Content Targeting Human Rights Defender in Peru

2025-012-FB-UA

Summary

The Oversight Board overturns Meta's decision to leave up content targeting one of Peru's leading human rights defenders. Restrictions on fundamental freedoms, such as the right to assembly and association, are increasing in Peru, with non-governmental organizations (NGOs) among those impacted. Containing an image of the defender that has been altered, likely with AI, to show blood dripping down her face, the post was shared by a member of La Resistencia. This group targets journalists, NGOs, human rights activists and institutions in Peru with disinformation, intimidation and violence. Taken in its whole context, this post qualifies as a "veiled threat" under the Violence and Incitement policy. As this case reveals potential underenforcement of veiled or coded threats on Meta's platforms, the Board makes two related recommendations.

About the Case

A member of La Resistencia posted a likely AI-manipulated image, in which the headshot of the leader of a human rights organization has been altered to show her face covered with blood. A caption in Spanish insinuates that non-governmental organizations (NGOs) are engaging in financial wrongdoing by receiving foreign funds, also accusing them of encouraging violent protests. At the time this post was shared, there were ongoing demonstrations by civilians in Peru against the government.

Viewed around 1,000 times, the post was reported. Meta determined there were no violations. The user who had appealed to Meta then came to the Board. Before the Board selected the case, Meta received a report from one of its Trusted Partners, a global network of NGOs, humanitarian agencies and human rights researchers that flag emerging risks from content on Meta's platforms. As a result, Meta reviewed the

account posting the image and disabled it for violating its Terms of Service, meaning this specific post is no longer on Facebook.

Key Findings

The Board has unanimously found this post qualifies as a "veiled or implicit" threat under the Violence and Incitement Community Standard. When threats are veiled, they require a threat signal – such as a retaliatory statement or call to action – and a context signal, including local experts confirming the statement could lead to imminent violence.

The AI-manipulated image has a target – the human rights defender who is clearly identifiable to many Peruvians. Her image has been edited to look like she has sustained physical injuries. The text sets out grievances against NGOs, including alleged financial wrongdoing. Together, these factors meet the requirement for a threat signal. The content also satisfies the need for a context signal, since attacks against human rights defenders, including by La Resistencia, are well reported in Peru. Additionally, the Trusted Partner report sent to Meta highlights how this post could have contributed to imminent violence.

Meta interpreted this image to be a human rights defender with "blood on her hands." The Board is unpersuaded and disappointed by this explanation, noting the image is altered to indicate a bloody head wound. Meta's internal teams could easily have discovered that the defender is recognizable through a search online, which would have brought up her original smiling headshot.

No intervention short of content removal would have adequately mitigated the risks to the human rights defender in this case. Recent reporting by the UN has discussed the unsafe environment for defenders, especially women, in Peru. The stigmatization of civil society groups has created an atmosphere of fear, and this dynamic has been exacerbated by legislative initiatives that seek to assert more control over NGOs and restrict peaceful assembly.

Finally, the Board has received reports that this content has been reposted by other accounts associated with the same user who originally posted it. Meta should ensure such posts are removed unless they are for condemnation or awareness-raising.

The Oversight Board's Decision

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

The Board also recommends that Meta:

- Clarify that "coded statements where the method of violence is not clearly articulated" are prohibited in written, visual and verbal form, under the Violence and Incitement Community Standard.
- Produce an annual accuracy assessment on potential veiled threats, including a specific focus on content containing threats against human rights defenders that incorrectly remains up on the platform and instances of political speech incorrectly being taken down.

Full Case Decision

1. Case Description and Background

In July 2024, a Facebook user in Peru posted a digitally altered headshot of a well-known leader of a human rights organization in Peru. She is clearly identifiable in the likely AI-manipulated image, which shows her face covered with blood that drips downwards. A caption in Spanish insinuates financial wrongdoing by non-governmental organizations (NGOs) in receiving foreign funds and accuses NGOs of encouraging violent protests. The post was shared around the time citizens demonstrated against the government in Peru's capital, Lima. It was viewed around 1,000 times and had under 100 reactions.

Three days after the content was posted, a user reported it for violating Meta's policies. A human reviewer determined the content was not violating and it was kept up on the platform. The user appealed Meta's decision, but that appeal was automatically closed without further review. The same user then appealed to the Board.

In the time between the user appealing to the Board and the Board selecting the case, the post was also reported to Meta through its Trusted Partner Program. This is a network of NGOs, humanitarian agencies and human rights researchers from 113 countries that reports content and provides feedback to Meta about its content policies and enforcement. Following this report, Meta's internal escalation teams reviewed the account associated with the post and found it violated Meta's Terms of Service because the user had multiple accounts under the same or similar name. Meta then disabled the account, making the content inaccessible on Facebook. Consequently, the content was not assessed further.

When the Board selected this case, Meta reviewed the post again, confirming its original decision that the content did not violate its policies.

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

Peru has faced an "acute political and social crisis," having had six different presidents and three legislatures since 2016 (see 2023 report of the Office of the United Nations High Commissioner for Human Rights (OHCHR)). Political instability came to a head in December 2022, when former President Pedro Castillo was removed from office and arrested on charges of abuse of authority. Dina Boluarte, then vice president, was appointed to the presidency by Congress. The OHCHR noted that the events of December 2022 "triggered protests throughout the country, channeling social discontent ... over the historical discrimination and political and socioeconomic marginalization faced by a large part of the population."

Since then, the UN and human rights groups have expressed concern about government restrictions on basic human rights in Peru, particularly the right to

assembly, including protest. A 2024 report by the UN Special Rapporteur on the rights to freedom of peaceful assembly and of association noted allegations of "excessive, indiscriminate and disproportionate use of force and firearms, extrajudicial executions, and mass arbitrary detentions" in the context of mostly peaceful protests that began in December 2022. Additional protests against the government have since occurred, including in July 2024 when the content in this case was posted. Another "major concern" is the "alleged stigmatization and criminalization of human rights defenders ... and the lack of effective protection for defenders at risk" according to the 2024 report. As restrictions on fundamental freedoms around assembly and association have increased, the UN Special Rapporteur on the situation of human rights defenders has documented how the position of human rights defenders has become more precarious. These trends particularly affect women who face threats as a result of both their activism and gender identity. The Special Rapporteur noted that "the type of harassment they suffered was often of a discriminatory, misogynistic and sexual nature." The "lack of a systemic and intersectional approach by the authorities" has also hampered women defenders when filing complaints and seeking remedy and reparation.

The work of human rights defenders has also been threatened by legislative proposals to limit NGOs' activities. In June 2024, the Peruvian Congressional Foreign Relations Commission proposed draft amendments to the 2022 law creating the Peruvian Agency for International Cooperation. These amendments would, Peruvian civil society actors warned, impede international funding for NGOs and restrict their freedom of expression. In March 2025, Peru's Congress passed them. If enacted, penalties also could be imposed against civil society organizations for taking legal action against the state on human rights abuses. Human Rights Watch has tracked how this type of "foreign influence style legislation" around the world can stigmatize independent civil society and "offer a handy tool to discredit" the promotion of human rights by equating it with "promoting the interests of a foreign power."

In Peru, these proposals have advanced alongside political lobbying by right-wing-groups, as well as social media campaigns targeting NGOs and human rights defenders

with accusations of terrorism (referred to as “terruqueo” in Peru) and inciting violence at protests (see PC-30930), as per this post. The user who created the post in this case is an influential member of La Resistencia, a loosely organized group of right-wing activists. Formed in 2018, the group has targeted journalists, NGOs, human rights defenders and public institutions through disinformation campaigns, intimidation and violence.

2. User Submissions

In their appeal to the Board, the user who reported the content said that the post was a “thinly veiled death threat” against a human rights defender. They added that the post should be interpreted within a context of “harassment and physical attacks” against human rights defenders in Peru and that it was shared in response to the July 2024 demonstrations. They explained that the user who posted the content is a member of a group known for inciting violence and that such online threats have escalated into offline violence.

3. Meta’s Content Policies and Submissions

I. Meta’s Content Policies

Meta’s Violence and Incitement Community Standard “aims to prevent potential offline violence that may be related to content on [its] platforms.” The policy rationale notes that Meta removes “language that incites or facilitates violence and credible threats to public or personal safety.” The company also tries “to consider the language and context in order to distinguish casual or awareness-raising statements from content that constitutes a credible threat to public or personal safety.” The policy says that Meta removes content including “threats of violence against various targets.” The company defines “threats of violence” as “statements or visuals representing an intention, aspiration, or call for violence against a target.”

Under a subheading stating that Meta requires “additional information and/or context to enforce,” the Community Standard notes that Meta removes “coded statements

where the method of violence is not clearly articulated, but the threat is veiled or implicit, as shown by the combination of both a threat signal and context signal." A threat signal can include a coded statement "shared in a retaliatory context" or a coded statement that "acts as a threatening call to action." A context signal can mean confirmation from local experts or information that the statement could lead to imminent violence. A context signal can also be the target of a threat reporting the content.

Internal guidelines to reviewers clarify how threats of violence can exist in visual form, including digitally generated or altered imagery. To determine if digitally altered or created imagery is targeting someone with a threat visually, Meta considers factors such as whether there is a target in the image or whether the imagery shows an intent to target the person with high-severity violence.

II. Meta's Submissions

Meta stated this post does not violate the policy line prohibiting threats against targets, including digitally altered imagery depicting visual violence, because it does not contain a clear threat. According to Meta, the text of the post "levies allegations of corruption and violence on the part of NGOs . . . This focuses on unspecified NGOs and their activities – it neither identifies a target nor makes a threat." Meta noted the challenges of moderating content that makes allegations of corruption or malfeasance directed at NGOs and human rights defenders. The company stated that it was "not in a position to determine the truth or falsity of allegations" and "does not want to impede on what may constitute political speech about corruption or wrongdoing." However, Meta did acknowledge that "in some instances, with additional context, these allegations can contribute to a risk of offline harm and should be removed in the interest of safety."

For Meta, the image in the post was a "closer call [to a violation], as it appears to be digitally altered to show [the human rights defender] covered in blood." In its analysis of the image, Meta found that "while there is blood in the imagery, the human rights

defender does not appear to be injured. Rather, she appears unruffled, even smiling slightly, and looks directly at the camera. There is no pain in her expression that suggests she is injured, nor is the blood coming from any visible cuts or injuries." Meta also emphasized that signs of high-severity violence can be important when determining what constitutes a visual threat. Meta said, "had [the image] been visually altered to show visible stab wounds or other high-severity injuries, that could constitute a visual threat." However, it concluded that when "understood alongside the caption," the "more obvious meaning of the image" is that the human rights defender has "blood on her hands" due to the NGO's "alleged actions described in the text."

Meta's prohibition on "coded statements where the method of violence is not clearly articulated, but the threat is veiled or implicit" is assessed on escalation only. Meta requires a threat signal and context signal to enforce the "veiled threats" policy line. Meta reported that, as the account had been disabled, it did not reach out to a broad cross-functional team or external parties to assess the post as a veiled threat because this would involve "resource-intensive investigatory steps." Meta said it would have done this more in-depth review had the content remained live on Facebook. The company did say that it "likely would not constitute a veiled threat because the image [of the human rights defender] could be understood as political critique, rather than visual violence." In response to a request for data on veiled threats, Meta stated it does not track the number of pieces of content reviewed for potential veiled threats.

Meta said that it "proactively engages with human rights defenders to understand their needs and strives to offer specific measures to protect their safety and mitigate the risks they face." According to the company's Human Rights Policy, the term "human rights defenders" includes "human rights organizations, members of vulnerable groups advocating for their rights; professional and citizen journalists; non-violent political activists, and any member of the public who raises a human rights concern."

The Board asked questions on the veiled threats framework; how Meta understands and assesses visual threats in images; the role Trusted Partners play in providing additional information and context for assessments of the likelihood of offline harm;

and how Meta protects human rights defenders on its platforms. Meta responded to all questions.

4. Public Comments

The Oversight Board received 65 public comments that met the terms for submission. Of these, 60 were submitted from Latin America and the Caribbean, three from Europe and two from the United States and Canada. To read public comments submitted with consent to publish, click here. Personally identifiable information has been redacted from public comments.

The submissions covered the following themes: the social and political context in Peru; the situation of human rights defenders; gendered dimensions of threats against defenders; recent legislative initiatives that impact the activities of NGOs in Peru; social media narratives accusing NGOs, human rights defenders and civil society groups of "terrorism"; the operations of La Resistencia; and how Meta should moderate content including possible veiled threats.

In January 2025, as part of stakeholder engagement, the Board consulted with advocacy organizations, academics, inter-governmental organizations and other experts on protecting human rights defenders online. This roundtable focused on threats that human rights defenders face and previous campaigns to implement policy recommendations for their protection at social media companies. Participants also discussed reporting content they thought was likely to cause offline harm through Meta's Trusted Partner program.

5. Oversight Board Analysis

The Board selected this case to examine how Meta's policies protect human rights defenders, particularly when threats of violence are veiled or implicit, require additional context to interpret, or occur within an environment of intimidation and harassment. This case falls within the Board's strategic priority of Elections and Civic Space.

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 unanimously finds that the post violates the Violence and Incitement policy. Viewed in context, the combination of the text with the image of the bloodied human rights defender meets Meta's definition of a prohibited threat. The Board is unanimous that the post qualifies as what is described by Meta as a "veiled" threat, which categorizes potentially ambiguous posts as threats if they have both a "threat signal" and a "context signal" that together make up an implied or disguised threat.

- Threat signal: The post satisfies the "threat" signal requirement specifically as "a threatening call to action" or an "expression of desire to engage in violence" in retaliation for alleged misdeeds by NGOs. The image contains a target in the form of the human rights defender, who would be clearly identifiable to Peruvian users. The text lays out grievances against NGOs, including financial wrongdoing and incitement of violence at protests, alongside an altered image of the human rights defender, edited to clearly depict her with blood and injuries sustained in an attack. The Board is unpersuaded and disappointed by Meta's surprising conclusion that the image of the human rights defender with blood dripping down her face signified "blood on her hands" and thus constituted "political critique." The human rights defender "appears unruffled" and "has no pain in their expression" because the image is a digitally altered version of a professional headshot in which she is smiling. The individual's hands are not even visible. Meta's internal teams could easily have discovered that the individual is recognizable. The Board is not aware of what reverse image search tools Meta makes available to its moderators but presumes that Meta has the technical sophistication to provide moderators with the information they need to assess

images. While no wound is visible in the altered version of the image, the pattern of blood dripping downwards from one side of the head and out of the subject's eyes indicates that it originates in a head wound.

- Context signal: The post satisfies the "context" signal because "[l]ocal context or expertise confirms that the statement in question could lead to imminent violence." This conclusion is based on contextual information that similar accusations in Peru have been credibly identified as the impetus for people to target human rights defenders with intimidation and violence. The OHCHR documented several attacks by La Resistencia on human rights organizations, during which the organizations were accused of being "pro-terrorist" and encouraging violence at protests. The Committee to Protect Journalists reported that at one La Resistencia gathering, participants shouted threats including "your days are numbered" and "you will die" at individuals inside a media outlet's office. The contextual risk was also highlighted in this case by a report from a Trusted Partner stating that the content could contribute to imminent violence. Human rights defenders have also made their concerns about threats of violence and abuse known to the company through reports, strategic litigation and stakeholder events.

Some Board Members consider that it is not necessary to rely on the "veiled threats" analysis to conclude that the post violates the Violence and Incitement policy. The Board has repeatedly emphasized that Meta must assess posts as a whole and in context (see Wampum Belt, Iran Protest Slogan, Violence Against Women and Statements about the Japanese Prime Minister). Meta's Violence and Incitement policy prohibits statements "representing an intention, aspiration, or call for violence against a target." For these Board Members, the image of the identifiable human rights defender covered in blood, presented with a caption that alleges wrongdoing, is a clear "expression of hope," "aspiration," or "call for action" in the form of "high-severity violence." For these Board Members, this can only mean that the individual depicted is being targeted with imagery representing a call for violence and showing intent to target the individual with high-severity violence.

II. *Enforcement Action*

This case raises concerns that the operational distinction between threats that do and do not require context to enforce is resulting in underenforcement and more veiled threats remaining on Meta's platforms.

The Board has previously recognized the challenges of enforcing against veiled threats of violence (see UK Drill Music, Protest in India against France and Knin Cartoon) because sufficiently deep contextual analysis is sometimes only undertaken on escalation. However, it has also recommended context-specific applications of Meta's policies on threats that could be applied at scale (see Statements About Japanese Prime Minister). Meta's current guidance for at-scale reviewers significantly limits the possibility of contextual analysis (see Violence Against Women). At-scale moderators are not instructed or empowered to identify content that violates the company's escalations-only policies, like the rule at issue in this case (see Sudan's Rapid Support Forces Video Captive). This means that the human reviewer in this case would not have been able to exercise discretion and judgment in evaluating the content when initially reported or to escalate the content to teams empowered to enforce the context-sensitive policy line.

Because Meta does not track the number of pieces of content reviewed for veiled threats, the Board could not assess the prevalence of veiled threats or the magnitude of underenforcement. However, even if veiled threats against human rights defenders are found to be a "low prevalence" issue, the impact is still high and acutely felt by human rights defenders who are threatened, discouraged from pursuing their work and faced with physical violence. To address this risk, Meta should invest in regular, high-quality assessments of its performance to identify opportunities to improve enforcement in this area. Meta should develop a better understanding of how prevalent veiled threats are on its platforms, and how accurately its systems detect and enforce against this content. This work could also serve as the basis for eventually generating more granular metrics, such as the prevalence of threats targeting human rights

defenders, and targeted evaluation mechanisms. As part of this process, Meta could experiment with developing an automated tool to identify potential veiled threats, to be enqueued for review by the relevant escalations team.

Finally, Trusted Partners play an important role in identifying potentially violating content, including veiled or coded threats, and providing information needed for accurate enforcement. The Board has previously addressed issues related to Meta's responsiveness to Trusted Partner reports (see Haitian Police Station Video). As Meta has indicated its intent to rely less on automated systems and more on user reports to detect violating content, the Trusted Partner program is an important channel for surfacing emerging risks and identifying mistakes. Meta should ensure that the program is adequately resourced and supported so that its internal teams are able to make enforcement decisions that benefit from Trusted Partners' expertise and contextual insights.

5.2 **Compliance With Meta's Human Rights Responsibilities**

The Board finds that the removal of the content from the platform, as required by a proper interpretation of Meta's content policies, is consistent with Meta's human rights responsibilities.

Freedom of Expression (Article 19 ICCPR)

Article 19 of the International Covenant on Civil and Political Rights (ICCPR) provides for broad protection of expression, including views about politics, public affairs and human rights (General Comment No. 34, paras. 11-12). 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, which Meta itself has committed to in its Corporate Human Rights Policy. 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).

I. Legality (Clarity and Accessibility of the Rules)

The principle of legality under international human rights law requires rules limiting expression to be clear and publicly accessible (General Comment No.34, at para. 25). Legality standards further require that rules restricting expression "may not confer unfettered discretion for the restriction of freedom of expression on those charged with [their] execution" and "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," (A/HRC/38/35, at para. 46). People using Meta's platforms should be able to access and understand the rules and content reviewers should have clear guidance on their enforcement.

The Board finds that Meta's prohibitions on "threats of violence against various targets" and "coded statement where the method of violence is not clearly articulated, but the threat is veiled or implicit" are sufficiently clear as applied in this case.

However, the Board notes that, while the Violence and Incitement policy's at-scale prohibition against threats stipulates that "threats of violence are statements or visuals," the escalation-only policy lines on veiled threats focus on "coded statements." The Board recommends that Meta clarify this language so it is clear that "coded statements" containing threats in written, visual and verbal forms are prohibited. The Board has previously called on Meta to develop policies and enforcement guidelines that treat posts containing text and image holistically (see Post in Polish Targeting Trans People). This is particularly important for content like the post in this case, where context is required to understand its meaning as a whole. Meta's analysis of the post does this inconsistently: it considers the text when interpreting the image but does not consider the image when interpreting the text. Moreover, the Board notes that Meta's

chosen terminology of "veiled threat" may be misleading for some users as it suggests a threat is disguised or, perhaps, even less severe. While the post in this case requires some interpretation to understand, it is clearly intended to deliver a threatening message.

II. Legitimate Aim

Any restriction on freedom of expression should also pursue one or more of the legitimate aims listed in the ICCPR, including the aim of protecting the rights of others. The Violence and Incitement Community Standard aims to "prevent potential offline violence" by removing content that poses "a genuine risk of physical harm." This policy serves the legitimate aim of protecting the rights to freedom of expression and assembly (Articles 19 and 21, ICCPR) and the right to life and the right to security of person (Article 6, ICCPR; Article 9 ICCPR).

III. Necessity and Proportionality

Under ICCPR Article 19(3), necessity and proportionality requires that restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected," (General Comment No. 34, para. 34).

When analyzing the risks posed by violent content, the Board is typically guided by the six-factor test described in the Rabat Plan of Action. Although the Rabat framework was created to assess advocacy of national, racial or religious hatred that incites acts of discrimination, hostility or violence, the test is useful for evaluating incitement to violence generally (see Iran Protest Slogan and Call for Women's Protest in Cuba). Based on an assessment of the relevant factors, especially the content and form of expression, the intent of the speaker and the context below, the Board finds that removing the content is a necessary and proportionate limitation on expression in order to protect the right to life and the security of the human rights defender. The post both identifies

the human rights defender and threatens her with violence. No intervention short of removal would adequately mitigate the risks the post presents.

Allegations of criminality and corruption targeting NGOs and human rights defenders in Peru have frequently been used to mobilize demonstrations by La Resistencia, which have resulted in attacks. The post both identifies the human rights defender personally and threatens her through i) an altered picture depicting the aftermath of a violent injury, and ii) a caption invoking narratives that have been used to mobilize such attacks.

The content was posted on the same day demonstrations against the Peruvian government were held to criticize the state's "infringements working against the interests of the great majority." The "alleged stigmatization and criminalization of human rights defenders, the persistent problematic practices of the state's response within the context of social protests, and the lack of effective protection for defenders at risk" have been noted as a major concern in Peru by the UN Special Rapporteur on the rights to freedom of peaceful assembly and association. Furthermore, the situation of human rights defenders should be understood within the context of threats, intimidation, harassment and physical attacks from groups like La Resistencia (see PC-30929, PC-30927, PC-30930 and PC-30932). The post was made by an influential member of La Resistencia with a significant social media following. Research commissioned by the Board highlighted the user's public role in organizing gatherings that target journalists and human rights defenders with death threats and intimidation. The Board finds that any threat of violence from such users risks likely near-term harm against depicted targets. The Board underscores that criticism of NGOs is permitted, but credible threats of violence are not.

Recent reporting corroborates the seriousness of threats that target human rights defenders. The UN Special Rapporteur on the situation of human rights defenders has said that defenders are increasingly being threatened and "death threats that often precede the killing of human rights defenders" are of particular concern. The Special Rapporteur also noted that "many threats are gendered," targeting women human rights defenders. In Peru, the Special Rapporteur has found that a "large number of

human rights defenders are unable to operate in a safe and enabling environment." Through a public information request it made to Peruvian authorities in 2023, the human rights organization Amnesty International noted that the state's Protection Mechanism for Human Rights Defenders registered 197 threats against human rights defenders and/or their families, 60 of whom were victims of physical and/or verbal abuse (see PC-30928). Amnesty International also confirmed the killing of at least four human rights defenders in Peru in 2023.

In addition to creating physical risks, content that targets human rights defenders with threats and is accompanied by narratives used to mobilize groups that have attacked NGOs, even if implicit and requiring context to understand, perpetuates an atmosphere of fear and fosters an environment in which the targeting of civil society groups more broadly is normalized. Practically, this makes it more difficult for human rights defenders to do the work of protecting the rights of others. Stigmatization, according to a UN Report, is "inherently connected" to the erosion of the underlying human rights for which defenders advocate. In Peru, this dynamic has been exacerbated by legislative initiatives that seek to assert more government control over NGOs and place restrictions on the rights to freedom of peaceful assembly and association at protests. As the International Center for Not-For-Profit Law (PC-30930) noted in its public comment: "The work of human rights defenders is essential for strengthening democracy and the rule of law . . . respect for human rights in a democratic society largely depends on effective and adequate guarantees for human rights defenders that enable them to carry out their activities freely."

Due to the fear of being targeted, threats can produce chilling effects on the freedom of expression of human rights defenders, especially women. Women often play a disproportionate role in advocating and organizing for equal rights, and they are disproportionately singled out for threats and abuse. Women human rights defenders have also raised concerns about the severe risks of overenforcement when Meta interprets political speech that uses violent metaphors or draws attention to human rights abuses (see Iran Protest Slogan). As the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression noted in a 2023 report, gendered disinformation and online abuse "may lead women in public life or journalism

to leave online spaces or may have a chilling effect on their freedom of expression and on their ability to continue their professional activities." Stakeholders participating in the Board's roundtable on "Protecting Human Right Defenders Online" shared that defenders, especially women, women of color and queer women, find themselves in a catch-22: they depend on Meta's products for their work, but simultaneously are subjected to harassment and threats on the company's platforms.

5.3 Identical Content With Parallel Context

The Board has received reports (PC-30929) that the content, despite being inaccessible following deactivation of the user's account, has been reposted from different accounts associated with the user. Following this decision, Meta should ensure that identical content is removed, unless it is shared in a condemning or awareness-raising context.

6. The Oversight Board's Decision

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

7. Recommendations

Content Policy

1. To ensure that its Violence and Incitement Community Standard clearly captures how veiled threats can occur across text and imagery, Meta should clarify that threats made out of "coded statements," even "where the method of violence is not clearly articulated," are prohibited in written, visual and verbal form.

The Board will consider this implemented when the public-facing language of the Violence and Incitement Community Standard reflects the proposed change.

Enforcement

2. To ensure that potential veiled threats are more accurately assessed, in light of Meta's incorrect interpretation of this content on-escalation, the Board recommends that Meta produce an annual assessment of accuracy for this problem area. This should include a

specific focus on false negative rates of detection and removal for threats against human rights defenders, and false positive rates for political speech (e.g., Iran Protest Slogan). As part of this process, Meta should investigate opportunities to improve the accurate detection of high-risk (low-prevalence, high impact) threats at scale.

The Board will consider this implemented when Meta shares the results of this assessment, including how these results will inform improvements to enforcement operations and policy development.

***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.