a video that shows two bleeding men who look like they could have been tied up and beaten. People around the frightened men ask them questions in one of Nigeria's major languages, Igbo. In response, one of the men responds with his name and explains, seemingly under coercion, that he was beaten for having sex with another man. The user who posted this content included an English caption mocking the men, stating they were caught having sex and that this is "funny" because they are married.

The video was viewed more than 3.6 million times. Between December 2023 when it was posted and February 2024, 92 users reported the content, the majority for violence and incitement or hate speech. Two human reviewers decided it did not violate any of the Community Standards so should remain on Facebook. One user appealed to Meta but, after another human review, the company decided again there were no violations. The user then appealed to the Board. After the Board brought the case to Meta's attention, the company removed the post under its Coordinating Harm and Promoting Crime policy.

Nigeria criminalizes same-sex relationships, with LGBTQIA+ people facing discrimination and severe restrictions on their human rights.

Key Findings

The Board finds the content violated four separate Community Standards, including the Coordinating Harm and Promoting Crime rule that does not allow individuals alleged to be members of an outing-risk group to be identified. The man's admission in the video of having sex with another man is forced, while the caption explicitly alleges the men are gay. The content also broke rules on hate speech, bullying and harassment, and violent and graphic content.

There are two rules on outing under the Coordinating Harm and Promoting Crime policy. The first is relevant here and applied at-scale. It prohibits: "outing: exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group." There is a similar rule applied only when content is escalated to Meta's experts. The Board is concerned that Meta does not adequately explain the differences between the two outing rules and that the rule applied at-scale does not publicly state that "outing" applies to identifying people as LGBTQIA+ in countries where there is higher risk of offline harm, such as Nigeria. Currently, this information is only available in internal guidance. This ambiguity could lead to confusion, preventing users from complying with the rules, and hindering people targeted by such abusive content to get these posts removed. Meta needs to update its public rule and provide examples of outing-risk groups.

This content was left up for about five months, despite breaking four different rules and featuring violence and discrimination. Human moderators reviewed the content and failed to identify that it broke the rules. With the video left up, the odds of

someone identifying the men and of the post encouraging users to harm other LGBTQIA+ people in Nigeria increased. The video was eventually taken down but by this time, it had gone viral. Even after it was removed, the Board's research shows there were still sequences of the same video remaining on Facebook.

When the Board asked Meta about its enforcement actions, the company admitted two errors. First, its automated systems that detect language identified the content as English, before sending it to human review, while Meta's teams then misidentified the language spoken in the video as Swahili. The correct language is Igbo, spoken by millions in Nigeria, but this is not supported by Meta for content moderation at-scale. If the language is not supported, as in this case, then content is sent instead to human reviewers who work across multiple languages and rely on translations provided by Meta's technologies. This raises concerns about how content in unsupported languages is treated, the choice of languages the company supports for at-scale review and the accuracy of translations provided to reviewers working across multiple languages.

The Oversight Board's Decision

The Oversight Board overturns Meta's original decision to leave up the content.

The Board recommends that Meta:

- Update the Coordinating Harm and Promoting Crime Community Standard's at-scale prohibition on "outing" to include illustrative examples of "outing-risk groups," including LGBTQIA+ people in countries where same-sex relationships are forbidden and/or such disclosures create significant safety risks.
- Conduct an assessment of the enforcement accuracy of the at-scale prohibition on exposing the identity or locations of anyone alleged to be a member of an outing-risk group, under the Coordinating Harm and Promoting Crime Community Standard.
- Ensure its language detection systems precisely identify content in unsupported languages and provide accurate translations of such content to language-agnostic reviewers.
- Ensure that content containing an unsupported language, even if this is combined with supported languages, is routed to agnostic review. This includes

giving reviewers the option to re-route content containing an unsupported language to agnostic review.

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

Full Case Decision

1. Case Description and Background

In December 2023, a Facebook user in Nigeria posted a video showing two men who are clearly visible and appear to have been beaten. They are sitting on the ground, near a pole and a rope, suggesting they may have been tied up, and are heavily bleeding. Several people ask the men questions in Igbo, one of the major languages in Nigeria. One of the men responds with his name and explains, seemingly under coercion, that he was beaten because he was having sex with another man. Both men appear frightened and one of them is kicked by a bystander. The user who posted the video added a caption in English mocking the men, saying that they were caught having sex and that this is "funny" because they are both married.

The content was viewed over 3.6 million times, received about 9,000 reactions and 8,000 comments, and was shared about 5,000 times. Between December 2023 and February 2024, 92 users reported the content 112 times, the majority of these reports under Meta's Violence and Incitement and Hate Speech policies. Several of the reports were reviewed by two human moderators who decided the content did not violate any of the Community Standards and therefore should remain on Facebook. One of the users then appealed Meta's decision to keep the content up. Following another human review, the company again decided the content did not violate any of its rules. The user then appealed to the Board. After the Board brought the case to Meta's attention, in May 2024, the company reviewed the post under its Coordinating Harm and Promoting Crime policy, removing it from Facebook. Following Meta's removal of the original video, upon further research, the Board identified multiple instances of the same video left on the platform dating back to December 2023, including in Facebook Groups. After the Board flagged instances of the same video remaining on the platform, Meta removed them and added the video to a Media Matching Service (MMS) bank, which automatically identifies and removes content that has already

been classified as violating. While this type of violation can result in a standard strike against the user who posted the content, Meta did not apply it in this case because the video was posted more than 90 days before any enforcement action was taken. Meta's policy states that it does not apply standard strikes to accounts of users whose content violations are older than 90 days.

The Board considered the following context in reaching its decision in this case:

LGBTQIA+ people in Nigeria and in several other parts of the world face violence, torture, imprisonment and even death because of their sexual orientation or gender identity, with anti-LGBTQIA+ sentiment on the increase (see public comment by Outright International, PC-29658). Discrimination against people based on their sexual orientation or gender identity limits everyday life, impacting basic human rights and freedoms. Amnesty International reports that in Africa, 31 countries criminalize same-sex relationships. Sanctions range from imprisonment to corporal punishment. Nigeria's Same Sex Marriage Prohibition Act not only criminalizes same-sex relationships but also prohibits public displays of affection and restricts the work of organizations defending LGBTQIA+ rights. In addition, colonial-era and other morality laws on sodomy, adultery and indecency are still enforced to restrict the rights of LGBTQIA+ people, with devastating outcomes.

In a 2024 report, the UN Independent Expert on protection against violence and discrimination based on sexual orientation and gender identity emphasized: "States in all regions of the world have enforced existing laws and policies or imposed new, and sometimes extreme, measures to curb freedoms of expression, peaceful assembly and association specifically targeting people based on sexual orientation and gender identity," (Report A/HRC/56/49, July 2024, at para. 2).

Activists and organizations supporting LGBTQIA+ communities can be subject to legal restrictions, harassment, arbitrary arrests, police raids and shutdowns, with threats of violence discouraging public support for LGBTQIA+ rights (see public comment by Pan-African Human Rights Defenders Network, PC-29657). Human rights organizations can struggle to document cases of abuse and discrimination due to fear of retaliation from public authorities and non-state actors, such as vigilantes and militias. Journalists covering LGBTQIA+ issues can also be targeted.

Social media is an essential tool for human rights organizations documenting LGBTQIA+ rights violations and abuses, and advocating for stronger protections. People share videos, testimonials and reports to raise awareness and advocate for governments to uphold human rights standards (see public comment by Human Rights Watch, PC-29659). Additionally, platforms can act as information hubs, providing people with updates on legal developments as well as access to legal support. Independent research commissioned by the Board indicates that social media platforms play a crucial role for LGBTQIA+ people in countries with restrictive legal frameworks. The research indicates that Facebook, for example, allows users to connect, including anonymously and in closed groups, to share resources in a safer environment than offline spaces.

Experts consulted by the Board noted that state authorities in some African countries also use social media to monitor and curtail the activities of users posting LGBTQIA+ content. The experts reported that in Nigeria, authorities have restricted access to online content about LGBTQIA+ issues. According to Freedom House, Nigeria has introduced legislation to regulate social media platforms more broadly, which could impact LGBTQIA+ rights online. Similarly, Access Now – a digital rights organization – reports that cybercrime laws in Ghana provide authorities with the ability to issue takedown requests or content bans that could restrict public discourse around LGBTQIA+ issues, and block documentation of human rights abuses as well as vital information for the community.

Non-state actors, including vigilantes, also target LGBTQIA+ people with physical assaults, mob violence, public humiliation and ostracization. For example, in August 2024, a transgender Tik-Tok user known as “Abuja Area Mama” was found dead after allegedly being beaten to death in Nigeria’s capital Abuja. LGBTQIA+ people can be targets of blackmail by other community members who discover their sexual orientation or gender identity. According to Human Rights Watch, Nigeria’s legal framework encourages violence against LGBTQIA+ people, creating an environment of impunity for those carrying out this violence.

2. User Submissions

In their statement to the Board, the user who reported the content claimed the men in the video were beaten solely for being gay. The user stated that, by not removing the video, Meta is allowing its platform to become a breeding ground for hate and

homophobia and that if the video was of an incident in a Western country, it would have been removed.

3. Meta's Content Policies and Submissions

I. Meta's Content Policies

Coordinating Harm and Promoting Crime policy

The Coordinating Harm and Promoting Crime policy aims to "prevent and disrupt offline harm and copycat behavior" by prohibiting "facilitating, organizing, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals." Two policy lines in the Community Standards address "outing." The first is applied at-scale, and the second requires "additional context to enforce" (which means that the policy line is only enforced following escalation). The first policy line applies to this case. It specifically prohibits "outing: exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group." This policy line does not explain which groups are considered to be "outing-risk groups." The second policy line, which is only enforced on escalation and was not applied in this case, also prohibits "outing: exposing the identity of a person and putting them at risk of harm" for a specific list of vulnerable groups, including LGBTQIA+ members, unveiled women, activists and prisoners of war.

According to Meta's internal guidance to content reviewers on the first policy line, identity exposure can occur through the use of personal information such as a person's name or image. Meta's internal guidelines list "outing-risk groups," including LGBTQIA+ people in countries where the affiliation to a group may carry an associated risk to the personal safety of its members. It also provides that "outing" must be involuntary: a person cannot out themselves (for example, by declaring themselves to be a member of an outing-risk group).

Violent and Graphic Content policy

The Violent and Graphic Content policy provides that certain disturbing imagery of people will be placed behind a warning screen. This includes: "Imagery depicting acts

of brutality (e.g., acts of violence or lethal threats on forcibly restrained subjects) committed against a person or group of people." However, if such content is accompanied by "sadistic remarks," the post will be removed. Sadistic remarks are defined in the public-facing rules as "commentary – such as captions or comments – expressing joy or pleasure from the suffering or humiliation of people or animals."

Bullying and Harassment policy

The Bullying and Harassment Community Standard aims to prevent individuals being targeted on Meta's platforms through threats and different forms of malicious contact, and that such behaviour "prevents people from feeling safe and respected." The policy prohibits content that targets people with "celebration or mocking of [their] death or medical condition." Meta's internal guidelines explain that medical condition includes a serious disease, illness or injury.

Hate Speech policy

Meta's Hate Speech policy rationale defines hate speech as a direct attack against people on the basis of protected characteristics, including sexual orientation. It prohibits content targeting people in written or visual form, such as: "Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image." Meta's internal guidelines define hate crimes as a criminal act "committed with a prejudiced motive targeting people based on their [protected characteristics]."

II. Meta's Submissions

After the Board selected this case, Meta found that the content violated the Coordinating Harm and Promoting Crime policy for identifying alleged members of an "outing-risk group" in a country where the affiliation with such a group may carry an associated risk to the personal safety of its members. Meta noted that the user's caption alleged that the men were gay, and the admission from one of the men in the video was potentially coerced, demonstrating that the "outing," by exposing their identity, was involuntary.

Meta recognized that reviewers were wrong in finding that the post did not violate any Community Standards and investigated why these errors occurred. In this case, it

appears reviewers only focused on the Bullying and Harassment policy related to "claims about romantic involvement, sexual orientation or gender identity" against private adults and found that the policy had not been violated, without considering other potential violations. This policy requires that the name and face of the user reporting the content match the person depicted in the content for it to be removed. Since the users reporting the content in this case were not depicted in the content, the reviewers assessed it as non-violating. Following its investigation, Meta's human review teams took additional steps to improve accuracy in applying the Coordinating Harm and Promoting Crime policy, sending policy reminders and conducting knowledge tests on the outing of high-risk individuals policy.

In response to the Board's questions, Meta confirmed the post also violated three other Community Standards.

The post violated the Violent and Graphic Content policy as the video included sadistic remarks about a depicted act of "brutality," with the men subjected to excessive force while in a position of being dominated. Without the sadistic remarks, the content would only have been marked as disturbing under this policy. It violated the Bullying and Harassment policy because the caption mocks both men by referring to their situation as "funny" while showing their serious injuries. Lastly, it violated the Hate Speech Community Standard, since the caption mocked victims of a hate crime, particularly the assault and battery motivated by prejudice against two men based on their perceived sexual orientation.

In response to Board questions, Meta confirmed that it conducted additional investigations that led to removals of other instances of the same video. The video was added to Meta's MMS banks to prevent future uploads of the content.

Meta also informed the Board it leverages its language detection and machine translation systems to provide support for content in Igbo through agnostic review at-scale. Meta has a few Igbo speakers who provide language expertise and content review for Igbo upon escalation (not at-scale). The company requires its human reviewers to have proficiency in English and "their relevant market language." Before confirming that the language spoken in the video was Igbo, Meta misidentified the language of the video as Swahili in its engagement with the Board. Finally, Meta explained that because the user's caption for the video was in English, the company's

automated systems identified the language of the content as English, routing it to English-speaking human reviewers.

The Board asked Meta 24 questions on enforcement of the Coordinating Harm and Promoting Crime Community Standard and other content policies, Meta's enforcement actions in Nigeria, Meta's detection of content languages and human review assignments, as well as governmental requests and mitigation measures the company has undertaken to prevent harm. Meta responded to all the questions.

4. Public Comments

The Oversight Board received seven public comments that met the terms for submission. Four of the comments were submitted from the United States and Canada, two from Sub-Saharan Africa and one from Europe. To read public comments submitted with consent to publish, click here.

The submissions covered the following themes: violence against LGBTQIA+ people in West Africa by state and non-state actors, and risks associated with the exposure of people's sexual orientation and/or gender identity; the impact of the criminalization of same-sex relationships on LGBTQIA+ people; the impact of this criminalization and other local laws in Nigeria, and West Africa more broadly, on the work conducted by human rights organizations, advocacy groups and journalists; and the importance of Meta's platforms, and social media more broadly, to communication, mobilization and awareness-raising among LGBTQIA+ people in Nigeria and West Africa.

5. Oversight Board Analysis

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

5.1 Compliance With Meta's Content Policies

I. Content Rules

The Board finds the content violates four Community Standards: Coordinating Harm and Promoting Crime, Hate Speech, Violent and Graphic Content, and Bullying and Harassment.

The content violates the Coordinating Harm and Promoting Crime policy prohibiting identifying individuals alleged to be members of an outing-risk group. The Board agrees with Meta that the video exposes the identity of the two men against their will, as they appear to have been beaten and are visibly frightened. The admission of one to having sex with another man is therefore forced and involuntary. Additionally, the caption to the video explicitly alleges the men are gay.

It also violates the Hate Speech policy prohibiting content mocking victims of hate crimes. The video captures the aftermath of violence against two men, which continues in the video, with their injuries clearly visible. One of the men explains they were beaten because they had sex with each other, with the video's caption further demonstrating the criminal battery and assault was motivated by the men's perceived sexual orientation. Because the post's caption ridicules the victims of this hate crime by saying it is "funny" they are apparently married, the Board believes it meets Meta's definition of "mocking."

The post's caption also violates the Bullying and Harassment policy, for mocking their visible injuries (a "medical condition") by referring to the situation as "funny."

Finally, the Board finds that the content violates the Violent and Graphic Content Community Standard too, since it includes "sadistic remarks" made about acts of brutality against the two men in a context of suffering and humiliation. In itself, and without other policy violations being present, this would warrant the application of a warning screen. However, as the caption contains "sadistic remarks" ridiculing the acts of violence and assault against the men, the policy requires content removal.

II. Enforcement Action

The Board is especially concerned that content depicting such severe violence and discrimination, and violating four Community Standards, was left up for about five months, and that sequences of the same video remained on the platform even after the original video was removed. After it was posted in December 2023, the video was reported 112 times by 92 different users, by which time this single instance had amassed millions of views and thousands of reactions. Three human moderators independently reviewed the reports and subsequent appeals. All three concluded there were no violations, seemingly because they were not reviewing the posts against all Community Standards. Additionally, these reviewers may not have been familiar with the Igbo language or able to perform an agnostic review, given that Meta's automated systems wrongly identified the language of the content as English and routed it to English-speaking reviewers.

5.2 Compliance With Meta's Human Rights Responsibilities

The Board finds that leaving the content on the platform was not consistent with Meta's human rights responsibilities in light of the UN Guiding Principles on Business and Human Rights (UNGPs). In 2021, Meta announced its Corporate Human Rights Policy, in which the company reaffirmed its commitment to respecting human rights in accordance with the UNGPs. Under Guiding 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" even if they have not contributed to those impacts.

In interpreting the UNGPs, the Board has drawn from the UN Special Rapporteur on Freedom of Expression and Opinion's recommendation that social media companies should consider the global freedom of expression standards set forth in the International Covenant on Civil and Political Rights (ICCPR) Articles 19 and 20, (see paras. 44-48 of the 2018 report of the UN Special Rapporteur on freedom of expression, A/HRC/38/35 and para. 41 of the 2019 report of the UN Special Rapporteur on freedom of expression, A/74/486).

Article 20, para. 2 of the ICCPR provides that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence is to

be prohibited by law." This prohibition is "fully compatible with the right to freedom of expression as contained in article 19 [ICCPR], the exercise of which carries with it special duties and responsibilities," (General Comment No. 11, (1983), para. 2). The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence is an important road map for interpreting Article 20, para 2 (A/HRC/22/17/Add.4, 2013, para. 29). It sets out six relevant factors for states to determine whether to prohibit speech: "Context of statement; speaker's status; intent to incite the audience against the target group; content of statement; extent of dissemination, and likelihood of harm, including imminence." The Board has been using these factors to determine the necessity and proportionality of speech restrictions by Meta. In this case, the Board is considering the same factors when assessing whether Meta should remove the content given its human rights responsibilities.

The Board finds that Meta's original decision to leave the content on the platform created a risk of immediate harm to the men in the video, thereby warranting removal. In countries like Nigeria, where societal attitudes and the criminalization of same-sex relationships fuel homophobic violence, LGBTQIA+ people who are outed online may be subjected to offline violence and discrimination. Meta's failures to take timely action on this video, allowing it to be shared so extensively, likely contributed to that hostile environment, creating risks for others (see public comment by Human Rights Watch, PC-29659). The Board also notes that the post amassed a great number of views (over 3.6 million), which increased the odds of someone identifying the men depicted in the video and of the post instigating users to harm LGBTQIA+ people, more broadly. Moreover, the Board highlights the sadistic remarks accompanying the video, which indicate the user's intention of exposing and humiliating the men, inciting others to discriminate and harm them. The great number of reactions (about 9,000), comments (about 8,000) and shares (about 5,000) indicate that the user managed to engage their audience, further increasing the likelihood of harm, both to the men depicted in the video and to LGBTQIA+ people in Nigeria.

Speech restrictions based on Article 20, para. 2 ICCPR should also meet ICCPR Article 19's three-part test (General Comment No. 34, para. 50). The analysis that follows finds that removal of the post was consistent with Article 19.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the ICCPR provides for broad protection of expression, including political expression and discussion of human rights, as well as expression that may be considered "deeply offensive," (General Comment No. 34, (2011), para. 11, see also para. 17 of the 2019 report of the UN Special Rapporteur on freedom of expression, A/74/486). 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. The Board does this both in relation to the individual content decision under review and what this says about Meta's broader approach to content governance. As the UN Special Rapporteur on freedom of expression has stated, although "companies do not have the obligations of Governments, their impact is of a sort that requires them to assess the same kind of questions about protecting their users' right to freedom of expression," (A/74/486, para. 41).

While the Board notes that multiple content policies are applicable to this case, its three-part analysis is focused on Meta's Coordinating Harm and Promoting Crime Community Standard, given this is the policy under which the company eventually removed the content.

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 that Meta's prohibition on "outing" individuals by exposing the identity or locations affiliated with anyone who is alleged to be a member of an outing-risk group is not sufficiently clear and accessible to users.

The Coordinating Harm and Promoting Crime Community Standard does not offer sufficient explanation for users to understand and distinguish between the two similar "outing" rules. The Board is particularly concerned that the Community Standard does not clearly explain that the at-scale rule prohibiting "outing" applies to identifying people as LGBTQIA+ in countries where the local context indicates higher risks of offline harm. Currently this information is only available in internal guidance to reviewers, making it impossible for users to know that persons alleged to belong to "at-risk" outing groups include LGBTQIA+ people in specific countries.

The Board is concerned that the ambiguity surrounding Meta's policies on content outing LGBTQIA+ individuals may result in user confusion and prevent them from complying with the platform's rules. It also creates obstacles to people targeted by abusive content who are seeking the removal of such posts. Meta should, therefore, update its Coordinating Harm and Promoting Crime policy line that prohibits "outing," and which the company enforces at-scale, to include illustrative examples of outing-risk groups, including LGBTQIA+ people in specific countries.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, which includes protecting the rights of others (Article 19, para. 3, ICCPR).

The Coordinating Harm and Promoting Crime policy serves the legitimate aim of "prevent[ing] and disrupt[ing] offline harm," including by protecting the rights of LGBTQIA+ people and those perceived as such in countries around the world where "outing" creates safety risks. Those rights include the right to non-discrimination (Articles 2 and 26, ICCPR), including in the exercise of their rights to freedom of expression and assembly (Articles 19 and 21, ICCPR), to privacy (Article 17, ICCPR), as well as to life (Articles 6, ICCPR), and liberty and security (Article 9 ICCPR).

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 Board finds that Meta's eventual decision to remove the content from the platform was necessary and proportionate. Research commissioned by the Board indicates that LGBTQIA+ people in Nigeria are continuously exposed to violence, arbitrary arrests, harassment, blackmail and discrimination, and risks of legal sanctions. The content itself depicts the aftermath of what appears to be corporal punishment for an alleged same-sex relationship. Under these circumstances, the Board determines that accurate enforcement of policies meant to protect LGBTQIA+ people is critical, especially in countries criminalizing same-sex relationships. Given these risks, the Board finds that content removal is the least intrusive means to provide protection to persons "outed" in this context. The damage from "outing" is immediate and impossible to undo; such measures can only be effective if implemented in a timely way.

The Board is concerned that Meta was not able to swiftly identify and remove clearly harmful content that involuntarily exposes the identities of persons alleged to be gay, which in turn perpetuates an atmosphere of fear for LGBTQIA+ people and fosters an environment where the targeting of marginalized groups is further accepted and normalized (see public comment by GLAAD, PC-29655). Even though the content violated four different Community Standards, was reported 112 times and reviewed by three different moderators, it was only after the Board selected the case for review that Meta removed the post and ensured similar content containing the video was taken down. The Board is particularly alarmed by the virality of the video, which was viewed over 3.6 million times, received about 9,000 reactions and 8,000 comments, and was shared about 5,000 times in a five-month period.

The Board understands that enforcement errors are to be expected in content moderation at-scale, however, Meta's explanations in this case reveal systemic failings. While Meta has taken additional steps to improve accuracy when enforcing the Coordinating Harm and Promoting Crime policy, and has sought to prevent similar errors through additional training, it did not provide details on measures implemented to ensure human reviewers assess content against all of Meta's policies. This is

particularly relevant in this case, in which the content was reviewed by three moderators who committed the same mistake, failing to assess the post against other relevant Community Standards. This indicates that Meta's enforcement systems were inadequate.

The Board finds Meta's enforcement error particularly alarming given the context in Nigeria, which criminalizes same-sex relationships. In order to improve implementation of its policies, and in addition to the measures the company has already deployed, Meta should conduct an assessment of the enforcement accuracy of the Coordinating Harm and Promoting Crime rule that prohibits content outing individuals by exposing their identity or locations. Based on the results of this assessment, Meta then needs to improve the accuracy of the policy's enforcement, including through updated training for content reviewers, given there should be no tolerance for this type of content.

The Board also examined Meta's enforcement practices in multilingual regions. In its exchanges with the Board, Meta initially misidentified the language of the video as Swahili, when it was actually Igbo. In response to a question from the Board, Meta noted that Igbo is not a language supported for content moderation at-scale for the Nigerian market, even if the company provides support for moderation of content in Igbo through agnostic review. According to Meta, the language is not supported because the demand for content moderation in Igbo is low. However, Meta informed the Board that when the content is in a language unsupported by the company's at-scale reviewers, such as Igbo, it is routed to language-agnostic reviewers (reviewers that work with content in multiple languages) who assess the content based on translations provided by Meta's machine translation systems. Meta also informed the Board that it has a few Igbo speakers who provide language expertise and content review for Igbo, although not at-scale, for the company.

The Board acknowledges that Meta has in place mechanisms to allow for moderation in unsupported languages, such as language-agnostic review and a few specialists with Igbo expertise. However, the Board is concerned that by not engaging human reviewers who speak Igbo in the at-scale moderation of content in this language, that is spoken by tens of millions of people in Nigeria and globally, the company's ability to effectively moderate content and mitigate potential risks is reduced. This could result in potential harm to user rights and safety, such as that experienced by the men shown in the video in this case. In light of its human rights commitments, Meta should reassess its criteria for selecting languages for support by the company's at-scale reviewers in order to be

in a better position to prevent and mitigate harms associated with the usage of its platforms.

Furthermore, Meta informed the Board that its automated systems detected the language as English, before routing the content for human review. According to Meta, this happened because the user's caption for the video was in English. While the caption was in English, the video is entirely in Igbo. Meta acknowledged that it wrongly identified the language of the content. The Board is concerned that bilingual content is being wrongly routed, potentially causing inaccurate enforcement.

In order to increase the efficiency and accuracy of content review in unsupported languages, Meta should make sure its language detection systems can precisely identify content in unsupported languages and provide accurate translations of that content to language-agnostic reviewers. Meta should also ensure that this type of content is always routed to language-agnostic reviewers, even if it contains a mix of supported and unsupported languages. The company should also provide reviewers with the option to re-route content containing an unsupported language to agnostic review.

The Board is very concerned that even after Meta removed the content in this case, the Board's research unearthed further instances of the same video dating back to December 2023, including in Facebook Groups, which had not been removed. This indicates that Meta must take much more seriously its due diligence responsibilities to respect human rights under the UNGPs. The Board welcomes the fact that this video was added to a MMS bank to prevent further uploads, after the Board flagged to Meta the remaining sequences of the video on Facebook. Given the severity of human rights harms that can result from Meta's platforms being used to distribute videos of this kind, Meta should make full use of automated enforcement to proactively remove similar violating content, in addition to using MMS banks to prevent new uploads.

6. The Oversight Board's Decision

The Oversight Board overturns Meta's original decision to leave up the content.

7. Recommendations

Content Policy

1. Meta should update the Coordinating Harm and Promoting Crime policy's at-scale prohibition on "outing" to include illustrative examples of "outing-risk groups," including LGBTQIA+ people in countries where same-sex relations are forbidden and/or such disclosures create significant safety risks.

The Board will consider this recommendation implemented when the public-facing language of the Coordinating Harm and Promoting Crime policy reflects the proposed change.

Enforcement

2. To improve implementation of its policy, Meta should conduct an assessment of the enforcement accuracy of the at-scale prohibition on exposing the identity or locations of anyone alleged to be a member of an outing-risk group, under the Coordinating Harm and Promoting Crime Community Standard.

The Board will consider this recommendation implemented when Meta publicly shares the results of the assessment and explains how the company intends to improve enforcement accuracy of this policy.

3. To increase the efficiency and accuracy of content review in unsupported languages, Meta should ensure its language detection systems precisely identify content in unsupported languages and provide accurate translations of that content to language agnostic reviewers.

The Board will consider this recommendation implemented when Meta shares data signaling increased accuracy in the routing and review of content in unsupported languages.

4. Meta should ensure that content containing an unsupported language, even if mixed with supported languages, is routed to agnostic review. This includes

providing reviewers with the option to re-route content containing an unsupported language to agnostic review.

The Board will consider this recommendation implemented when Meta provides the Board with data on the successful implementation of this routing option for reviewers.

***Procedural Note:**

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

- Under its Charter, the Oversight Board may review appeals from users whose content Meta removed, appeals from users who reported content that Meta left up, and decisions that Meta refers to it (Charter Article 2, Section 1). The Board has binding authority to uphold or overturn Meta's content decisions (Charter Article 3, Section 5; Charter Article 4). The Board may issue non-binding recommendations that Meta is required to respond to (Charter Article 3, Section 4; Article 4). Where Meta commits to act on recommendations, the Board monitors their implementation.

- For this case decision, independent research was commissioned on behalf of the Board. The Board was assisted by Duco Advisors, an advisory firm focusing on the intersection of geopolitics, trust and safety, and technology. Memetica, a digital investigations group providing risk advisory and threat intelligence services to mitigate online harms, also provided research. 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.