The Lima Hub of AI for Good's Young AI Leaders Community represents a pioneering initiative in Latin America that seeks to harness the transformative potential of artificial intelligence to advance social good, strengthen democratic institutions, and promote inclusive development.

Based in Peru, the Hub convenes experts, policymakers, civil society actors, and the private sector to foster responsible AI practices aligned with human rights and the rule of law. Its mission is to ensure that technological innovation is not only accessible but also oriented toward solving pressing social challenges—from access to justice and education to transparency, accountability, and civic participation.

The Lima Hub stands as a regional reference point for ethical and human-centered AI, amplifying Latin American voices in global debates about the governance of digital technologies. Upon the Oversight Board's request for opinions in the case 2025-056-FB-MR, the Lima Hub has unanimously decided to contribute.

I. Freedom of expression on social media: citizen scrutiny of public officials

Ever since social media became widely accessible, citizens have turned to these platforms as a coping mechanism and a collective space to rally. For many, it offered a buffer against the indifference and silence of public officials. Therefore, social media became a space where the voices of the people could finally be heard.

Over the past decade, social media has evolved far beyond a mere tool for social engagement, due to becoming a global beacon of freedom of expression, enabling individuals from different cultures, nationalities, economical status, ethnicities, as well as different communities to amplify their concerns, demand accountability, and defend human rights, democracy and rule of law with a global reach and impact previously unimaginable.

Furthermore, the existence of social media guarantees the possibility of a fluid dialogue between civil society and those in power, that "enables citizens to immediately and visibly denounce injustices, irregularities, arbitrary acts, or corruption perpetrated by public officials. On several occasions, this has resulted in the public administration being forced to take concrete measures to remedy the situation as quickly as possible"[1].

In current times, when authoritarian regimes want to silence civil society, they resort to drastic measures such as shutting down access to the internet nation-wide or ban specific social media platforms such as Facebook, Instagram, Whatsapp and X. For that reason, the Special Rapporteur on the rights to freedom of peaceful assembly and of association considers that "ending shutdowns has become a human rights imperative both to allow people to exercise their rights online and offline and to safeguard democratic governance in the digital era"[2].

Moreover, public office, by its very nature, subjects those who hold it to heightened scrutiny. Although officials are ordinary individuals, the functions they perform are of undeniable public interest. It should

[1] Diaz Giunta, R. (2022). El Derecho a la Libertad de Expresión y las Redes Sociales. In: Diaz Giunta, R., & Roel Alva, L. (Coordinators). *Athina: Edición Especial Bicentenario*, p. 103.
[2] Voule, C. (2021). *Ending Internet shutdowns: a path forward*. Geneva: United Nations, p. 17.

therefore come as no surprise that citizens turn to social media to voice criticism, expose misconduct, or denounce potential wrongdoing. Enduring such scrutiny is not an exception, but an inherent condition of serving in public office, where accountability and tolerance are essential to democratic life.

A public official's legitimacy does not just come from being elected or appointed, but from always being accountable to the people. Today, a phone and an internet connection are enough to record an abuse and share it instantly, so online complaints force responses and dismantle the idea of immunity.

Social media has also exposed cases of favoritism in vaccination campaigns, misuse of funds, and nepotism. Public pressure online has compelled many officials to provide explanations, resign, or face legal proceedings. Criticism in this sense is not an attack but a safeguard to ensure power is exercised properly.

Courts have further recognized that official social media accounts function as public forums. In Garnier v. O'Connor-Ratcliff, the Ninth Circuit held that blocking critics on an official account violated free speech rights, affirming that democratic debate extends into the digital sphere[3].

For this reason, public servants should avoid mixing personal posts with official messages. Establishing clear participation rules and responding openly to citizens' concerns not only reduces legal risks but also builds trust in institutions. Rather than fearing online scrutiny, officials who listen and adapt often gain credibility.

Ultimately, online conversation functions like a modern town square: a space where ideas are exchanged and leaders are held accountable. Promoting respectful and open dialogue helps build a culture of transparency that strengthens democracy.

II. Discourse inciting violence against family members of public officials, with special emphasis on the protection of minors

In order to best resolve the case, it is necessary to define **what is meant by incitement to violence**. According to Meta's Transparency Centre, "language that incites or leads to acts of violence and credible threats to public or personal safety is removed. This includes incitement to violence directed at a person or group of people on the basis of their protected characteristics"[4]. In this regard, from the review of the standards of the Inter-American Human Rights System, there are two main elements for identifying violence and incitement online, which follow a similar logic to Meta's Community Standards. First, there must be a clear intention to incite violence or similar behaviour. This intent is not subjective, internal, or psychological. Rather, it must be understood as the actual capacity to generate a real risk or cause harm. In other words, it is an objective requirement. In other words, there must be a "clear intent to commit a crime and the actual, real, and effective possibility of achieving its objectives"[5].

In the case of incitement to violence online, the actual capacity to cause harm or pose a real risk can be determined based, first, on the level of organization of the actors engaging in hate speech. For example, whether it is an organized group through a Facebook group, or an Instagram or WhatsApp channel; or

---

[3] Harvard Law Review. (2023). Garnier v. O'Connor-Ratcliff: Ninth Circuit finds First Amendment violation in school district officials' blocking of parents on social media. Cambridge, MA: Harvard Law Review Association, p. 1485.

[4] Meta, Transparency Center, Community Standards. More information in: https://transparency.meta.com/en-us/policies/community-standards/violence-incitement/
https://www.oas.org/es/cidh/informes/pdfs/violenciapersonaslgbti.pdf

[5] Comisión Interamericana de Derechos Humanos. "Violencia contra Personas Lesbianas, Gay, Bisexuales, Trans e Intersex en América", párr. 235. Fecha: 12 de noviembre de 2015.
https://www.oas.org/es/cidh/informes/pdfs/violenciapersonaslgbti.pdf

whether it is an ordinary person through their personal account; or, conversely, an influencer. Secondly, the actual capacity can be identified through the magnitude and impact of the message, video, audio or image. How many people does the channel have? How many people reacted to the photo? How many times was the content shared? Did people watch the entire video or just a few seconds? What was the main reaction that motivated that content?

In second place, incitement to violence on the Internet follows a pattern, targeting specific groups: vulnerable groups, or individuals belonging to these groups, who are subjected to violence because of their relationship with that community. Specifically, these are historically discriminated groups, such as people of African descent, women, migrants, the LGTBIQ+ community, indigenous communities, people with disabilities, human rights defenders, among others. It is based on this reasoning that the American Convention on Human Rights states, in Article 13, paragraph 5, that 'any advocacy of national, racial or religious hatred that constitutes incitement to violence [...] against any person or group of persons shall be prohibited.'

Thus, if online behaviour meets both requirements, it would be considered violent or inciting violence. In other words, it would be behaviour not protected by freedom of expression.

On the other hand, it is important to define the **relevance of public discourse and scrutiny of public officials in such cases, as well as whether this extends to their family members, including minors**. In this regard, it is worth recalling the jurisprudence of high human rights courts, such as the Inter-American Court of Human Rights and the European Court of Human Rights. Noteworthy is the 2011 case of Fontevecchia and D'Amico v. Argentina[6], in which the Inter-American Court emphasized that public figures could not have the same expectation of privacy regarding matters relevant to public affairs, especially indications of corruption or abuse of power. In that case, after applying a three-part test, it was determined that the dissemination of information and photographs involving the official's son was a matter of public interest, since there were indications of irregularities in the exercise of power, such as illicit enrichment, and therefore fell within the scope of freedom of expression. Also noteworthy is the case of KCouderc and Hachette Filipacchi Assoc. v. France, heard before the ECHR in 2015[7]. This case also shows that the father's status as a public figure may imply a certain degree of public scrutiny of events relating to his children, provided that the connection involves a matter of public interest, such as transparency in public administration or the use of public resources. In that case, it concerned the publication of news stories with photos of a child who was allegedly the head of state's illegitimate son.

On the other hand, with regard to the **standards for protecting minors against online hate speech and incitement to violence**, International Human Rights Law (IHRL) has developed a specific legal framework to protect minors against online hate speech. As Reyes, Miranda, Ruiz and Pulido, point out, "the normative body of IHRL has been expanding over time due to various factors[8]," incorporating treaties that recognize the needs of vulnerable groups. Children and adolescents constitute one of these "historically discriminated groups whose special situation of vulnerability has led to the adoption of international treaties aimed at ensuring they exercise their rights and freedoms effectively"[9].

---

[6] *Caso Fontevecchia y D'Amico vs. Argentina* (Corte IDH, Serie C No. 238). Sentencia de 29 de noviembre de 2011. Párr. 71-72. https://corteidh.or.cr/docs/casos/articulos/seriec_238_esp.pdf#:~:text=p%C3%BAblico%20y%20que%2C%20adem%C3%A1s%2C%20era,acto%20que%20no%20es%20una

[7] Tribunal Europeo de Derechos Humanos (TEDH), *Couderc and Hachette Filipacchi Associés v. France*, no. 40454/07, Grand Chamber, Judgment of 10 November 2015. https://hudoc.echr.coe.int/fre#{%22itemid%22:[%22001-158861%22]}

[8] Reyes, V., Miranda Cerna, P. G., Pulido Ramírez, D., & Ruiz, Y. (2023). *Nuevas tecnologías y derechos humanos: Impactos, desafíos y oportunidades en la era de la conectividad digital* [Informe de investigación]. IDEHPUCP & Fundación Konrad Adenauer.

[9] Ibid.

From this foundation, the universal system has established concrete protection standards through the Committee on the Rights of the Child, which recognizes the need to balance digital security and privacy. The Committee establishes that States must implement "an approach that integrates both security and privacy from the design phase in relation to anonymity,"[10] warning that these practices should not "be systematically used to hide harmful or illegal behaviors, such as cyber-aggression, hate speech or sexual exploitation and abuse"[11].

In parallel, the Inter-American system complements these universal standards through specific pronouncements on regional realities. REDESCA and RELE of the IACHR have condemned manifestations of online violence against minors, including death threats, harassment, defamation campaigns, and messages that falsely link them to illegal armed groups. These rapporteurs recognize that such speeches violate fundamental rights and generate intimidating effects that silence youth participation in public debates[12].

Together, these protection standards from the universal and Inter-American systems configure a specific corpus that balances the protection of minors with the guarantee of safe digital spaces. Both the privacy and security by design approach of the Committee on the Rights of the Child, and the Inter-American system's condemnation of threats, harassment and defamation campaigns against children and adolescents, establish concrete state obligations to prevent hate speech without compromising fundamental rights such as privacy and democratic participation of minors in digital environments.

Therefore, in the case under analysis, three issues should be taken into consideration: (i) the definition of online violence or incitement to violence, (ii) standards on the protection of children's rights, and (iii) the extension of public scrutiny to family members, such as children, of public officials when a matter of public interest is involved, such as public management or the administration of public resources.

III. Procedures and response mechanisms for the removal of content of public interest

We recommend adopting a **graduated protocol that applies proportional and verifiable measures** to address requests for the removal of public interest content involving or referencing children and young people (CYP). The protocol components are outlined below.

A. Detection and Trigger Signals

A hybrid system of automated and external signals is advised, ensuring reliance is not placed exclusively on algorithms. Automated signals may include:

- Detection of images depicting CYP
- Sudden spikes in engagement and other indicators of virality
- Identification of terms that may constitute “rhetorical threats”[13]

[10] Comité de los Derechos del Niño. (2021, 2 de marzo). *Observación general núm. 25 (2021) relativa a los derechos de los niños en relación con el entorno digital* (CRC/C/GC/25). Naciones Unidas.
[11] Ibid.
[12] Relatoría Especial sobre Derechos Económicos, Sociales, Culturales y Ambientales & Relatoría Especial para la Libertad de Expresión. (2025, 16 de septiembre). *REDESCA y RELE condenan la violencia en línea contra niñas, niños, adolescentes y jóvenes que defienden el ambiente y el clima* [Comunicado de prensa]. Comisión Interamericana de Derechos Humanos.
[13] We recommend the use of multilingual datasets enriched with cultural annotations, while avoiding literal translations, to reduce both false positives and false negatives.

These should be complemented by external signals such as notifications from authorities, affected individuals, or mass user reports. A diverse set of triggers supports early identification and prioritization of posts requiring urgent action.

B. Temporary Measures

When content is identified as potentially harmful to CYP, immediate temporary safeguards are recommended rather than outright removal pending review. Suggested measures include:

- Restricting algorithmic amplification, including recommendations and trending placement
- Automatically blurring CYP faces through computer vision to protect their image, privacy and intimacy. In case of error, review mechanisms must be available.

C. Human Review

Within 48 hours, content should undergo human review by a dual team: one moderator and one regional language or cultural expert, to mitigate bias and ensure contextual understanding. For example, a Tagalog phrase could be clarified as either a call for accountability or a veiled threat of violence. Automated systems alone are insufficient due to limitations in accurately interpreting political, cultural, and linguistic nuances.

The review process may follow a checklist to assess:

- Whether the post constitutes public interest (e.g., involving a public official in office)
- Whether its tone is informative or incites action against the official
- Whether it poses risks to CYP, such as direct identification, calls for action against them, or exposure of personal data that could enable doxxing or offline harm

If risks to CYP are confirmed, measures should include editing the author's content to remove harmful elements. Where editing is not feasible, removal should be reserved for cases involving explicit threats, exposure to danger, or dissemination of CYP's personal data.

Conclusion

Given the growing prevalence of digital threats to CYP, and the need to preserve spaces free from undue censorship or restrictions on freedom of expression, this protocol is designed to ensure proportionate responses where CYP may be at risk. We recommend META adopt or update its response mechanisms in line with these guidelines.

SS.

RENZO DÍAZ GIUNTA
ROMMEL INFANTE ASTO
RUBIELA GASPAR CLAVO
XIMENA CUZCANO CHÁVEZ
LUIS PEBE MUÑOZ
ANDREA APOLÍN VARGAS