CELE
Centro de Estudios en Libertad de Expresión

Documento de posición

31

CELE submission to the Committee of Ministers' public consultation on the Draft Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression

Vladimir Garay y
Matías González Mama

12 de agosto de 2025

Vladimir Garay. Matías González Mama, "CELE submission to the Committee of Ministers' public consultation on the Draft Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression", Documento de posición No. 31 (ENG), Centro de Estudios en Libertad de Expresión (CELE), Buenos Aires (2025)

UP
Universidad de Palermo

Facultad de Derecho

Centro de Estudios en Libertad de Expresión

CELE submission to the Committee of Ministers' public consultation on the Draft Guidance Note on the Implications of Generative Artificial Intelligence for Freedom of Expression

Vladimir Garay

Matías González Mama

CELE/UP

cele@palermo.edu

Buenos Aires, Argentina

August 12, 2025

The Center for Studies on Freedom of Expression and Access to Information (CELE) is an academic research center affiliated with Universidad de Palermo in Argentina. The Center provides technical and legal analysis on issues affecting this fundamental right, and since 2012 has been studying freedom of expression on the Internet as a specific 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 the public written consultation on the draft text of the Guidance Note by the Committee of Experts on the Implications of Generative Artificial Intelligence For Freedom Of Expression (MSI-AI). We value the effort made by the MSI-AI and are grateful for the opportunity to provide feedback on their proposal.

The Guidance Note addresses a crucial topic in the context of rapid technological evolution. However, to fully safeguard freedom of expression and human rights, it is essential that it incorporates and addresses certain considerations presented hereinafter.

Any restriction on freedom of expression, even when mediated by AI, must pass a rigorous analysis of legality, legitimate aim, necessity, and proportionality. The Guidance Note should explicitly emphasize that the regulation of content generated or mediated by AI, if considered a state-mandated restriction (directly or indirectly), must meet this standard. Vague definitions do not comply with the principle of legality, which can lead to excessive discretion and over-removal of content.

Pluralism and Diversity in AI

One of the main concerns expressed throughout the document is that the widespread adoption of Artificial Intelligence technologies could pose a threat to pluralism and undermine diversity. This could lead to the standardization of human expression[1] by homogenizing both the content and the way in which ideas are structured and expressed, and the risk of overshadowing unique voices by repetitive or statistically standardized content;[2] and a lack of diverse "opinions, perspectives, and sources" in the media "that reflect the plurality of Society",[3] eroding "the right to access to pluralistic information from a variety of sources"[4] which is vital for a robust public debate.

As stated in the draft document, there are several reasons for this, such as lack of representativity of datasets, including limited linguistic and cultural diversity,[5] and partial or biased data sets,[6] and varying levels of concentration in the AI industry,[7] which could end up arbitrarily limiting the possible answers for a certain prompt.

The document offers a couple of solutions to this problem, for example, investing in data strategies fostering the development of accessible, diverse, and representative public data sources,[8] and incentivizing measures to encourage the availability of more diverse Generative AI-powered product choice and viable technical alternatives.[9]

Both ideas could be complemented by the implementation of public service digital information infrastructures,[10] like Digital Public Goods (DPG), digital resources that are owned or provided for public use including "open-source software, open standards, open data, open AI systems, and open content collections that adhere to privacy and other applicable laws and best practices, do

[1] par. 42 b
[2] par. 48
[3] par. 66
[4] par. 23
[5] par. 50
[6] par. 66
[7] par. 42 f
[8] par. 91
[9] par. 94
[10] par. 89

no harm, and help attain the Sustainable Development Goals".[11] DPGs can support the availability of multiple AI systems, each trained on different datasets, built with different emphases and purposes, and guided by distinct content moderation practices, leading to a greater diversity of responses and solutions to the same prompts, encouraging a robust public debate. DPGs can also support the development of smaller, specialized AI models designed to better perform complex tasks. Adopting policies that promote DPGs can help mitigate some of the challenges associated with open-source software, such as long-term maintenance. It also offers a way to counter market concentration in the AI sector. [12]

Promotion of AI Literacy

The Guidance expressed concern about Generative AI systems producing content that is poor in quality, erroneous, factually incorrect, misleading, or biased.

Generative AI systems can generate "new text, images, audio, video or actions, or a combination of these, and transform content in various modalities and formats",[13] and "holds a significant potential to influence opinion and expression and feeds into public debate, knowledge dissemination, content creation and distribution".[14] But if the systems are trained on poor quality, unrepresentative, or biased data[15] they can reproduce and amplify existing inequalities, which may undermine freedom of expression.[16]

Among the potential harms identified in the Guidance is sycophancy, the tendency of AI systems to prioritize user approval by producing outputs that mirror the user's beliefs, fostering engagement over factual accuracy and pluralistic viewpoints.[17] The document states that AI systems can produce false answers or cite non-existent sources, a phenomenon commonly referred to as "hallucinations", which may pose a risk to the right to access reliable information and lead to the widespread dissemination of inaccurate content.[18] While the concern about sycophancy is valid in certain contexts, it may be more accurate to describe the phenomenon as a "mirror" or "bubble" effect, in which the AI reflects the user's preferences or assumptions rather than actively seeking to flatter or gain favor. Unlike a human actor engaged in sycophancy, who flatters strategically to obtain a benefit, particularly in asymmetrical power dynamics, the AI does not possess intent or motivation to please (at least that we know so far).[19] Instead, it responds

[11] https://www.digitalpublicgoods.net/digital-public-goods
[12] par. 71
[13] par. 6
[14] par. 9
[15] par. 20
[16] par. 7
[17] par. 23
[18] par. 51
[19] Ayres, Ian and Balkin, Jack M., The Law of AI is the Law of Risky Agents without Intentions (June 01, 2024). University of Chicago Law Review Online (forthcoming 2024), Yale Law & Economics Research Paper, Yale Law

based on patterns in the user's input and interaction history. The challenge, then, lies less in intentional flattery and more in the reinforcement of cognitive or informational biases, which can indeed compromise access to diverse, accurate, and pluralistic information. Yet these are not new risks, but rather extensions of dynamics that have long existed in human communication and media ecosystems. Biases have shaped the circulation of information throughout history, well before the emergence of AI.

The document also mentions that "Generative AI tools and applications behaving like echo-chambers hold the potential to impair the right to hold opinions and to access and receive accurate and plural information and ideas. Effective enjoyment of the right to freedom of expression involves access to pluralistic information from a variety of sources."[20]

While generative AI tools that behave like echo chambers may raise legitimate concerns, it is important to distinguish between the right to freedom of expression and the right to hold opinions, which are closely related but not identical. The document seems to conflate these two rights, which could obscure important legal and conceptual differences between them. The right to hold opinions is absolute under international human rights law and cannot be restricted, whereas freedom of expression can be subject to certain limitations. It is not self-evident that the use of generative AI systems impairs the right to form one's own opinion. Nor does it necessarily undermine access to pluralistic information, since AI is not the only means through which individuals inform themselves. People still have access to a broad range of sources, including the open internet, social media platforms, newspapers, television, radio, academic journals, public libraries, podcasts, community events, and interpersonal conversations. While it is crucial to remain attentive to the potential of AI systems to shape information environments, these systems are only one part of a much broader media and knowledge ecosystem.

AI systems are trained on massive datasets using statistical learning techniques to identify patterns, correlations, and structures in the data; they then generate content by determining what is most likely to appear based on the distribution of the data they were trained on. The new content is built from various unidentified sources, potentially including inaccurate ones, [21] and may also alter the original material or misattribute sources, [22] which can diminish people's ability to evaluate the reliability and trustworthiness of the outputs.

Although several technological refinements try to correct the inaccuracies of Generative AI,[23] this remains a difficult problem to solve. Particularly, the Guidelines say in note 32 that "Generative

---

School, Public Law Research Paper, Available at SSRN: https://ssrn.com/abstract=4862025 or http://dx.doi.org/10.2139/ssrn.4862025

[20] par. 23

[21] par. 52

[22] par. 42 c

[23] par. 51

AI-based systems generate possible next words and sentences mimicking human productions, as such, it can also be misinformation or disinformation." However, while concerns about the non-factuality of generative AI outputs are valid, particularly when systems produce inaccurate information or fabricated sources, it is important to recall that, under international freedom of expression standards, access to reliable information is a key element, but not the only one. The right to freedom of expression, as protected under Article 19 of both the ICCPR, encompasses not only the right to seek and receive information but also to impart ideas, including those that may be erroneous or controversial.[24] In this regard, equating AI-generated inaccuracies with "misinformation" or "disinformation" may be conceptually misleading, as these terms presuppose intent to mislead, an element that generative AI systems, lacking consciousness or intent, do not possess. While the risk to the public's right to access reliable information should not be underestimated, especially at scale, such outputs must be addressed with measures proportionate to the nature of the risk and with a clear understanding that not all inaccurate content constitutes disinformation under international law. The best way forward appears to be raising public awareness of this inherent characteristic of AI systems through AI literacy.[25] It is vital that users understand what AI tools can and cannot do, including their experimental nature, accuracy levels, and error rates. Recommendations should focus on creating accessible public resources and comprehensive education programs, both in schools and in the workplace, to improve understanding of GenAI's implications for freedom of expression.

Agency and Opinion Formation

The Guidance Note expresses concern about the impact of AI systems on users' ability to exercise their agency, as they may be influenced, coerced, misled, manipulated, or deceived.

The Guidance Note stated that, when interacting with users, AI Systems can prefer engagement over factuality, creating "content specifically tailored for that individual",[26] for example by "retaining information from past interactions, which reveals details about the users' identity and preferences, then used to influence future interactions or outputs".[27] According to the document, this could "influence personal opinions and beliefs and be misused to obtain large-scale automated opinion shifts or manipulation",[28] potentially leading users to adopt biases without realizing it.[29] The document express concern that this "subtle but pervasive influences" could threaten informed decision-making and undermine foundational principles of the freedom to hold and form opinions through pluralistic debate,"[30] leading to "fully automated forms of

[24] Calvet-Bademunt, Jordi, Freedom of Expression in Generative AI – A Snapshot of Content Policies (February 28, 2024). Available at SSRN: https://ssrn.com/abstract=5176290 or http://dx.doi.org/10.2139/ssrn.5176290
[25] par. 95 a
[26] par. 8
[27] par. 26
[28] par. 42 d
[29] par. 48
[30] par. 64

online radicalisation, coercion, and emotional attachment",[31] impairing the "cognitive autonomy of individuals and their right to form opinions and take independent decisions".[32]

The document also highlights the risk of losing cognitive abilities related to engaging meaningfully with information and forming opinions, and a weakening of critical thinking due to the frequent use of AI tools.[33] Another risk is the potential of techniques like deepfakes and voice cloning and other "practices of false or mimicked online identities" that can be used "for deceptive purposes" including to manipulate public debate, weakening "public trust and corrode the ecosystem of factual, reliable and diverse information".[34]

However, while concerning, there are still no conclusive studies that definitively demonstrate public debate manipulation by this type of content.[35] The Guidance Note should adopt a more measured approach, based on solid evidence and not on speculation, recognizing that discussions about AI are often polarized between "technosolutionism" and "panic".

AI technologies are described as "experimental" by their developers. The Guidance Note should reflect this uncertainty, avoiding presumptions about long-term impacts that have not yet been fully demonstrated or understood. This is essential for a technical, evidence-based public policy approach and for avoiding the pitfalls of risk-based frameworks, which tend to relegate rights to a secondary and marginal role.[36]

The Fundamental Role of Strong Public Institutions

The Guidance Note should explicitly include the possibility that State actions constitute a source of risk for human rights in the context of AI and platform governance. States must establish clear, human-rights-based legal frameworks that ensure transparency in the acquisition, implementation, and use of AI systems, with robust mechanisms for control, accountability, and independent oversight.

---

[31] par. 60
[32] par. 59
[33] par. 61
[34] par. 58
[35] Garay, Vladimir and Gonzalez Mama, Matias, ¿Prohibir lo falso?: deepfakes y libertad de expresión (June 14, 2025), available at: https://www.lasillavacia.com/red-de-expertos/red-de-democracia-y-tecnologia/prohibir-lo-falso-deepfakes-y-libertad-de-expresion/
[36] 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

The Guidance Note should urge States to be transparent about the acquisition, development, use, and oversight mechanisms of AI and GenAI. This includes establishing clear guidelines when AI is used to conduct state activities (i.e., from a chatbot to communicate with its citizens to surveillance mechanisms to ensure safety in public spaces), and that companies selling AI services to States conduct human rights impact assessments and be transparent about the use and impact of these services. Companies should disclose any state-mandated conduct with regard to their products, especially when freedom of expression and privacy may be affected by State action. By setting standards for AI, States could inadvertently promote ideological conformity under the guise of neutrality, for example, by removing references to "ideological" topics or requiring systems to reflect a notion of "truth" free from "social engineering." This could "chill" speech and distort the information ecosystem. The Guidance Note should include safeguards against this.

Mitigation measures

The Guidelines recommend that member States integrate freedom of expression considerations into human rights risk and impact assessments for Generative AI systems and applications, building on existing tools such as the Council of Europe's HUDERIA Methodology. These assessments should be systematic, iterative, and flexible, covering the entire Generative AI technology stack from end to end.[37] They should be co-developed by states, technology actors, and those directly affected, including through participatory due diligence in public procurement processes. A documented and auditable trail should be created, detailing the intended purpose, safeguards, design, and data choices, stakeholder engagement, and mitigation strategies.[38] In addition, clear, accessible, and meaningful information about how Generative AI systems operate, their implications for freedom of expression, and the safeguards in place to address risks should be made publicly available to citizens and civil society. This approach promotes a targeted, inclusive, and transparent framework for addressing the impact of Generative AI on freedom of expression.

However, it should be noted that mitigation measures themselves can also affect freedom of expression. Therefore, any such measures should be subject to the three-part test established under international human rights law: legality, necessity, and proportionality.[39] The current proposal does not clarify how, or by whom, the adequacy of these mitigation measures will be assessed in light of freedom of expression standards. Ensuring that mitigation efforts do not unintentionally infringe on human rights is essential to preserving the very freedoms these assessments aim to protect.

---

[37] par. 84
[38] par. 85
[39] Gonzalez Mama, Matias and Pérez Aledda, Agustín and Palacios Ramírez, Lina, Global AI Governance: Who Regulates, with what Approach, and for Whom? (June 01, 2025). CELE Research Paper No. 67, Available at SSRN: https://ssrn.com/abstract=5311517 or http://dx.doi.org/10.2139/ssrn.5311517

Copyright, diversity, and market concentration

Throughout the document, the Guidance Note expressed the need for "the copyright and intellectual property rights" of the content used in AI training to be "clearly established and respected".[40] The document establishes that Generative AI-based systems rely on high-quality, up-to-date content to produce accurate outputs that "often involve the unauthorised use of media content,"[41] which could represent a threat to the economic sustainability of the media and other creative industries. There is another similar concern regarding AI systems' «appropriation» of an author's style, "undermining and diluting the value and originality of their work".[42]

Regarding this, it is necessary to make some clarifications. First, styles are not copyrightable; no one can own a style and, consequently, no person or machine can appropriate one either.

The document states that AI systems generate content by synthesizing responses statistically, often blending multiple sources without explicit attribution.[43] This characterization is only partially accurate. Generative AI-based systems produce text by predicting the most probable sequence of tokens based on statistical patterns learned during training. This process does not involve patchworking content from specific sources, but rather generating new content that is statistically consistent with the patterns in the training data. While the outputs may reflect influences from multiple sources, they are not direct composites or pastiches of existing works. Potentially, the outputs could infringe the copyright of an existing work if they are proven to be derivative of it, but that needs to be assessed on a case-by-case basis. There are currently several legal cases worldwide concerning the copyright status of materials used in training AI systems, but no definitive ruling has yet been issued on whether this constitutes infringement or falls under a permitted exception.

It's important to understand that an over-protectionist approach, which expands copyright restrictions to AI training, would undermine the ability of new actors to build and establish AI systems, thereby consolidating the concentration of the AI market: if training an AI system becomes too onerous, only established actors will be able to afford it. This will affect the diversity and plurality of the information ecosystem and, ultimately, the exercise of freedom of expression.

Differentiation and Specificity in AI System Regulation

The Guidance Note focuses on Generative AI (GenAI) and its layers (foundation, tool, product). However, AI regulation must differentiate between different systems and their uses. For example, the Guidance Note should propose differentiated solutions for different types of generative AI

[40] par. 45
[41] par. 67
[42] par. 53
[43] par. 42 c

(i.e. AI for content moderation, a writing aid, and the use of GenAI by governments to conduct behavioural analysis or crises management are not the same and may probably require different regulatory approaches, oversight, risk identification, and human rights impact assessments).

The Guidance should also recommend prohibiting or strictly limiting the use of AI-based surveillance (such as facial recognition) in contexts that may unduly restrict privacy, freedom of expression, and assembly, especially in protests. Furthermore, AI systems should never have final decision-making authority where responsibility, ethical judgment, or legal liability are involved.