Board to Review for First Time Meta Approach to Disabling Accounts

Digital Rights Foundation
3 February 2026

In 2025, Meta permanently disabled an Instagram account for repeatedly violating the company's community standards. The account posted five posts the year before it was permanently disabled, with multiple posts related to graphic threats and harassment for women journalists. Amongst these other posts included anti-gay slurs against prominent politicians and content that depicted them in a sexual act, which allegedly also pointed towards misconduct against minorities. According to Meta, the posts violated Violence and Incitement, Bullying and Harassment, Hateful Conduct, and Adult Nudity and Sexual Activity community guidelines. Meta removed each post from the platform and applied a strike to the account after each violation.

The account was flagged by Meta staff and escalated to internal specialists for further review. Their assessment found a sustained pattern of policy violations over the past year and identified the account as a safety concern, as several posts incited violence with the potential to cause loss of life. Although the account had not yet reached the threshold for automatic enforcement action, the cumulative risk posed by repeated violations led Meta to permanently disable the account.

Meta's Account Integrity policy allows the company to disable accounts that repeatedly breach its rules, as well as those that show a clear intent to do so. In its explanation, Meta clarified that enforcement actions are not limited to the strike system and that accounts may be disabled on a case-by-case basis, based on an overall assessment of a user's behavior and activity.

While the account has been posting objectionable content, ensuring due process and fairness for individuals whose accounts are penalized or permanently disabled requires greater transparency and meaningful opportunities to contest enforcement decisions. Currently, users are often not informed of the specific post that resulted in a strike or the particular community guidelines they are accused of violating, leaving them unable to understand the basis of the action or effectively challenge it. This lack of clarity undermines the right to appeal, as users cannot meaningfully argue their case without knowing what content triggered enforcement. Furthermore, when appeals are submitted, responses are frequently inadequate or opaque, offering little insight into whether the appeal was properly reviewed or considered. Together, these gaps raise serious concerns about the effectiveness and fairness of platform appeal processes.

Another critical concern relates to the effectiveness of social media platforms' measures to protect public figures and journalists from accounts that engage in repeated abuse and threats of

violence, particularly women in the public eye, who face disproportionate and gendered attacks. In practice, harmful content that re-emerges months or even years after it was first posted is often deprioritized or deemed out of scope during escalation processes, including by trusted partners. This creates a significant gap in protection, as older content can still incite harassment, revive coordinated abuse campaigns, or pose real safety risks, especially when it targets women journalists, politicians, or activists. It is reasonable to assume that individuals self-reporting face similar, if not greater, barriers.

Additionally, platform moderation frequently focuses on individual posts while overlooking comment sections, which often become concentrated spaces for hate speech, threats and gender-based violence. Even when the original post does not violate platform policies, the ensuing comments can rapidly escalate into coordinated abuse that remains insufficiently moderated. Although platforms rely on keyword and hashtag detection, users increasingly employ moderation-evasion tactics, including coded language, evolving slang, emojis and visual cues that automated systems fail to recognize.

Moreover, abusive campaigns are rarely confined to a single platform. Cross-platform posting and tactics such as directing users through “link in bio” features allow harassment and threats to migrate seamlessly across services, undermining platform-specific enforcement efforts. These dynamics highlight the need for more holistic, context-aware and survivor-centered protection mechanisms that account for persistence, amplification and coordination in online abuse, particularly where the safety of women public figures and journalists is at stake.

A significant challenge in assessing threats against public figures and journalists lies in platforms’ limited ability to identify and meaningfully account for off-platform context, where online threats often intersect with real-world risks. Trusted Partners play a crucial role in flagging these dangers early, yet they are frequently asked to provide extensive evidence of imminent physical harm before content is removed. Such a standard is both unrealistic and counterproductive, as proving imminent harm is often only possible after violence has already occurred, the very outcome Trusted Partners are working to prevent.

Additionally, platform assessments frequently overlook the need to adequately incorporate cultural context and local sensitivities, particularly in regions outside the West. When Trusted Partners flag accounts or content as unsafe, their expertise is grounded in an understanding of local political climates, linguistic cues, religious and gendered dynamics, and historical patterns of violence. Discounting this contextual knowledge weakens risk assessments and leaves public figures and journalists, especially women and marginalized individuals, exposed to harm. Platforms must place greater trust in Trusted Partners’ localized expertise and treat their

assessments as credible indicators of risk, rather than requiring narrowly defined or post-facto proof of physical violence.

It is worth noting that research examining the effectiveness of punitive measures in shaping online behavior suggests that punishment alone is insufficient and often fails to meet the needs of those harmed by online abuse. A 2021 text-message-based survey[1] of 832 adolescents and young adults in the United States, aged 14 to 24, found that punitive platform responses rarely create space for justice, accountability, or meaningful reparation for targets of harassment. Participants were twice as likely to distrust social media companies' ability to deliver a fair resolution (41%) than to trust them (20%), highlighting a significant confidence gap in platform enforcement systems. Notably, 62% of respondents indicated a preference for receiving an apology from the person responsible for the harassment, highlighting the potential value of restorative or complementary approaches, alongside or instead of purely punitive interventions.

Good industry practice requires robust transparency around account enforcement decisions and the handling of related appeals, particularly where permanent penalties are imposed. Current mechanisms fall short of this standard, as evidenced by persistent shortcomings in the new escalation portal for Trusted Partners, including delayed responses and a lack of clarity or consistency in how cases are assessed and prioritized. At the same time, enforcement systems struggle to address recidivism effectively, allowing bad actors to repeatedly return using new accounts or VPNs to evade detection. In contrast, journalists, human rights defenders, and independent media outlets, especially those operating in remote or underrepresented regions, are frequently and incorrectly penalized, often for alleged violations involving CSAM, violence, or encouraging self-harm. These accounts are commonly permanently disabled without meaningful explanation or access to appeal, resulting in irreversible harm to public interest reporting and civic discourse.

Oversight board case:
https://www.oversightboard.com/news/board-to-review-for-first-time-meta-approach-to-disabling-accounts/

[1] https://dl.acm.org/doi/10.1145/3449076