CELE
Centro de Estudios en Libertad de Expresión

Documento de posición

29

CELE submission to the Office of the Special Rapporteur for Freedom of Opinion and Expression

Agustina Del Campo,
Vladimir Garay y
Nicolás Zara

21 de julio de 2025

Agustina Del Campo, Vladimir Garay, Nicolás Zara, "CELE submission to the Office of the Special Rapporteur for Freedom of Opinion and Expression", Documento de posición No. 29 (ENG), Centro de Estudios en Libertad de Expresión (CELE), Buenos Aires (2025)

Facultad de Derecho

Centro de Estudios en Libertad de Expresión

UP
Universidad de Palermo

CELE submission to the Office of the Special Rapporteur for Freedom of Opinion and Expression[1]

Buenos Aires, Argentina

Santiago, Chile

July 21, 2025

The Center for Studies on Freedom of Expression and Access to Information (CELE) is an academic research center aliated with Universidad de Palermo in Argentina. The Center provides technical and legal analysis on issues aecting this fundamental right, and since 2012 has been studying freedom of expression on the Internet as a specic research area. The Center is a leading voice on the promotion and protection of freedom of expression nationally, regionally, and internationally.

This document was prepared in response to an informal consultation conducted by the Oce of the Special Rapporteur for freedom of opinion and Expression at the United Nations, Dr. Irene Khan, to provide inputs for her upcoming report on "threats to freedom of expression online, emerging from content moderation and content curation policies and practices of digital platforms, including in particular recent backsliding on human rights commitments by major companies and States at a time when the international human rights system is already under immense pressure." The report is meant to build on "the latest thematic report to the Human Rights Council on Elections in the digital age A/HRC/59/50 and latest General Assembly report on Global threats to freedom of expression arising from the conict in Gaza Document Viewer)".

[1] This submission was prepared by Agustina Del Campo (Director), Vladimir Garay (Senior Coordinator) and Nicolás Zara (Researcher)

Content moderation online: legal frameworks and Human Rights obligations

1. In recent years, there has been growing interest in the content moderation and curation policies and practices of digital platforms, particularly concerning their potential impact on the exercise of human rights in the digital environment, especially freedom of expression online.

   There are concerns about the role of digital platforms in shaping public debate — especially the handful of tech companies often labeled as 'big tech', most of which are based in the United States — and their ability to mold online speech through their internal policies, Terms of Service (ToS), and decisions. There has been a push, including from this oce, for companies to align their practices with international human rights standards per the UN Guiding Principles on Business and Human Rights (UNGPs).

2. The power of online platforms regarding public debate largely stems from a normative model that encourages self-regulation, supported by legal protections —such as Section 230 of the U.S. Communications Decency Act— that shields companies from liability for most user-generated content, provided that they act in good faith when moderating content and comply with applicable laws. ("Good Samaritan" protection). This model granted companies' broad freedom to set their own rules, enabling them to host vast amounts of user-generated content without needing to pre-approve each piece. It also allowed companies to independently decide how to limit, curate, and moderate content without fear of being sued for that which they decided to host, allow, and not remove.

3. "Good Samaritan" protections, while allowing the broadest business and speech freedoms, also turned companies into targets for advocacy, coming from both the public

and the private sector. Digital platforms are prot-driven companies. And their drive to maximize prot and shareholder value makes them susceptible to various forms of pressure from States. This susceptibility is geographically uneven and proportional to each company's interests in a particular region.

4. States actors and civil society alike —though not necessarily as eectively— have pressured companies to "voluntarily" shape the content they host, their Terms of Service (ToS), and their moderation practices in specic ways, compelling internet intermediaries to act beyond their legal obligations. In the last few years, companies became a new forum where traditionally interested parties —NGO's, political parties, State representatives, etc.— argued and advocated for dierent interpretations of State-led law, renegotiation of freedom of expression's scope and limits, and expansion of censorship, even beyond what States would be legally able or allowed to enforce.

5. This phenomenon is not new. In the past, companies had already adjusted their policies in response to mounting public pressure or specic government demands. After the 2016 elections and the crisis of trust triggered by the Cambridge Analytica scandal, Meta —then Facebook —, for example, revised its privacy settings and terms of service.[2] During the COVID-19 crisis, they also yielded to pressure from the U.S. Executive Branch to curb content seen as undermining their public health policies.[3] The fact that even powerful companies remain vulnerable to government pressure is troubling. This oers

[2] Mark Zuckerberg, March 21, 2018, https://www.facebook.com/zuck/posts/10104712037900071. Erin Egan and Ashlie Beringer, It's Time to Make Our Privacy Tools Easier to Find, March 28, 2018, https://about.fb.com/news/2018/03/privacy-shortcuts/

[3] Mark Sweney, Mark Zuckerberg says White House 'pressured' Facebook to censor Covid-19 content, The Guardian, August 27, 2024, https://www.theguardian.com/technology/article/2024/aug/27/mark-zuckerberg-says-white-house-pressured-facebook-to-censor-covid-19-content

an interesting perspective on the relative power dynamics between private companies and states.

6. Internet platforms' content moderation practices are never neutral but result from a complex mix of competing interests and tensions, including their unwillingness to compromise operations or prot margins, and a willingness to go to certain lengths to remain in good standing with certain governments. This applies to the broader content moderation practices adopted by every internet platform over time. The shift within Twitter around 2018, for example, when they turned "speak truth to power" to "promoting a healthy conversation" was also the result of increasing pressure and tensions. In this sense, every content governance decision — past, present, and future — must be seen as a reection of a dynamic— and, sometimes, tense — interplay of corporate, political, and social pressures rather than purely principled or technical reasoning. In some cases, this has led to content decisions that appear aligned with dominant geopolitical narratives, especially when platforms operate in politically or economically strategic markets. In others, it has triggered pushback.

State action and platform self-regulation

7. It is paramount to distinguish the private, company-made rules that govern interactions within a platform from the rules and actions of the States, for private companies and States have dierent responsibilities. While the former are legally entitled to change their rules and editorial positions, the latter are not allowed to put pressure on them to do so.

8. States can and must pass legislation to govern the online ecosystem under their responsibility to protect and guarantee the enjoyment of human rights. But they must do

so following strict human rights standards, which include freedom of expression and compliance with the three-part test: legality, necessity, and proportionality. Limitations must be established by law. Negotiations with platforms or undue pressures behind closed doors to make them change the way they handle third-party-posted content, therefore, violate International Human Rights Law. So could state-led use of content moderation mechanisms to monitor and silence certain voices or points of view, including through public ocials' use of private agging mechanisms. Furthermore, state-mandated "trusted aggers" mechanisms -like the ones mandated through Article 16 of the European Digital Services Act- ought to exclude state actors or dependents from applying and acquiring such status. These mechanisms must not be used by the State to "*de facto* sidestep legal requirements to request content removal" or to intervene in content moderation without a legal basis, by agging potential breaches of the platform's terms of service[4].

9. Platform policies that are more eective in tackling disinformation or curbing oensive expression may come at the expense of curtailing legally protected —albeit rude or distasteful— speech. It is important not to conate what private platforms can do under their own "hateful speech policies" — which usually covers both protected and unprotected expression under International Human Rights standards— with International Human Rights-based hate speech rules. Narrowing the scope of the former does not necessarily entail allowing the kind of expression referred to in Article 20 of the ICCPR.

[4]Jacob van de Kerkhof, "Article 22 Digital Services Act: Building trust with trusted aggers," Internet Policy Review: Journal on Internet Regulation, Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin, vol. 14(1), 2025, p. 8-9, retrieved from https://www.econstor.eu/bitstream/10419/315582/1/1923068466.pdf

State-led codes of conduct (such as those on Hate Speech and Disinformation under Article 45 of the DSA) deserve closer scrutiny, as they may allow regulators to impose restrictions on freedom of expression while circumventing human rights standards. These codes contain obligations that go way beyond what is required by International Human Rights Law, and therefore include actions that the states could not legitimately ask platforms to take[5]. While they are presented as legitimate eorts against dangerous expressions since platforms undertake such compromises formally voluntarily, their voluntary nature is disputed: from the outset, authorities pressured platforms to sign them under the threat of stricter regulations and set the course their content should follow[6]. Moreover, there are both positive and negative incentives that make it too costly –almost inescapable– for platforms to stay out of them[7].

10. Ultimately, the International and regional systems for the protection of human rights make States, not companies, internationally responsible for failing to protect the rights of their citizens. However, States could be found in breach of these obligations should they fail to take appropriate steps to prevent, investigate, punish, and redress abuse by private parties to persons under their jurisdiction[8].

[5] Natalie Alkiviadou, Why we should be concerned about the "Illegal hata speech" code of conduct +, blogpost at The Bedrock Principle, February 5, 2025, https://www.bedrockprinciple.com/p/why-we-should-be-concerned-about

[6] Rachel Grin and Carl Vander Maelen, "Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges" (May 30, 2023). Available at SSRN: https://ssrn.com/abstract=4463874 or http://dx.doi.org/10.2139/ssrn.4463874, also see Rachel Grin and Carl Vander Maelen, Twitter's retreat from the Code of Practice on Disinformation raises a crucial question: are DSA codes of conduct really voluntary?, blogpost at the DSA Observatory Blog, June 12, 2023, https://dsa-observatory.eu/2023/06/12/twitters-retreat-from-the-code-of-practice-on-disinformation-raises-a-crucial-question-are-dsa-codes-of-conduct-really-voluntary/

[7] Íbid.

[8] United Nations, Commentary to Principle 1, in Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, 2011

State and government influence on companies' practices

11. Recent changes in Meta's content moderation policy, just a few days after the Trump administration took oce, have largely been interpreted as a strategic move to curry favor with the U.S. government. Most of Silicon Valley's big tech companies have made similar gestures. In return, the new administration demonstrated a much more protectionist approach and a willingness to defend the interests of this industry, for example, adopting retaliations against countries that sought to introduce tougher legal obligations for internet platforms, such as Brazil and the European Union, or implementing greater restrictions to foreign companies operating in their territory, such as TikTok. These kinds of changes of heart by internet intermediaries have raised alarms across the globe, and dierent groups have highlighted dierent reasons. Our point, however, is mostly that regardless of the substantive changes, the timing and alignment of those changes to t the new government's narrative is by far the most compelling evidence of editorial self-censorship.

12. The U.S. is, however, not the only country or government that inuences platforms. The latest General Assembly report on Global threats to freedom of expression arising from the conict in Gaza[9] oers some recent examples of these tensions as well. On one hand, the report characterizes platforms' content moderation practices as "[o]verly restrictive, unbalanced", "opaque and inconsistent" against Palestinian speech, that "seem to be the result of inherently biased policies (...), heavy reliance on automated tools to moderate and translate and a permissive approach to State requests for content removal".[10] According to the report, "the large platforms have tended to be more lenient regarding

[9] United Nations. *Global Principles for Information Integrity: Recommendations for Multi-Stakeholder Action*. A/79/319. New York: United Nations, 2024. https://docs.un.org/en/A/79/319.
[10] Ibid., par. 60

Israel",[11]leading to the "disproportionate censoring of Palestinian content".[12] Furthermore, the report highlights numerous ways in which certain States have suppressed "protests, advocacy and critical expression in a disproportionate and discriminatory manner against Palestinian groups",[13]including restrictions to demonstrations in support of Palestinian people, arbitrary detentions, excessive use of force by the police against demonstrators, and other repressive measures in Germany, France, Belgium, Canada, Greece, Italy, the Kingdom of the Netherlands, and the United States.[14] There have also been reports of assaults against academic[15]and intolerance of artistic[16]freedom in Western Europe and North America.

Even more disturbing is the fact that the Government of Israel invested millions of dollars in online advertisement dehumanizing Palestinians, published in online platforms (YouTube, Meta) "even when they appear to violate the terms of service of platforms".[17] Israeli government ocials have used online platforms to publish Statements against Palestinians in dehumanizing terms. As Statesd in the report, "[s]uch language may amount to incitement to genocide".[18]

13. Furthermore, other States have criminalized the display of Palestinian symbols as signs of antisemitism and support for Hamas, including the Palestinian national ag, the *keffiyeh* (traditional black and white scarf),[19]and the phrase "From the River to the Sea, Palestine

[11]Ibid., par. 51
[12] Ibid., par. 53
[13] Ibid., par. 32
[14]Ibid., par. 34-36
[15]Ibid., par 38-45
[16]Ibid., par 46-48
[17]Ibid., par 62
[18]Ibid., par. 62
[19]Ibid., par. 71

Will Be Free".[20] These blanket bans have been implemented based on counter-terrorism laws that often include vague and poorly dened oenses that fail to meet the international requirement of legality.[21] Coincidentally, the report cites a study from December 2023 regarding English-language takedowns by Meta, documenting the censorship and suppression of legal content in support of Palestinians: "Phrases such as ′From the River to the Sea, Palestine will be free', ′Ceasere Now′, and ′Stop the Genocide′ have been repeatedly removed by Meta's platforms as ′spam′. The ′emoji′ of the Palestinian ag has been hidden as ′potentially oensive′, and phrases such as ′Palestinian′ and ′praise be to God′ have been translated as ′Palestinian terrorists′.[22]

14. If online platforms act in ways that enable censorship, discrimination and genocide, this cannot be explained solely by internal policy failures. States have a duty to respect, protect, and guarantee human rights on and oine. States must abstain from pressuring companies to censor legal speech. They must act within the rule of law and may not unduly inuence or otherwise use either informal application of ToS or formal pressure to excessively restrict freedom of expression, especially without due process or transparency.

Companies' duties under the UNGPs

15. Transnational companies may sometimes be as impactful as States in the lives of communities and individuals and could aect "virtually all internationally recognized

[20] Ibid., par. 73
[21] Ibid., par .70
[22] Ibid., par. 54

rights".[23] This may include internet intermediaries.[24] This is what motivated the United Nations to issue the United Nations Guiding Principles for Business and Human Rights in 2011. Under these Guidelines, companies have a duty to identify risks to human rights, mitigate their potential impact, and remediate where possible.

16. Inspired by the human rights due diligence obligations arising from the UNGPs and concerned by their lack of legal force, lawmakers around the world started to issue domestic regulations that kept the immunity rule for third-party-posted content described above for internet intermediaries, but placed some due diligence obligations on top of that. While innovative and well-intentioned, some of these laws conate duties to identify and mitigate human rights risks with larger, overall social risks, dangerously broadening the language of the UNGPs and risking the existing consensus and agreements on human rights law worldwide.

17. The European Digital Services Act (DSA) is a pioneer law among those, and one that many bills are being modeled after elsewhere.[25]It has a particular approach to content moderation that requires very large online platforms and search engines to identify, assess and mitigate a series of systemic risks stemming from the design or use of their service. However, this is dierent from the UNGPs insofar as it requires platforms to curb *systemic risks* such as the spread of illegal content, gender-based violence, negative eects on fundamental rights, negative eects on the well-being of minors, elections and civic discourse, public health, etc.

[23]J. Ruggie, «Protect, Respect and Remedy: a Framework for Business and Human Rights. Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises». Human Rights Council, Geneve. A/HRC/8/5. April 7, 2008

[24] Alex Warofka, An Independent Assessment of the Human Rights Impact of Facebook in Myanmar - About Facebook, About Facebook, November 6, 2018, https://about.fb.com/news/2018/11/myanmar-hria/ [25] Matías González Mama and Ramiro Álvarez Ugarte, "Modelización regulatoria: palpitando la inuencia de la digital services act en América Latina," Latin American Journal of European Studies 5, no. 1 (2025): 400 et seq.

18. The scope of the obligations arising from the UNGPs are narrower, and limited only to Human Rights aectations; the DSA, on the contrary, covers a much broader range, and under its framework, certain expressions that are perfectly legal could be considered problematic and trigger "mitigation" obligations for the companies. Second, the DSA is specic to an industry that deals with processing and handling human expression and articulates a new paradigm: that online legal speech generates risks that need to be managed and mitigated. Finally, the DSA does not consider States as a "risk-generating" factor under this paradigm.[26]

19. Laws containing this kind of mandate to mitigate certain "lawful but awful" expressions under general "due diligence" obligations, often including vague and overbroad categories of expression, are prone to abuse. Authoritarian governments or ocials could easily make use of them to further erode the institutional safeguards and silence dissidents in their jurisdictions.[27]

20. Where those laws are already in the books, it is important that any vague key terms are dened; that open-ended provisions are interpreted through the prism of international human rights law to avoid overbreadth; that any such due diligence obligations are applied under the light of the Business and Human Rights framework; and that the role of the States as a potential source of risks to freedom of expression is acknowledged.

[26] Del Campo, Agustina and Zara, Nicolas and Álvarez-Ugarte, Ramiro, Are Risks the New Rights? The Perils of Risk-based Approaches to Speech Regulation (February 28, 2025). CELE Research Paper No. 64. Forthcoming in the Journal of Intellectual Property, Information Technology and Electronic Commerce, vol. 16, No. 2 (2025), CELE Research Paper No. 64, Available at SSRN: https://ssrn.com/abstract=5161173 or http://dx.doi.org/10.2139/ssrn.5161173, p. 11-12

[27] David Kaye. 2024. The Risks of Internet Regulation. Foreign Aairs. Retrieved March 25, 2025 from https://www.foreignaairs.com/united-States/risks-internet-regulation, Anupam Chander, When the DSA goes global, Berkeley Technol. Law Journ. 38, 3 (January 2024), p. 1067-1088, Rachel Grin, EU Platform Regulation in the Age of Neo-Illiberalism (March 29, 2024). Available at SSRN: https://ssrn.com/abstract=4777875 or http://dx.doi.org/10.2139/ssrn.4777875

Information Integrity and Content Moderation

21. The underlying concerns that gave rise to the DSA can also be found in recent United Nations guidelines on content moderation and the governance of internet platforms, most prominently through the Information Integrity agenda.

22. "Information integrity" is a loosely dened term adopted in recent years by the United Nations and its aliated institutions to address the impacts of digital technologies on the communication ecosystem. While recognizing that the mass adoption of digital technologies has created opportunities for "connecting individuals and communities on a previously unthinkable scale and presenting unparalleled opportunities for the diusion of knowledge, cultural enrichment and sustainable development",[28]the Information Integrity agenda is concerned with how that same technology has given an outlet for dierent forms of speech that it considers disquieting although not necessarily illegal.

23. Ultimately, the Information Integrity agenda appears to ask what it would take to preserve all the positive outcomes of the so-called digital revolution while avoiding its "negative" consequences. With that goal, signicant eort has been devoted to drafting recommendations and guidelines to achieve a "perfect" model for content moderation. But this may prove to be an impossible task. Not only is every content moderation policy the product of a complex interplay of interests within specic contexts, but the idea of limiting content not based on its legality, but on its perceived "problematicity" risks triggering overreach and censorship, especially given that achieving a nal consensus on what constitutes "harmful" or "acceptable" content may be unattainable. The very

[28] United Nations. *Global Principles for Information Integrity*. New York: United Nations, 2024. https://www.un.org/sites/un2.un.org/les/un-global-principles-for-information-integrity-en.pdf , p.2

denition of terms like “harmful,” “acceptable,” “fringe,” “risky,” and “desirable” are not neutral; they are contextual, contested, and deeply political.

24. Moreover, the impulse to restrict problematic-but-legal content online may stem from a misunderstanding of the nature of the issues and the role of online platforms. In the eld of political philosophy, many authors have argued that dissent is a critical component of democracy, including John Stuart Mill, Chantal Moue, and Jacques Rancière. From dierent perspectives, they all assign value to unusual, fringe, and even problematic ideas within the context of public debate. It is precisely through conicting oppositions that democracy can give rise to momentary forms of deliberation and consensus. In her critique of Habermas's concept of the “public sphere,” Nancy Fraser argues that, rather than a single, overarching, rational, egalitarian, and culturally neutral space, there exists a multiplicity of competing “counterpublics”, each with uneven levels of power and inuence. “The public sphere was always constituted by conict”.[29] Fraser concedes that some of these counterpublics can be “explicitly anti-democratic and anti-egalitarian”.[30]

25. Following this line of thinking, the emergence of fringe, problematic, or even risky speech is not a problem caused by the mass adoption of social media or the result of awed content moderation practices, but rather a feature of democratic dynamics itself — not a bug. The main transformation introduced by digital platforms is a shift in the visibility regime of dierent counterpublics. These new dynamics not only allow more people to encounter—and potentially adopt—ideas that would have been harder to access oine, but also increases awareness within hegemonic spheres of speech they oppose or reject,

[29] Nancy Fraser, “Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy,” *Social Text*, no. 25/26 (1990): 56–80, https://www.jstor.org/stable/466240. p.61
[30] Ibid., p. 67

leading to dierent proposals on how to limit and combat the ones they nd more alarming.

The emergence of problematic speech is neither a new phenomenon nor merely a technical issue caused by digital platforms and awed moderation practices. And it cannot be eliminated without compromising democracy itself, even by shutting down all online media platforms. You can only hide it better.

26. An alternative way forward — a more democratic approach— is to embrace conict and dissent as essential components of democratic deliberation. This involves promoting diverse content moderation practices that respect international human rights standards—neither falling short of, nor exceeding their mandate—and limiting the ability of states to exert pressure in either direction.

Recommendations

27. The immunity of digital platforms for third-party-posted content is a powerful guarantee of freedom of expression in the digital age and a free internet. States should refrain from passing laws that discard or carve out those protections.

28. The ability of platforms to self-govern is an important tool for business and speech freedoms. However, ToS, community guidelines or equivalents should not be equated to law. Self-regulation regimes should always develop within the framework provided by human rights law.

29. When self-regulation regimes are legally promoted, adopted, and in place, States should carefully craft policies and design mechanisms to separate self-governing structures from State-led action. These should include at least strong transparency mandates for public

ocials' communications with companies, uses of ToS to ag content online for private action, clearly distinguish public accounts from public accounts (whenever that distinction may be made, i.e. the account of a municipality or state oce from the account of a public servant working in that jurisdiction).

30. State entities and/or actors should not be set up to be or be accepted by companies to be legitimate or trusted aggers.

31. Companies do not operate in a vacuum. As prot-driven entities, they must comply with government requirements to continue operating. Clear distinctions must be drawn between companies' obligations under the UNGPs to identify risks to human rights and mitigate those and other obligations that exceed the scope and reach of companies and entail State duties. Companies may have a duty under the UNGPs to ght against state legislation or policies that violate human rights. The Global Network Initiative is a good example of best practices in the eld of driving corporate accountability for their actions and reactions against state conduct.

32. States have the duty to protect the human rights of people within their territories. In accordance with it, they must adopt concrete measures to guarantee their full enjoyment online. Any such measure must be implemented by law, pursue a legitimate interest, and be necessary and proportionate. Laws designed to protect a certain interest could have an undesired or collateral impact on human rights –for instance, the adoption of rules obliging platforms to reduce the exposure of users to certain contents could have a detrimental eect on the access of users to legitimate sources of knowledge–.

33. When states adopt laws that impose due diligence measures on internet service providers, they must ensure that they do not require, encourage or incentivize the removal,

deranking, demonetization, etc., of legal content, even when it is unpleasant or

disturbing. Laws with the potential to aect the ToS or content moderation practices of providers (even collaterally) must not impose on them any criteria for content treatment other than legality.

34. When such laws are adopted, they ought not to include vague terminology, which gives both platforms and state actors in charge of enforcement too much discretion in their interpretation, which conicts with the principle of legality under Article 19 of the ICCPR. Enforcement agencies and actors must set clear guidelines regarding what they expect from the regulated entities and how any open-ended term will be interpreted for the purpose of their enforcement. Any vague provisions must be interpreted in such a way that makes them compatible with international human rights law standards.

35. Any due diligence obligations imposed on internet intermediaries must be interpreted in light of the Business and Human Rights framework and must not encourage or nudge them to act in any way upon legally protected expression.

36. While the States have a role to play in protecting and guaranteeing human rights in the digital environment, it must be acknowledged that they could also pose a potential risk to freedom of expression. The laws designed to protect the rights of internet users must take that into account and contemplate democratic oversight mechanisms of the state role.