Public Comment for Oversight Board
Content Targeting Human Rights Defender in Peru

Prepared by Ricki-Lee Gerbrandt and Jeffrey Howard (University College London)

Summary

- The post at issue should likely have been removed as a "veiled threat" violation of the Violence and Incitement Policy. However, because the strength of the relevant signals required to establish a veiled threat is mixed, the post's permissibility under the rules is ambiguous. In such cases, we argue that whether enforcement should be taken depends on the risks of harm. Given the serious vulnerability of human rights defenders, the risks of harm to such persons are high; so the post should have been treated as a threat and removed.
- The post at issue should have also been analysed as a potential violation of the Misinformation Policy, given the connection between falsehoods about human rights defenders and real-world violence against them. It should also have been assessed for using AI-manipulated media without disclosure.
- More broadly, Meta must be proactive in protecting human rights defenders and journalists on its platforms, both to protect individual targets' safety and to uphold freedom of expression, which is compromised when these actors are intimidated.

Background: Human Rights Defenders as a Vulnerable Population

It is essential to recognise the context in which human rights defenders currently work. Like journalists, human rights defenders are at increased susceptibility to abusive and threatening conduct, online and offline. Both groups engage in advocacy and communication concerning political issues, conduct investigations of powerful public and private people and institutions, and are engaged in difficult public-facing work. The risks to psychological, financial, and physical well-being that people face to carry out this work can be substantial–with serious knock-on effects for freedom of expression and democratic governance.

Reports Without Borders (RSF) reports that in Peru 'Journalists have continued to be the targets of attacks by far-right activists since 2018', and the situation has reportedly worsened since the political and social crisis that began in 2022.[1] The Oversight Board's Case Description notes that the user who posted the content is alleged to be a member of a group 'known for inciting violence against human rights defenders and journalists in Peru, and that such online threats have escalated into offline violence'. Scholarly empirical research at the global level has documented that online abuse and associated

[1] https://rsf.org/en/country/peru

disinformation tactics are typical tactics deployed to silence journalists and human rights defenders, to chill others from engaging in that work, and can destabilise the public's information environment and generally erode democratic norms. These studies have also documented the connection between online and offline violence.[2]

The content at issue here therefore must be assessed in light of the recurrent targeting of journalists and human rights defenders online, which demonstrates their vulnerability as a population. This vulnerability, we will argue, properly bears on our determination of the dangerousness of speech targeting them.

Background: Protections for Human Rights Defenders and Journalists in IHRL

International Human Rights Law (IHRL) protects the right to freedom of expression as a fundamental principle of democracy. In particular, it protects the rights of journalists and human rights defenders to speak – so they can investigate state officials and hold them accountable (the 'watchdog function'); and so they can educate and inform the public, enabling citizens to participate fully in democratic governance. When human rights defenders and journalists endure abuse, harassment, or threats, or are targeted with disinformation campaigns, those tactics wrongly intimidate their direct targets. But they also intimidate the wider population who engage in similar work, potentially chilling their speech. Democracy suffers, then, when these citizens are targeted.

The current case therefore engages the freedom of expression of the specific human rights defender subject to the content at issue. But it also engages the freedom of expression of others engaged in that work. And it further implicates the public's rights to receive information necessary for democratic governance and participation, which cannot be fulfilled without citizens willing to do the work of holding power to account.

Since 2017, the UN General Assembly and UN Human Rights Council began adopting resolutions condemning online abuse against journalists and specifically the online abuse of women journalists.[3] The UN Special Rapporteur on the Promotion and

[2] *See* Chris Tenove, Ahmed Al-Rawi, Juan Merchan, Manimugdha Sharma, and Gustavo Villela, 'Not Just Words: Reputational attacks against journalists', Global Reporting Centre (20 June 2023) <https://globalreportingcentre.org/reputational-attacks/>;
Julie Posetti, Jackie Harrison, and Silvio Waisbord. "Online Attacks on Women Journalists Leading to 'Real World' Violence, New Research Shows." ICFJ, (25 November 2020)< https://www.icfj.org/news/online-attacks-women-journalists-leading-real-world-Violence-new-research-shows>;
Julie Posetti and Nabeelah Shabbir, 'The Chilling: A global study of online violence against women journalists' *International Center for Journalists/UNESCO* (2 Nov 2022) https://www.icfj.org/sites/default/files/2023-02/ICFJ%20Unesco_TheChilling_OnlineViolence.pdf)

[3] The most recent resolutions are:
- UN General Assembly, The safety of journalists and the issue of impunity, A/RES/76/173 (2021);

Protection of the Right to Freedom of Opinion and Expression has also recently stepped in to condemn online violence against journalists and human rights defenders.[4] Further, international human rights tribunals, including the Inter-American Court of Human Rights, have held that states have positive obligations to protect media workers, to investigate wrongdoing, and to hold perpetrators to account to reduce impunity.[5]

Similarly, the European Court of Human Rights has also held that Member States have positive obligations to protect and investigate threats and abuse of journalists and must also ensure that the state *considers any impact on threats or violence against journalists on the journalists' ability to exercise their freedom of expression.* In particular, the State has *enhanced obligations to protect and investigate threats and attacks on journalists' physical and psychological safety* pursuant to art 10 ECHR.[6]

Given Meta's commitment to grounding content moderation in international human rights norms for freedom of expression, Meta should treat attacks on human rights defenders and journalists with the utmost seriousness. Such attacks inflict physical and psychological harm on their direct targets; but they also intimidate the wider community of human rights defenders and journalists, setting back free expression.

Importantly, users' freedom of expression plainly includes the prerogative to *criticize* human rights organizations and media organizations; so these groups are not immune from negative feedback. But such criticism must not take the form of threatening, inciting, and abusive attacks.

Violence & Incitement Policy

Having clarified the importance of protecting human rights defenders and journalists from attack, we now consider the post at issue in detail, and whether it should be understood as a violation of Meta's rules.

The post at issue depicts the leader of a Peruvian human rights organization, with blood running down their face, alongside a caption alleging wrongdoing. The user who reported the post alleged that it was a "thinly-veiled death threat" against the depicted human rights defender. Explicit threats are disallowed under Meta's Violence &

- UN Human Rights Council, 'The safety of journalists' A/HRC/RES/51/9 (2022); UN General Assembly, 'The safety of journalists and the issue of impunity' GA A/RES/78/215 (2023).

[4] Irene Khan, 'Reinforcing media freedom and the safety of journalists in the digital age' *UNHRC, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression* (20 April 2022) UN Doc A/HRC/50/29 118-121.

[5] *Bedoya Lima Y Otra v Colombia*, (IACtHR 26 August 2021) https://www.corteidh.or.cr/docs/casos/articulos/seriec_431_esp.pdf; and *Vélez Restrepo and family v Colombia* [2012] IACtHR, Ser. C, No. 248.

[6] *Khadija Ismayilova v Azerbaijan* App Nos. 65286/13 and 57270/14 (ECtHR, 10 January 2019) at [153].

Incitement Policy; implicit or "veiled" threats are disallowed only if certain further signals are established. Specifically, a veiled threat is disallowed if at least one *threat signal* is present and at least one *contextual signal* is present.

Was a *threat signal* satisfied in this case? One kind of threat signal involves speech "shared in a retaliatory context" (e.g., "expression of desire to engage in violence against others in response to a grievance…"). Given the caption, the intimation could be that the target deserves to be harmed in retaliation for the alleged wrongdoing. Another kind of threat signal involves "references to historical or fictional threats of violence." Given the aforementioned history of violence of human rights defenders, it is possible a post depicting violence recalls precisely that history (though this may be a stretch). A further kind of threat signal involves "acts as a threatening call to violence." One possible interpretation of putting blood on the target's face is that it is calling for audiences to bring that about. Thus there seem to be multiple potential threat signals; though none is individually strong, taken together they have some force.

Was a *contextual signal* satisfied in this case? One kind of threat signal is that "local context or expertise confirms that the statement in question could lead to imminent violence." Because the account that posted this content was suspended for independent reasons, Meta "did not reach out to a broad cross-functional team or external parties for additional input to inform its decision" (OSB Case Description). Given the aforementioned vulnerability of human rights defenders, it is conceivable that such an effort could well have established the relevant contextual signal.

Accordingly, there is a reasonable case to be made that the post at issue involved a veiled threat that violated Meta's rules. However, it must be recognized that the case isn't a knock-down, unequivocal one, given the contestable strength of the relevant signals. One can easily imagine the rebuttal: "The use of blood dramatically and hyperbolically insinuates that the relevant NGO is guilty of wrongdoing–and users must be free openly to discuss their views on such questions, as a matter of freedom of expression."

This case is thus an example of an *ambiguous threat*–where one plausible interpretation of a post is that it violates the rules against threats, whereas another plausible interpretation of a post is that it *doesn't* violate the rules. (The same phenomenon arose in the Iran Protest Slogan Case, though the Board didn't see it that way.) What should be done in such cases? Elsewhere one of us has argued that in such cases, the determination should hinge on the *potential harmfulness* of the speech. *In cases of ambiguous threats where the risk of harm to the target is low, the speech should be allowed to stand. In cases of ambiguous threats where the risk of harm is high, the speech*

should be treated as violating.[7] (In practice, this may mean lowering the required confidence of the relevant classifier for such cases.)

Putting our point another way: the *contextual signal* that a post poses a substantial real-world danger should be accorded special significance in determining whether the post should be removed as a veiled threat. So long as there is some minimal *threat signal* (such that there is a reasonable interpretation of the post's meaning that makes it a threat, even if not the only reasonable interpretation), the post should be removed if the *contextual signal* is very strong.

Is it fair to hold a user accountable for an ambiguous threat, given the possibility that the user did not *intend* to be understood to be issuing a threat? We think it can. Given the vulnerability of certain targets, *it is reasonable to demand that speakers take due care in how they express themselves.* When one reasonable interpretation of an utterance is that it constitutes a threat, speakers ought to express themselves more precisely. The use of blood in connection with a vulnerable population, as in this case, is unacceptable.

Misinformation Policy

While the central question of the case is whether the post is regulable under the Violence & Incitement Policy, it is worth exploring whether it is potentially regulable under other policies as well – specifically, the Misinformation Policy. Two aspects of the policy are relevant: one concerning violence, another concerning AI.

Misinformation and violence

The Misinformation Community Standard states: "We remove misinformation where it is likely to directly contribute to the risk of imminent physical harm. We also remove content that is likely to directly contribute to interference with the functioning of political processes. In determining what constitutes misinformation in these categories, we partner with independent experts who possess knowledge and expertise to assess the truth of the content and whether it is likely to directly contribute to the risk of imminent harm. This includes, for instance, partnering with human rights organisations with a presence on the ground in a country to determine the truth of a rumour about civil conflict."

It is obviously beyond our expertise to assess whether the allegations in the post – that the human rights defender is engaged in financial wrongdoing and is also inciting violent protest – are correct or not. Given the apparent lack of evidence offered (at least

[7] One of us co-defends this view in Sarah A. Fisher and Jeffrey W. Howard, "Ambiguous Threats: 'Death-to' Statements and the Moderation of Online Speech-Acts," *Journal of Ethics & Social Philosophy* 28, 2 (2024): 208-229.

based on the Case Description), there seems to be a real possibility that they are spurious, baseless rumours. If on-the-ground experts verified that such rumours posed a risk of inspiring real-world violence, this could be a reasonable basis for moderating the speech under the misinformation policy.

Rather than allow the post, then, Meta's human reviewer should have escalated it so that Meta could have considered the likelihood that the content would contribute to imminent harm – again given the special risks that human rights defenders and journalists endure in Peru. These contextual decisions again require Meta to partner with local human rights experts to gauge the social and political situation in that country to help determine the veracity of the content and make these important determinations. Here, whether this misinformation policy is applicable depends on whether trusted third parties have indeed established that allegations of financial wrongdoing by human rights NGOs (intimated in the post) are false.

Misinformation and AI

Meta's Misinformation Policy also "require[s] people to disclose, using our AI-disclosure tool, whenever they post organic content with photorealistic video or realistic-sounding audio that was digitally created or altered, and we may apply penalties if they fail to do so." Given the claim in the Case Description that the image is seemingly AI-manipulated, the post could have had penalties applied for failure to disclose AI-manipulation.

Note also that Meta has a policy whereby it "may also add a label to certain digitally created or altered content that creates a particularly high risk of misleading people on a matter of public importance". It appears it did not do so in this case, despite the fact that the subject matter of this post is indeed on a matter of public importance.

Crucially, the fact that the image was AI-generated does nothing to affect the analysis above re: whether it counts as a veiled threat or violence-promoting misinformation. As one of us has argued before in relation to the Altered Video of President Biden case, what matters when it comes to most violations is *the message* conveyed by a post, not the technology used to produce it.[8] Even so, there is a duty to disclose AI-manipulated media, and this provides another policy lever with which Meta could have addressed such content.

[8] See UCL Digital Speech Lab - Public Comment 18036, available at https://osbcontent.s3.eu-west-1.amazonaws.com/PC-18036.pdf, which was cited by the Board in its decision. The reasoning here is spelled out more systematically in Sarah A. Fisher, Jeffrey W. Howard, and Beatriz Kira, "Moderating Synthetic Content: The Challenge of Generative AI," *Philosophy & Technology* 37 (2024): 133.

Further Points

Note that we pointed above to the provisions that prohibit misinformation when linked to real-world physical violence. But we stress that misinformation can cause harm to human rights defenders and journalists beyond endangering their physical safety. Defamatory smears can cause serious harm to targeted individuals. Further, if users post messages containing such serious allegations, without any underlying facts or evidential basis, that can significantly harm and deteriorate the information environment. These are long-established insights in defamation law theory.

While it may not be feasible for Meta to have generally enforceable community standards against defamation (something not at issue here), our point is simply that "imminent violence" is not the only real-world harm to which misinformation can lead. With regard to vulnerable groups like human rights defenders, Meta should consider revising its misinformation policy to allow a wider array of relevant harms to be considered.

(One alternative place to address such a concern is in the policy on Bullying & Harassment. Had the target of this post been subjected to "directed mass harassment" on the basis of this post, the post might have violated the Bullying and Harassment Policy's special protections for "human rights defenders." But it appears the account was suspended before this was possible.)

Conclusions

Our narrow conclusions for this case:

- The post at issue should likely have been removed as a "veiled threat" violation of the Violence and Incitement Policy. However, because the strength of the relevant signals required to establish a veiled threat is mixed, the post's permissibility under the rules is ambiguous. In such cases, we argue that whether enforcement should be taken depends on the risks of harm. Given the serious vulnerability of human rights defenders, the risks of harm to such persons are high; so the post should have been treated as a threat and removed.
- The post at issue should have also been analysed as a potential violation of the Misinformation Policy, given the connection between falsehoods about human rights defenders and real-world violence against them. It should also have been assessed for using AI-manipulated media without disclosure.

Our broader guidance for rethinking and developing policy:

- The Oversight Board Meta must be proactive in protecting human rights defenders and journalists on its platforms, both to protect individual targets and to uphold freedom of expression.[9] It should:
    - conduct risk assessments to ascertain the risks that human rights defenders and journalists face on each Meta platform, taking into account jurisdictional differences;
    - analyse those risk assessments and take measures to mitigate risks;
    - provide a quick process for human rights defenders and journalists to report abuse and to have that content quickly actioned;
    - implement policies and resources to detect and prevent repeat attackers from continued abuse and assuming new identities;
    - develop and implement tools and other resources to predict when human rights defenders and journalists will face an onslaught of abuse and take appropriate action;
    - provide data to researchers, including those working on developing tools to monitor and track online abuse;
    - compile and provide information to the human rights defender or journalist and, as appropriate, to police.

Submission Prepared By:

Ricki-Lee Gerbrandt is Fellow in Law and Platform Governance at University College London, based in the Digital Speech Lab.

Jeffrey Howard is Professor of Political Philosophy & Public Policy at University College London, where he directs the Digital Speech Lab, and Senior Research Associate at the Oxford Institute for Ethics in AI.

About the Digital Speech Lab

The Digital Speech Lab hosts a range of research projects on the proper governance of online communications. Its purpose is to identify the fundamental principles that should guide the private and public regulation of online speech, and to trace those principles' concrete implications in the face of difficult dilemmas about how best to respect free speech while preventing harm. It is funded by a Future Leaders Fellowship awarded to Jeffrey Howard from UK Research and Innovation. Thanks to UKRI (grant reference MR/V025600/1) for enabling this work. Find out more at www.digitalspeechlab.com.

[9] For development of some of these points, see Ricki-Lee Gerbrandt, "Media Freedom and Journalist Safety in the UK Online Safety Act," *Journal of Media Law* 15, 2 (2023): 179-212.