Account Ban for Targeting Public Figures

2026-006-IG-MR, 007-IG-MR, 008-IG-MR, 009-IG-MR, 010-IG-MR

Summary

The Oversight Board has due process concerns relating to how Meta disables user accounts and to the company's approach to account governance more broadly. In this case, the Board finds that the company was correct to permanently disable an account because of severe threats of violence it posted against a journalist. However, the Board has serious questions regarding the effectiveness of Meta's responses to such threats that the company must address to ensure respect for human rights.

The Board outlines key principles to guide Meta and other social media platforms in their approach to account governance, including respecting users' rights to transparency on account enforcement rules and providing a fair appeals process.

This Meta referral is the Board's first case assessing the permanent disabling of a user's account, piloting an extension of its mandate beyond individual content decisions.

Why This Matters

The Board is addressing an urgent concern for users – the disabling of accounts. It received more than 750 public comments meeting the terms for submission on this case. Those are added to the innumerable complaints it has received since the Board started taking cases in 2020 from users who have lost access to their accounts.

The Board highlights the importance of ensuring due process and protecting the free expression rights of people whose accounts are disabled and those targeted with abusive content. This case illustrates the disproportionate online abuse against women in the public sphere and its connections to physical violence. Such online attacks take place across all forms of social media.

About the Case

In 2025, Meta permanently disabled an Instagram account with more than 70,000 followers for violating the company's Community Standards.

In its referral, Meta included five of the account's posts. Two posts included visual threats of violence against a female journalist, showing a bullseye over her face. A third post used the abusive term "whore" against the same journalist, accusing her without evidence of engaging in sexual activity with a public figure. The journalist complained to Meta employees she knew, leading to escalated review. Meta initially removed the "whore" accusation post. On a second review, it identified the posts threatening violence and permanently disabled the account under the Account Integrity and Violence and Incitement policies due to the safety risks to the journalist.

Meta referred two other posts it said were illustrative of the account's broader behavior. One featured the anti-gay slur "fag" directed against prominent politicians. Another showed a photograph of a couple on public transportation who appeared to be engaged in oral sex, alleging they were from a religious minority and calling for their deportation. Meta removed each post shortly after their posting. The company cited these violations among numerous other strikes the account received in the 12 months prior to the permanent disablement as additionally justifying the account's removal under its policies.

Key Findings

The Board concludes that Meta was correct to permanently disable the account because of the severe threats of violence in two posts and to remove the three other posts that violated its policies. These actions were consistent with Meta's human rights responsibilities. The company must communicate a robust and deterrent position against threats of violence on its platforms.

Yet, the case raises concerns over due process and proportionality regarding account governance and the clarity of Meta's rules on permanent account disablement. The case raises serious questions about the efficacy of Meta's response to credible threats of severe violence

against people, including journalists and human rights defenders, that the company must address. Improving safety can improve people's right to freedom of expression.

This account's substantial following was seemingly gained through spreading conspiratorial posts and harassment of prominent people. For some Board Members, there are questions about how Meta's platform design rewards provocateurs' behavior and increases the risk of broadly spreading violent threats. Meta should review how its design choices promote the very behavior its policies prohibit.

Upholding this account disablement is not a blanket endorsement of the company's approach to banning and restricting accounts.

The case highlights the following human rights concerns at the systematic level:

- Due process for the targets of violent threats - The Board is seriously concerned that Meta did not review either of the two clear and credible threats swiftly when posted, delaying removal and exposing the targeted journalist to intolerable risk for a prolonged period.
- Due process for users whose accounts are disabled - Delays in providing justice undermined due process for the account holder, which could have provided more opportunities to correct his behavior.
- Unclear account suspensions and disablement policies – Accessible, consistent and comprehensive information about Meta's approaches must be ensured for users. Meta's expansion of AI assistants' role may help. Coherence between the policies and assistants' responses is important. The restrictions Meta applies to Instagram accounts prior to disabling are not public.
- Lack of a clear framework guiding decisions to permanently disable an account for "egregious" safety concerns – A more detailed but still flexible framework than currently available is needed for reviewers. This could improve consistency and avoid missing credible threats.
- No intermediate options between permanent disablement and leaving the strikes system to run its course for egregious policy violations – Where mitigating factors are present, a significant but time-bound suspension may be more proportionate.

The Board outlined the following principles to guide Meta and other social media platforms in their approach to account disablement:

i. Respect users' right to access clear and comprehensive guidance on the rules governing decisions to permanently disable social media accounts.
ii. Respect users' right to detailed reasons for account disablement, including providing: easily accessible, reliable information and appeal options; information on the role of any government request for either the review or disabling of an account, and of automation in the reviews.
iii. Social media companies should coordinate to create a program sharing information about accounts that credibly threaten serious violence.
iv. Allow effective and fair appeals, including the opportunity for users to provide written reasons for their appeal and prioritizing human review in edge cases.
v. Provide detailed transparency reports with meaningful information about account disablement enforcement trends.

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

Full Case Decision

1. Case Description and Background

In 2025, Meta permanently disabled an Instagram account with more than 70,000 followers for violating the company's Community Standards. Meta referred this decision to the Board, pointing to the challenges of respecting political speech while following its account disablement rules when users engage in patterns of abuse, including against public figures and threats against female journalists.

In its referral, Meta included five posts from the Instagram account made in the year before the company permanently disabled it. Of the referred content, two posts included visual threats of violence against a female journalist, showing a bullseye over her face. One of those two posts compared political elites in the country to autocrats, while the other contained no other intelligible information beyond the visual threat. A third post used the abusive term "whore" against the same journalist, accusing her without evidence of engaging in sexual activity with a public figure. These posts were reviewed together only after the journalist leveraged her connections to complain directly to Meta staff. Those employees escalated her complaints to internal teams twice. On the first escalation, the post using the term "whore" was removed and the account received a strike. On the second, Meta identified the posts threatening violence against the journalist and permanently disabled the account under its Account Integrity and Violence and Incitement policies due to the safety risks posed to the journalist, while noting the account was also near the disablement threshold under its strikes policy. Meta also disabled other accounts linked to the same user, two of which had also posted content targeting the journalist.

Meta referred two other posts, which it said were illustrative of the account's broader behavior. One post featured the anti-gay slur "fag" directed against prominent politicians. Another showed a photograph of an identifiable couple on public transportation who appeared to be engaged in oral sex, alleging they were members of

a religious minority and calling for their deportation. Neither of these last two posts involved the journalist. The company removed each post from the platform shortly after they were posted and applied a strike to the account upon each removal. Meta cited these violations among numerous other strikes the account received in the 12 months prior to the permanent disablement as additionally justifying the account's removal under its policies.

In taking this case, the Board aims to highlight impacts on the due process and free expression rights of both people whose accounts are disabled and those targeted with abusive content. The Board also notes the impact of online attacks on those targeted by them and that those attacks take place across all forms of social media. This case is emblematic of the disproportionate level of online abuse against women in the public sphere, including women journalists, and the concerning links between that abuse and physical violence. As part of the Board's own due diligence, some case details have been abstracted to avoid worsening the impacts of threats and harassment against the journalist.

2. User Submissions

Meta referred this case to the Board, it was not a user appeal and the Board did not receive a statement from the account holder.

In a statement to the Board, the journalist targeted by the account alleged that the account holder has a history of violence and that he had been harassing her through multiple social media accounts and offline stalking. She said his posts included death threats, defamation and incitement of violence to his followers, leading to a barrage of threats and abuse across multiple platforms. She said that this activity had severe emotional and psychological impacts, forcing her to take several security measures that have restricted her personal and professional life.

The journalist complained to Meta employees in her network who had the account reviewed. The Board has found that the account holder continued to post on other non-Meta platforms, though he has a much smaller audience on them.

3. Meta's Content Policies and Submissions

Individual Content Removals

Meta removed the two posts showing a bullseye over the journalist's face for violating its Violence and Incitement Community Standard. That policy prohibits "threats of violence that could lead to death (or other forms of high-severity violence)." Both posts met Meta's definition of threats of violence, which includes "visuals representing an intention, aspiration or call for violence against a target." The policy rationale also states that Meta will "remove content, disable accounts and work with law enforcement when we believe that there is a genuine risk of physical harm or direct threats to public safety."

Meta found that the post making sexual allegations against the journalist violated its Bullying and Harassment Community Standard for directing the term "whore" at the journalist. This policy prohibits content that is meant to degrade or shame, including attacks through derogatory terms related to sexual activity, listing this word as an example of violating content.

Meta removed the post using an anti-gay slur for violating its Hateful Conduct Community Standard. The policy prohibits content that targets people with slurs, defined as "words that inherently create an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic, often because these words are tied to historical discrimination, oppression and violence." Meta confirmed the term used in the post is on its slur list.

Meta removed the post depicting a sex act for violating the Adult Nudity and Sexual Activity Community Standard. This policy prohibits imagery of explicit and implicit sex

acts as well as impending sexual activity. The policy explains that Meta defaults to "removing sexual imagery to prevent the sharing of non-consensual or underage content."

Account Disablement

Meta's Account Integrity policy sets out that "to maintain a safe environment and empower free expression," the company will "restrict or remove accounts that are harmful to the community." Meta explained that its "approach to account enforcement aims to balance user education and rehabilitation, transparency, proportionality and timeliness." This approach is also outlined in the Transparency Center, including in pages on disabling accounts, restricting accounts and counting strikes.

Under the Account Integrity policy, Meta may disable accounts for a single violation, so long as it is sufficiently egregious, including for "posing severe safety risks." The Transparency Center sets out that "in some cases, a violation may be severe enough, such as posting child sexual exploitation content, that we'll disable your account … after one occurrence." The Violence and Incitement Community Standard also explains that content posing a genuine risk of physical harm can also lead to Meta permanently disabling the posting accounts. In its submissions, Meta explained that, in addition, illicit activity concerning high-risk drugs and human exploitation can also result in disablement for a single violation. According to Meta's submissions, these assessments, separate from the strikes system described below, are made on a case-by-case basis. To determine whether to disable an account (as opposed to simply removing the content), Meta considers factors such as "user behavior and recent activity," allowing it to "remain flexible and respond swiftly to high-risk accounts." When users report accounts, Meta reviews them holistically by assessing their profile picture, name and strike history as well as a limited number of recent posts by the account.

For persistent or repeated violations that are not egregious, Meta removes the content and notifies the user about the nature of the violation and its policies to help them understand the situation and adjust their future behavior. It may also apply a strike to

the account. Whether a strike is applied depends on factors such as the severity of the violation and the context in which it was shared. Strikes are not applied for every violation. For example, the company does not apply strikes to violating content "posted over 90 days ago for most violations or over four years ago for more severe violations." Strikes expire after one year. A first strike usually results in a warning with no further penalties but, as additional strikes are accrued, an account may receive increasing levels of restriction. The Account Integrity policy links to an explanation in the Transparency Center of Meta's approach to account restrictions, which specifies the thresholds at which accrual of strikes for repeat violations correspond to escalating penalties (which primarily only apply on Facebook), in addition to the removal of content. Under the strike system, accounts that persistently violate Meta's policies will be warned before losing access to certain features, such as going live on Instagram or being removed from recommendations. Facebook accounts can also be suspended for increasing periods of time after additional violations, but this penalty is not applied to Instagram accounts. Accounts that continue to accumulate strikes through violations will be disabled. At scale, Meta disables accounts that reach a specified threshold for standard strikes or a separate threshold for severe strikes.

In this case, when the account was escalated for review by internal experts, it was approaching the threshold for at-scale removal but had not crossed it. However, because the two posts threatened high-severity violence against an identified journalist which could lead to death, Meta's assessment of the risk the account posed led to its conclusion that the permanent disablement of the account was the only option consistent with the company's policies on egregious violations. Meta noted the account "demonstrated a persistent pattern of repeated violations" over the previous year, including violations of the Bullying and Harassment, Violence and Incitement, Hateful Conduct, Child Sexual Exploitation, Abuse and Nudity, Adult Nudity and Sexual Activity, and Adult Sexual Solicitation and Sexually Explicit Language policies. This pattern included, as examples, the additional three posts referred to the Board. Each of these violations would have been accompanied by messaging linking to the relevant policy and warning the user against further breaches of Meta's rules. Meta informed the Board that the user's account was subjected to account restrictions on 10 separate occasions,

seven of which were for 30 days each. These restrictions prevented the account from going live.

Given the account's escalating behavior, with two posts posing an egregious safety risk, ongoing violations despite repeated warnings and proximity to the automatic removal threshold, Meta determined that disabling the account was warranted under its Account Integrity policy and rationale, and consistent with its human rights responsibilities.

The Board asked Meta questions about the individual account, Meta's systems and procedures for identifying and assessing accounts for removal, the automated strike system and what action it takes on accounts that pose a safety risk. Meta responded to all of the Board's questions.

4. Public Comments

The Board received a substantial number of public comments on this case, with more than 750 meeting the terms for submission, and a selection of the most relevant have been published with this decision. These comments came from around the world and largely consisted of users explaining that they did not understand the reason their accounts were being permanently disabled, or believed the reason was wrong. Many commenters complained of having their accounts permanently disabled for allegedly posting child sexual exploitation content, even though they had never posted anything even remotely related to such content. Others pointed to seeming inconsistencies in Meta's enforcement as well as what they believe to be biased decision making. Some users claimed they were banned for interacting with content Meta had allowed on its platforms, or said they had seen other people post far more harmful content without any apparent penalty. Other commenters complained that their accounts were disabled after they were compromised and used by strangers.

Many commenters wrote about systems failing to work, saying they were unable to appeal Meta's decision to disable their account, that they never received any

explanation for why their account was disabled or that they were unable to download their content. Many of these users also noted that the decisions appeared to have been made automatically, with no human oversight, even on appeals against the disabling of longstanding and widely followed accounts. Others wrote that even when they created a new account specifically so they could pay for access to the Meta Verified program, which includes "24/7 access to email or chat agent support," they could not get meaningful assistance. Some said their new accounts were disabled for violating the prohibition on creating an account after being banned.

Commenters also wrote about the profound impact these decisions have had on them. Many lost access to networks of friends, contacts and followers in both personal and business contexts. Users forcefully expressed how losing their accounts harmed their social lives, mental health or financial wellbeing.

Since the Board's inception in 2020, Board Members have received innumerable complaints from Meta users through social media and other avenues who have lost access to their accounts. Under the current Charter and Bylaws, the Board does not have the authority to take these complaints as case appeals. However, Board Members are concerned by the volume and urgency of these requests, which have been broadly consistent with the public comments received on this case. Many users feel that Meta has treated them unfairly and that they lack a satisfactory way to appeal its decisions.

In January 2026, as part of ongoing stakeholder engagement, the Board consulted with trust and safety experts on the difficulties involved in setting account governance policies and tackling the problems of abusive accounts without overbroad restrictions on users' freedom of expression. The Board also consulted with journalists and representatives of advocacy organizations, academics, inter-governmental organizations and other experts on the impact and prevalence of online attacks, especially against women in the public eye, and the effectiveness of measures platforms use to protect their users from such attacks.

5. Oversight Board Analysis

This is the Board's first case assessing the permanent disabling of a user's account, piloting an extension of the Board's authority beyond individual content decisions. This referral from Meta requires the Board to analyze whether it was correct to disable the account and whether it was correct to remove the referred posts.

The Board concludes that it was correct to permanently disable the Instagram account because of the severe threats of violence against a journalist in two of the posts. The three other posts Meta referred also violated Meta's policies and were correctly removed. These actions were consistent with Meta's human rights responsibilities (in line with United Nations (UN) Guiding Principles on Business and Human Rights and Article 19, para 3, of the International Covenant on Civil and Political Rights (ICCPR)). At the same time, this case raises due process and proportionality concerns about Meta's overall approach to account governance, in addition to concerns around the clarity of Meta's rules governing permanent account disablement. There are also serious questions about the efficacy of Meta's response to credible threats of severe violence against people, including journalists and human rights defenders, that the company must address.

I. Permanent Disabling of the Account

Two of the posts targeting the journalist show her face with a bullseye superimposed over it. These are credible visual threats to shoot her and exhortation to others to do the same. These threats also motivated the account's substantial number of followers to engage in threats, creating a risk of inciting imminent violence against her. Such threats of high-severity violence are prohibited by Meta's Violence and Incitement policy. The removal of both posts was consistent with Meta's rules.

Permanently disabling the account was consistent with the Violence and Incitement policy, which specifies that Meta will "disable accounts and work with law enforcement when we believe that there is a genuine risk of physical harm or direct threats to public

safety." This is also reflected in the Account Integrity policy, which states that Meta will disable accounts that violate its policies in ways that involve "egregious harms, including those that we refer to law enforcement due to the risk of imminent harm to individual or public safety." The Board notes that, while Meta did not refer these threats to law enforcement, the company stated that this was only because by the time they were reviewed, the journalist had already reported them herself.

Meta's Corporate Human Rights Policy recognizes that journalists are human rights defenders and are a high-risk group that the company seeks to protect (see Section 4). This reflects Meta's responsibility to address credible threats of violence on its platforms, which can have particularly adverse human rights impacts when targeting journalists and other human rights defenders. Those impacts extend not only to the journalist herself, but also to all those engaged in reporting facts to the public who can be chilled by such violence and to the public whose right to information is infringed when journalists are intimidated out of doing their jobs.

This case is emblematic of a broader trend of women journalists being threatened online. The link between online attacks on women journalists and offline harm is well documented, with a recent report finding that over 40% of those surveyed experienced offline harm linked to attacks that started online, more than double the 20% surveyed in 2020. Online attacks have a chilling impact on the free expression of women targeted, diminishing access to information for all, as society hears less from women as a result. A 2026 report by the UN Special Rapporteur on human rights defenders noted that platforms have failed to provide adequate mechanisms to protect journalists and other human rights defenders from online abuse and attack. This speaks to the necessity and proportionality of the actions Meta took in this case and demonstrates that the trade-off between voice and safety is not zero-sum.

The removal of both posts was therefore required to meet Meta's responsibilities, and permanent disablement complied with the principles of necessity and proportionality.

The Board has found that analogous removals of visual threats of violence, including against women targeted due to their work, comply with Meta's responsibility to respect human rights, including freedom of expression (see Content Targeting Human Rights Defender in Peru). Meta's rules on Violence and Incitement, as applied to these posts, are sufficiently clear to satisfy the requirement of legality, and removal is necessary and proportionate to the legitimate aim of protecting others' right to life and security (ICCPR Article 19, para. 3; UN Human Rights Committee, General Comment No. 34, para. 25).

In the context of political expression, the Board has found that threats of violence are sometimes figurative or rhetorical and should be permitted (see Statements About the Japanese Prime Minister and Iran Protest Slogan). However, the Board finds no indication that the two bullseye posts were figurative or rhetorical threats.

Neither post engages in public interest commentary that could warrant leaving these egregious violating posts on the platform under Meta's newsworthiness allowance and the company's related freedom of expression responsibilities.

Firstly, while the comparison between the depicted public figures and totalitarian autocrats in one of the two posts is political opinion, it did not appear related to any contemporary news reporting or online discussions about these public figures. It revealed no facts that could initiate such a discussion. The visual threat against the journalist in the post is completely out of context and unrelated to the rest of the content. The other post contains no other information besides the threat and does not warrant a public interest exception.

Secondly, as the user is not a political leader or a public figure, there was no heightened interest in other users hearing from him that could warrant leaving these threats on the platform. Even in the Former President Trump's Suspension and Cambodian Prime Minister cases, the Board warned against using the newsworthiness allowance to keep prohibited threats of violence by politicians or other public figures on the platform. There is substantially more public interest in

hearing from political leaders than from the account holder in this case. This account's substantial following was seemingly gained through spreading conspiratorial posts and through harassment of prominent people. On balance, the Board finds that his relatively large following contributes less toward a finding that he speaks in the public interest and more toward a finding that his threats could produce imminent harm. For some Board Members, this raises questions about how Meta's platform design rewards provocateurs' behavior and increases the risk of violent threats reaching a broad audience, as also discussed in the Former President Trump's Suspension decision. Meta should review how its design choices promote the very behavior its policies prohibit.

The targeted journalist's submissions to the Board support the credibility and imminence of the threats. Not all of those details can be disclosed in this decision in the interests of the journalist's safety. These details would not necessarily have been available to Meta at the time the posts should have been removed. However, when she escalated the concerns to her contacts in the company, they presumably had as much opportunity to inquire into the facts as the Board did in this case. It was only on the second review following an escalation that the permanent disablement decision was made.

Similarly, the account's prior behavior in response to strikes for earlier violations is relevant, albeit not dispositive. The volume of strikes the user had received for previous violations indicates this user was not responsive to warnings or escalating sanctions, showing no commitment to change his behavior. Combined with the severity of harm in the two posts reviewed above, this factor further supports the proportionality of permanent disablement as applied to the facts of this case.

These factors point towards the conclusion that permanent disablement was the least intrusive means to achieve the aim of ensuring the safety of the journalist targeted by this account. Given Meta's clear responsibility to respect human life and given the vulnerabilities of human rights defenders and journalists to threats against their lives, it is important that the company communicates a robust and deterrent position that it does not tolerate this type of abuse on its platforms. This is commensurate with the

other situations in which its policies sanction permanent disablement based on single violations.

It is important that the decision to uphold Meta's account disablement in this case is not taken as a blanket endorsement of the company's approach to banning and restricting accounts. The case highlights for the Board the following broader human rights concerns at the systemic level around legality and due process, as well as the proportionality of available penalties for egregious violations in the governance of accounts.

a. *Improve Due Process for the Targets of Violent Threats*

The Board is seriously concerned that Meta did not review either of these clear and credible threats swiftly when they were posted, delaying removal and exposing the targeted journalist to intolerable risk for a prolonged period. Instead, it required the journalist to leverage her connections at Meta to compel the company to act. Any user in such a situation without similar connections would be faced with no recourse.

b. *Improve Due Process for Users Whose Accounts are Disabled*

Delays in providing justice to the targeted journalist also undermined due process for the account holder. Had Meta reacted to his violations sooner, the account holder may have had more opportunities to understand the nature of his wrongdoing and correct his behavior, including during additional periods of account restriction. Instead, several of these violations were actioned during an escalated review of his account. As discussed below, it is concerning that Instagram accounts are not subject to temporary suspensions following repeated policy violations and instead are only restricted from going live as an intermediate penalty before being permanently disabled.

c. *Clarify Policies on Account Suspensions and Disablement*

While Meta had a clear basis in its policies to take the actions it did in this case, clarity to users about account disablement and restriction policies could be significantly

improved. There are two tracks to permanent disablement of an account. One, as in this case, is a "one and done" approach for egregious violations. The other is the strikes system, with increasing penalties corresponding to the volume of violations an account amasses over time, with distinctions between regular strikes and severe strikes. The latter can be referenced as a factor for decisions under the former, but for the most part, it is useful to think of them operating in parallel.

While information about permanent disablement is included in the Violence and Incitement policy, a full picture of Meta's two-system approach requires consulting the Account Integrity policy, and the Restricting Accounts, Counting Strikes and Disabling Accounts pages in the Transparency Center. Even after such consultation, the distinction between the two systems across these resources can be difficult to follow. The examples of "egregious" violations that can result in permanent disablement are not comprehensive or easily distinguishable from "severe" strikes that can lead to swifter serious penalties under the strikes system. Worse still, some of the information in these sources is contradictory. For example, the Disabling Accounts page states that users may receive a 30-day restriction of creating content after five strikes. In contrast, the Restricting Accounts page says the same penalty will not be imposed until 10 strikes. While there are hyperlinks between some of these pages, they are not all easy to find.

There is a clear need for Meta to harmonize these various approaches and explanations into a single accessible resource that clearly distinguishes and ensures consistent information about Meta's two main approaches. For egregious violations that result in permanent disablement, much more comprehensive information is needed, including hyperlinks to all policies that, if violated, can result in permanent disablement. The Board notes that Meta's expansion of the role of AI assistants on its platforms may help better explain rules to users at key moments. Ensuring coherence between the underlying policies and responses these assistants provide to user queries is important.

Additionally, Meta does not publicly set out all the restrictions it can apply to Instagram accounts for violations before and up to disabling them. The Disabling Accounts page

discusses 30-day restrictions but does not specify which of Meta's platforms are subject to such penalties. The restrictions listed on the Restricting Accounts page only apply to Facebook, although strikes are also counted on Instagram. The Instagram Help Centre does not provide any clarifying information. Meta informed the Board that it does not temporarily suspend Instagram accounts in response to policy violations. Instead, it only restricts Instagram accounts from going live for set spans of time, which lengthen based on the severity of the violation and the enforcement history of the account. There is a significant gulf in severity between suspending a user from live broadcast features and permanently disabling their account. Additional graduated penalties between these escalating sanctions would allow a more proportionate approach. The Board also notes the "go live" feature is only available to users who have more than 1,000 followers, and many do not use this feature. Its effectiveness as a penalty, especially in response to violations not committed as part of a live broadcast, is doubtful. For violations in permanent posts, a penalty that directly corresponds to violating behavior by suspending a user's ability to post (e.g., by putting their account in read-only mode for a set period) would have a greater chance of influencing behavior.

d. Create a Framework to Guide Permanent Disablement Decisions

While Meta reached the right conclusion in this case, the Board is concerned that the company lacks a clear framework guiding decisions to permanently disable an account for "egregious" safety concerns. Although case-by-case assessment is appropriate, factors such as "user behavior and recent activity" are extremely broad and open-ended and the escalation of these assessments must follow clearly structured paths. They could create unjustifiably different outcomes depending on the reviewers' perspectives.

There will always be a tension between seeking consistency through specific guidance that anticipates the various ways serious threats can manifest on Meta's platforms, on the one hand, and allowing reviewers to properly consider context for aggravating and mitigating factors, on the other. This balance between prescription and discretion can be difficult: too much prescription can lead to a robotic analysis where key indicators

of harm are missed or misunderstood, while too much discretion can lead to different reviewers reaching unjustifiably divergent outcomes based on the same information. Adopting a more detailed but still flexible framework for reviewers with criteria that still preserve the necessary discretion to engage in contextual analysis could help ensure greater consistency and avoid situations where credible threats, including against journalists or human rights defenders, are missed. The absence of more specific guidance may explain why Meta failed to act at various points on the threats against the journalist in this case. Publicly sharing information about such a framework would also increase transparency for Meta's users.

e. *Ensure Proportionate Penalties*

The Board is concerned that for egregious violations that pose credible safety threats to individuals or broader society, Meta's policy provides no intermediate options between permanent disablement and leaving the strikes system to run its course. Under the strikes system, repeated 30-day suspensions are the only penalty that limits full account access prior to a permanent disablement. Where mitigating factors are present, a significant but time-bound suspension, such as those the company adopted for restrictions on public leaders' accounts during civic unrest, may be more proportionate to achieve the same end of ensuring individual or community safety, while retaining the possibility of rehabilitation for the user.

II. Other Content Decisions

The analysis of the remaining three posts is separated because it does not fundamentally alter the Board's key conclusions on the permanent disabling of the account. Nevertheless, one of the three posts displays clear gender-based animus towards the same journalist targeted by the threats. All three posts reinforce the conclusion that there was little public interest in the posts that would justify a departure from Meta's regular policies, including permanent disablement.

a. *Bullying and Harassment Case*

This post attacked the same female journalist targeted in the above posts, calling her a "whore." The Board finds this slur is prohibited by the Bullying and Harassment policy's prohibition on derogatory terms related to sexual activity, justifying removal. This determination is supported by Meta's human rights responsibilities. The prohibition is clear, meeting the requirement of legality as applied to this post. Removal was necessary and proportionate to the legitimate aim of respecting the rights of the journalist targeted, including equality and non-discrimination. The Board has previously found that sexual allegations against women in the public eye can have serious implications, including for their safety (see Explicit AI Images of Female Public Figures). Harassment also has serious impacts on the freedom of expression of those it targets. Research has shown the chilling impact that harassment and other forms of online attacks can have on women journalists, with 30% of those surveyed engaging in self-censorship and 20% withdrawing from all online interaction as a result. Here, the allegation against the journalist appears to be the sole point of the post; it was not incidental to some broader commentary or political message.

b. *Hateful Conduct Case*

This post negatively targeted politicians with an anti-gay slur, as prohibited under Meta's Hateful Conduct policy. The Board finds this term is on Meta's slur list and meets the policy's definition of being a word that inherently creates "an atmosphere of exclusion and intimidation against people on the basis of a protected characteristic." This prohibition is clear as applied to the facts of this case and pursues the legitimate aim of protecting the rights of others, including equality and non-discrimination. The removal of this post was necessary to protect the rights of LGBTQIA+ people on Meta's platforms, given the dehumanizing effects of slurs that include incitement of imminent discrimination (see Emojis Targeting Black People). The term has a long history connected to discrimination and violence against LGBTQIA+ people, which is perpetuated by its continued use.

c. *Adult Nudity and Sexual Activity Case*

This post contains images of two adults on public transport with calls for their deportation based on their religion. The post states that the incident was in the account holder's country, though the Board's research has shown that to be false. From the positioning of the people pictured, they appear to be engaged in an oral sex act, postings of which are prohibited by Meta's Adult Nudity and Sexual Activity policy, requiring removal. This is consistent with Meta's human rights responsibilities. This prohibition is sufficiently clear as applied to this post to meet the requirements of legality, and the Board has previously found that the policy broadly pursues the legitimate aim of protecting the rights of others (see Images of Partially Nude Indigenous Women). Removal was necessary to protect the rights to privacy and dignity of the people pictured. There is no public interest in the post's false claims that would justify leaving it on Instagram, notwithstanding the violation. It misrepresented the location of the incident to call for the deportation of the couple involved, underpinned by unsupported assumptions about their ethnicity, religion and immigration status. No less restrictive option than removal was available to achieve this objective.

6. Account Governance Good Practices

Given that this case is a pilot, it does not include formal recommendations to Meta but offers guiding principles to the industry more broadly. The Board expects to issue formal recommendations in future decisions on account cases.

The number of public comments received on this case demonstrates the significant demands for greater fairness from Meta's users whose accounts have been permanently disabled. There is also a need to ensure fair treatment for users who are targeted by threats and harassment from abusive accounts. Here, some basic principles on permanent account disablement are set out, principally in response to users' experiences on Meta's platforms. These may have broader relevance across industry for the governance of accounts on other platforms.

Respecting Users' Rights to Clear and Comprehensive Rules on Account Disablement

Companies should respect users' right to access clear and comprehensive guidance on the rules governing decisions to permanently disable social media accounts.

The rules should:

a. Clearly explain the approach to sanctioning accounts for content they post or other behaviors they engage in.
b. Clearly explain which rules, if violated, will be met with notice and graduated sanctions (e.g., a strikes system), and which can result in immediate account disablement (e.g., a "one and done" rule).
c. For graduated penalties, publicize the thresholds for escalating penalties, including the threshold for permanent account disablement. These should clearly distinguish any tiering of severity between different types of violation (e.g., regular strikes vs. severe strikes) and provide multiple escalating levels of penalties to make it more likely that a proportionate option is available.
d. Explain policy frameworks and factors internal enforcement teams consider when permanently disabling accounts.
e. If sanctions vary between platforms a company owns, or for violations on different surfaces of a platform (e.g., publicly vs in private), explain these distinctions.

Respecting Users' Rights to Detailed Reasons for Account Disablement

When users' accounts are sanctioned for content or behavior, they are entitled to be notified of the rule they violated, and to maintain access to the history of violations and current account status.

This requires:

a. Providing users with dashboards through which they can easily access reliable information on current account status, past violations, information on available appeal options, and the status of any pending appeals. This information should be downloadable for record-keeping purposes.

b. Clear, prominent and timely notifications to users, in a language they understand, of any warning or penalty at the time it was imposed, specifying the rule violated, the sanction imposed and options to appeal.
c. Information on the role of any government request for either the review or disabling of an account.
d. Information on the role of automation in the review of content or behavior and the imposition of warnings or penalties.
e. Ensuring that users who lose access to their account due to a permanent disablement decision can still access the reasons for a decision and appeal mechanisms, for example, through account status dashboards.

Coordination for Addressing Violent Threats Across Platforms

Often, credible threats of violence against individuals are coordinated between actors and across social media platforms. This creates particular challenges for the targets of threats to manage their own safety, in particular where in-app reporting tools fail. Coordination among social media platforms has been important for ensuring effective responses to harmful behaviors in the context of countering terrorist threats (for example, the Global Internet Forum to Counter Terrorism) and in tackling threats to child safety (for example, the Lantern program). Social media companies should create a program to share information about accounts that credibly threaten serious violence.

This program should:

a. Enable the sharing of credible threat information between platforms, including details about the target and source of the threat, in addition to any relevant off-platform context.
b. Establish dedicated channels for high-risk targets of violence and their representatives, including journalists and human rights defenders.
c. Share good practices on addressing threats, including on establishing the credibility of threats and distinguishing rhetorical and non-credible threats, and on referring threats, when appropriate, to law enforcement and on the preservation of evidence for accountability purposes.

Effective and Fair Appeals

Many public comments cited experiences of account disablements being accompanied by reasons they did not understand, as well as frustration when appeals were denied within moments of submission. While anecdotal, many of these experiences appear to relate to "one and done" type violations of the kind reviewed in this case.

The following principles should guide evaluations of whether these systems can be made more effective and fairer:

a. Appeal mechanisms should provide users with an opportunity to provide written reasons for their appeal, supported by continued access to the company's policy justification for the disablement and any available exceptions within the policy that may apply. Providing users with tags for how their appeals might fit within applicable exceptions (e.g., "it's news reporting" or "it's satire") should be explored.
b. Appeals of account disablement are not effective if they run the same decision through the same automated process, expecting different results with the same inputs. While advances in artificial intelligence should be leveraged to ensure more accurate and comprehensive reviews, taking into account users' written submissions, prioritizing of human review in edge cases will be important. This is especially important for "one and done" violations, where any error in automated decision-making has very serious consequences beyond the removal of a post or imposition of a strike.

Transparency Reporting on Account Disablement

Platforms should provide detailed transparency reports that allow the public to access meaningful information about enforcement trends. This reporting should include:

a. The total number of accounts disabled for violating platform policies, disaggregated between accounts disabled under "one and done" rules,

specifying which rule resulted in those disablements, and those disabled under any graduated enforcement system (e.g., strikes).

b. The number of accounts that were disabled following a request to review them from a government or law enforcement agency.
c. The number of disabled accounts that were also referred to law enforcement and the policies they violated.
d. The ability to break down this data by region and language.