CEJIL
CENTRO POR LA JUSTICIA Y EL DERECHO INTERNACIONAL

February 3, 2026

Members of the Oversight Board
META

Ref: Submitting public comments
Account Ban for Targeting Public Figures

In order to respond to the request for comments in the case "Account Ban for Targeting Public Figures", the Center for Justice and International Law (CEJIL),[1] submits the following observations, drawing primarily on standards developed by the Inter-American System of Human Rights. These comments are further informed by the Esperanza Protocol,[2] a tool developed by CEJIL that systematizes international human rights standards applicable to threats and intimidation against human rights defenders. Based on this framework and CEJIL's longstanding experience in the protection of human rights defenders in the region, the present submission addresses: A) the effectiveness of measures used by social media platforms to protect public figures and journalists from accounts engaged in repeated abuse and threats of violence; and B) the challenges in identifying and considering off-platform context when assessing threats against public figures and journalists.

I. The effectiveness of measures used by social media platforms to protect public figures and journalists from accounts engaged in repeated abuse and threats of violence

While the present call refers specifically to public figures and journalists, the comments submitted address more broadly the measures that social media platforms should adopt to protect human rights defenders in general, particularly the need for account level enforcement when repeated conduct undermines their safety and freedom of expression. Under the standards developed by the Inter-American System of Human Rights and the United Nations Declaration on Human Rights Defenders, human rights defenders are understood in a broad sense as any individual or group who, through peaceful means, promotes, protects, or strives for the realization of human rights and fundamental freedoms.[3] Journalists therefore constitute a recognized category of human rights defenders due to their essential role in informing the public and enabling democratic debate.

Importantly, the protection owed to defenders does not depend exclusively on their level of public notoriety, but on the nature of their work and the risks it entails. While some defenders become

[1] The Center for Justice and International Law is a civil society organization with over 30 years of experience, whose mission is to contribute to the full enjoyment of human rights in the Americas through the effective use of the Inter-American System's tools and other international human rights protection mechanisms. Website: https://cejil.org/en/
[2] Protocol for an Effective Response to Threats Against Human Rights Defenders (Esperanza Protocol), available at: https://esperanzaprotocol.net/en/
[3] CIDH. *Tercer informe. Situación de personas defensoras de derechos humanos en las Américas.* OEA/Ser.L/V/II, Doc. 119/25, April 15th, 2025, at 29. United Nations. *Declaration on human rights defenders,* Doc. ONU A/RES/53/144. March 8, 1999, Annex – Article 1.

public figures because of their visibility and face heightened risks as a result,[4] many others carry out crucial work without public recognition and nevertheless experience comparable levels of danger due to their involvement in specific organizations.[5] Limiting protective measures to visibility-based categories risks leaving other defenders exposed to the same patterns of intimidation and silencing.

Measures used by platforms to protect human rights defenders must take into account that threats, harassment, and intimidation occurring online constitute a serious interference not only with their individual rights, but also with the rights of society as a whole, which is prevented from learning the truth about the situation of respect for or violation of human rights.[6] The Inter-American Court of Human Rights has consistently held that such acts undermine freedom of expression and the public's right to seek, receive, and impart information.[7] Persistent threats generate a chilling effect that extends beyond the individual targeted, discouraging others from engaging in public debate, investigative journalism, or human rights advocacy.[8] Likewise, Esperanza Protocol emphasizes that the cumulative nature of threats can erode democratic space and silence critical voices, even when no physical violence ultimately occurs.[9] Where platforms allow such cumulative threats to persist, they contribute to the very chilling effects that international human rights law seeks to prevent, triggering a heightened responsibility to intervene effectively.

The Meta Oversight Board has translated these human rights principles into concrete enforcement guidance in previous cases. In its 2025 decision "Content Targeting Human Rights Defender in Peru",[10] concerning a veiled threat against a human rights defender, the Board overturned Meta's initial decision to leave the content online. It found that the post constituted a threat when assessed in its broader context, including patterns of hostility and the heightened risks faced by defenders. The Board underscored the importance of contextual analysis and recommended that Meta clarify its policies to explicitly address veiled threats and patterns of intimidation directed at human rights defenders.

From this perspective, although the Board analyzed only a specific post in this case, its contextual analysis invites an approach in which the effectiveness of platform measures is not limited to the removal of individual pieces of content. Both the Inter-American Court, the Inter-American

[4] *See, for example,* I/A Court H.R., *Case of García Andrade et al. v. Mexico*. Preliminary Objection, Merits, Reparations and Costs. Judgment of August 22, 2025. Series C No. 563, at para. 73.
[5] IACHR. Resolution 26/2024. MC 438-15, *Integrantes del Programa Venezolano de Educación-Acción en Derechos Humanos (PROVEA), Venezuela,* April 29, 2024, at para. 25. IACHR. Resolution 12/2021. MC 1051-20 34, *Miembros identificados del Periódico Digital El Faro, El Salvador*, February 4, 2021, at para. 43.
[6] I/A Court H.R., *Case of Bedoya Lima et al. v. Colombia*. Merits, Reparations and Costs. Judgment of August 26, 2021. Series C No. 431, at para. 107. *See also,* United Nations Human Rights Council, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Doc. A/HRC/20/17. June 4, 2012, paras. 3 and 4.
[7] *Ibid.,* at para. 110. *See also,* I/A Court H.R., *Case of Ivcher Bronstein v. Peru.* Merits, Reparations and Costs. Judgment of *February 6, 2001. Series C No. 74, at para. 148, and I/A Court H.R., Case* of Carvajal *Carvajal* et al. v. Colombia. Merits, Reparations and Costs. Judgment of March 13, 2018. Series C No. 352., at 172.
[8] I/A Court H.R., *Case of Palamara Iribarne v. Chile.* Merits, Reparations, and Costs, Judgment of November 22, 2005, Series C No. 135, para. 68, and I/A Court H.R., *Case of Vélez Restrepo and family v. Colombia.* Preliminary Objection, Merits, Reparations, and Costs. Judgment of September 3, 2012. Series C No. 248., para. 139.
[9] Esperanza Protocol, at 4.
[10] Oversight Board, "Content Targeting Human Rights Defender in Peru", May 27, 2025.

Commission on Human Rights, and the Esperanza Protocol emphasize that repeated acts of harassment, intimidation, or threats, even when individually ambiguous or of low intensity, may place human rights defenders in a situation of aggravated risk. [11] Importantly, this assessment does not depend on the repetition of identical messages, but on the foreseeable impact of a pattern of conduct over time. When accounts have repeatedly demonstrated an intent to intimidate or harass human rights defenders, allowing them to remain active perpetuates fear, normalizes intimidation, and fails to respond to the heightened duty of protection triggered by such risk. In such circumstances, account level measures, including disabling, are not exceptional responses but necessary and proportionate tools to prevent ongoing harm and foreseeable violations of rights.

A particularly illustrative example is the case of Francisco Vera, a young Colombian environmental activist who has been subjected to constant online harassment, originating primarily from within Colombia, including false accusations and offensive narratives portraying him as a member of illegal armed groups.[12] While many of these messages might appear minor or non-actionable when viewed in isolation, their sustained and massive dissemination created an environment of escalating hostility that ultimately forced him to leave Colombia for his safety. This case demonstrates that individually low-grade messages, which may not meet a high threshold of severity on their own, can nonetheless produce serious harm when accumulated over time, resulting in severe real-world consequences and generating an immediate, structural risk to the life and integrity of a young defender. In this sense, disabling accounts that repeatedly violate community standards would help reduce the volume and visibility of such recurring messages, particularly given that much of this harmful content is commonly amplified by a limited number of especially influential accounts.

The importance of addressing repeated messages that have an impact because of their number and recurrence is further reinforced by the Oversight Board's reasoning in the case concerning the "Brazilian General's Speech".[13] In that decision, the Board expressed concern about Meta's failure to act on identical content circulating in parallel contexts prior to the January 8 riots in Brazil, emphasizing that repeated dissemination of content with the same meaning and inciting potential cannot be assessed in isolation. While that case focused on identical content, its underlying rationale is equally applicable to situations involving threats and intimidation against human rights defenders. From a human rights perspective, harm does not depend on strict identity of content. As reflected by the jurisprudence of the Inter-American System on Human Rights, repeated dissemination of highly similar messages, narratives, or insinuations may cumulatively amount to

[11] I/A Court H.R., *Case of Bedoya Lima et al. v. Colombia*. Merits, Reparations and Costs. Judgment of August 26, 2021. Series C No. 431, at para. 123 and 150-153, Esperanza Protocol, at 14; CIDH. *Tercer informe. Situación de personas defensoras de derechos humanos en las Américas.* OEA/Ser.L/V/II, Doc. 119/25, April 15th, 2025, at para. 238.

[12] Defensoría del Pueblo de Colombia, "Protección integral y respeto para la labor de Francisco Javier Vera Manzanares", September 2025, available at: https://x.com/franciscoactiv2/status/1973011311979335681

[13] Oversight Board, *Brazilian General's Speech,* June 22, 2023.

intimidation,[14] even when the wording varies. Such patterns can generate fear and chilling effects, particularly when directed at the same defender or group of defenders.

Accordingly, Meta's approach should extend beyond identical content to encompass patterns of substantially similar conduct. Where such content originates from an account that has repeatedly violated platform rules and demonstrated an intent to intimidate, account level measures, including disabling, are necessary to prevent ongoing harm. This approach is consistent with the Oversight Board's emphasis on preventing foreseeable harm in escalating contexts and with international human rights standards requiring effective protection of human rights defenders.

Taken together, these standards lead to a clear conclusion. Disabling accounts that engage in repeated abuse and threats against journalists and other human rights defenders is a necessary and proportionate measure under international human rights law. As recognized by the Inter-American System of Human Rights, the cumulative impact of threats can result in self-censorship, withdrawal from public life, and the silencing of critical voices. Measures that fail to disrupt these patterns are therefore ineffective and incompatible with Meta's responsibility to prevent foreseeable harm to human rights defenders.

II. Challenges in identifying and considering off-platform context when assessing threats against public figures and journalists

A central challenge in assessing threats against journalists and other human rights defenders lies in the tendency to evaluate individual posts in isolation, without adequate consideration of their broader context. This fragmented approach is inconsistent with the standards developed by the Inter-American Court, which has repeatedly held that threats must be assessed considering surrounding circumstances. For example, in the case "Jineth Bedoya Lima et al. v. Colombia", regarding the kidnapping, torture, and sexual violence inflicted on a journalist as a reprisal for her work, the Court emphasized that the seriousness of threats must be evaluated considering patterns of aggression, prior acts of violence, and the specific risks faced by the journalist.[15]

The Esperanza Protocol reflects this same analytical framework by underscoring that conduct which may appear minor, ambiguous, or non-threatening, when viewed in context, can amount to serious intimidation. Repeated hostile messages, insinuations, or symbolic acts, particularly when directed at the same person, may be designed to exhaust, frighten, and silence human rights defenders, even in the absence of explicit threats or physical violence.[16] From a human rights perspective, it is precisely this cumulative effect that gives such conduct its coercive power. Against this, social media platforms should engage in context-specific moderation globally, considering different languages, usage, and contexts.[17]

---

[14] I/A Court H.R., *Jorge Luis Salas Arenas and his family members, Peru,* Provisional Measures, Sept. 4, 2023, at para. 62, and IACHR. Resolución 76/2021 MC No. 475-21, *Bertha María Deleón Gutiérrez, El Salvador*, Sept. 19, 2021, at para. 30-33.
[15] I/A Court H.R., *Case of Bedoya Lima et al. v. Colombia*. Merits, Reparations and Costs. Judgment of August 26, 2021. Series C No. 431, at para. 126.
[16] Esperanza Protocol, at 38.
[17] *Ibid.,* at 19.

Off-platform context is often critical to understanding the real gravity of online threats. Human rights defenders are frequently targeted online in direct connection with their investigations, advocacy, or litigation. Digital harassment may coincide with offline surveillance, stigmatization, smear campaigns, or prior acts of violence. In such contexts, even indirect or veiled threats acquire heightened significance. The Inter-American Court has recognized that where there is a foreseeable risk of harm, the failure to take threats seriously can contribute to violations of the rights to freedom of expression and personal integrity.[18]

The Oversight Board's decision in the case "Content Targeting Human Rights Defender in Peru", illustrates this challenge in practice.[19] The Board found that Meta's initial failure resulted from an overly narrow reading of the content, divorced from the broader environment of hostility and risk faced by defenders. By incorporating contextual factors, including patterns of intimidation, the Board concluded that the content constituted a threat and should have been removed, and it recommended clearer policy guidance to address veiled threats.

This approach aligns with international human rights standards requiring proactive measures to prevent foreseeable harm. When platforms fail to integrate off-platform and cumulative context into their assessments, they risk enabling coordinated harassment campaigns to persist under the guise of isolated incidents. In this regard, account-based enforcement measures are essential, as they allow platforms to address patterns of conduct rather than isolated expressions, and to prevent ongoing intimidation that undermines the work and safety of journalists and other human rights defenders.

III. Recommendations to Meta for Account Integrity Policy

In light of the observations presented, we suggest the Oversight Board recommend Meta to:

1. Explicitly recognize human rights defenders, including journalists and other defenders regardless of public notoriety, as a category requiring heightened protective measures due to their exposure to chilling effects from threats and harassment.
2. Incorporate explicit provisions for disabling or restricting accounts that demonstrate persistent patterns of abuse, harassment, or threats against defenders, even if individual posts do not independently meet a high threshold of explicit violence.
3. Mandate contextual and cumulative analysis in enforcement decisions, acknowledging that patterns of hostile conduct and relevant off-platform information may elevate the harmful potential of content.
4. Clarify in Community Standards that veiled, implicit, or coded threats are prohibited when directed at human rights defenders and journalists, as recognized by the Oversight Board's case "Content Targeting Human Rights Defender in Peru", pending compliance, and ensure guidance for moderators reflects this requirement.

[18] I/A Court H.R., *Case of Bedoya Lima et al. v. Colombia*. Merits, Reparations and Costs. Judgment of August 26, 2021. Series C No. 431, at para. 145.
[19] Oversight Board, "Content Targeting Human Rights Defender in Peru", May 27, 2025.