Algorithmic Gatekeepers:
The Human Rights Impacts of LLM Content Moderation

VII. Right to Participate and Remedy

European Center for
Not-for-Profit Law

Acknowledgements:

Author: Marlena Wisniak, ECNL.

We extend special thanks to Isabelle Anzabi, who contributed significantly during the early stages of the research process. We express gratitude to the ECNL team—Karolina Iwanska, Vanja Skoric, and Francesca Fanucci—for their thoughtful review and feedback on the report.

Valuable input and review were provided by Evani Radiya-Dixit from the American Civil Liberties Union (ACLU); Lindsey Andersen from Business for Social Responsibility (BSR); Mona Elswah and Aliya Bhatia from the Center for Democracy and Technology (CDT); independent researcher and policy expert Luca Belli; and Roya Pakzad from Taraaz.

Insightful contributions through interviews and consultations came from representatives of Meta's Human Rights Team, the Policy and Safety Machine Learning Teams at Discord, and the Research Team at Jigsaw.

We extend our sincere gratitude to everyone who generously contributed their invaluable time, insights, and expertise to the preparation of this report. Your thoughtfulness and creativity have greatly enriched the quality and depth of our findings. We thank Betsy Popken of the UC Berkeley Human Rights Center; Corynne McSherry from the Electronic Frontier Foundation (EFF); Daniel Leufer and Eliska Pirkova from Access Now; Dave Willner of Stanford University; Dunstan Alison Hope; Jonathan Stray from UC Berkeley; Justin Hendrix of Tech Policy Press; Mike Masnick of Techdirt; Paul Barrett from New York University; Sabina Nong of Stanford University; Tarunima Prabhakar from Tattle; and Vladimir Cortes.

Design for the publication was created by Sushruta Kokkula and Andrea Judit Tóth. The illustrations featured in the report are based on the work of Balázs Milánik, Rozalina Burkova (The Greats) and Daniela Yankova (The Greats).

We thank the Omidyar Network for their generous support.

April 2025

Table of contents

Applicability of international human rights law for AI governance

International human rights law, grounded in instruments like the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social and Cultural Rights (ICESCR), provides a tested and globally recognised framework for assessing the potential risks and benefits of AI systems and content moderation—and enables a right to remedy. Human rights principles recognise inalienable rights, such as privacy, non-discrimination, freedom of expression, and freedom of peaceful assembly and association, which must be protected from undue interference. While these protections were historically focused on government obligations, the UN Guiding Principles on Business and Human Rights (UNGPs) have established that businesses—including AI companies— also have a responsibility to respect and uphold human rights.

As AI governance frameworks proliferate, many companies rely on ethics-based or trust and safety-driven approaches to responsible AI. While these frameworks often emphasise fairness, accountability, and harm mitigation, they typically lack consistency, international legitimacy, and are voluntary. By contrast, a human rights-based approach, legally binding for States, offers a universal, internationally recognised, and adaptable framework that applies across jurisdictions and industries and provide a right to remedy.

Given that AI-driven content moderation impacts human rights, integrating these principles into AI development, use, and governance can help AI companies navigate trade-offs and mitigate harm. Ultimately, it will help them protect and promote human rights in their products, services, and activities. International human rights also serve as a common baseline that enables meaningful collaboration between AI developers, deployers, regulators, and civil society, making them an essential foundation for evaluating and addressing risks in generative AI and developing rights-respecting products.

This report aims to highlight the key human rights impacts of using LLMs for content moderation, with a focus on core civic freedoms. While it doesn't follow the methodology of a human rights impact assessment (HRIAs) under the UNGPs or a fundamental rights impact assessment (FRIAs) under the DSA or EU AI Act, our goal is to surface potential positive and negative impacts on a sector-wide level, to guide future HRIAs and FRIAs carried out by AI developers and deployers.

Right to Participate

Legal basis

The right to participate in public decision-making is a fundamental aspect of international human rights law, enshrined in Article 25 ICCPR and further interpreted by the Human Rights Committee in General Comment No. 25.[1] This right extends beyond electoral contexts to include participation in public administration and policy formulation at various levels.[2] It enables inclusive and accessible participation mechanisms, particularly for marginalised or historically excluded groups, such as women, indigenous peoples, and persons with disabilities.[3] When applying this principle to the development and deployment of AI systems, public bodies and states have an obligation to ensure that stakeholders—including those potentially affected by AI technologies—are meaningfully involved in decision-making processes.

As for companies that develop or use LLMs systems, the UNGPs[4] and the OECD Due Diligence Guidance for Responsible Business Conduct[5] both call for stakeholder engagement as part of companies' responsibility to conduct human rights due diligence (including human rights impact assessments and risk mitigation measures) as well as enable access to remedy. In cases where direct consultation is not feasible, the UNGPs recommend engaging alternative representatives, such as independent experts or civil society organisations. Leveraging the right to participate in this way ensures that AI development and use are aligned with human rights principles, fostering accountability and inclusivity.

1 United Nations Human Rights Committee. (1996). General comment adopted by the Human Rights Committee under article 40, paragraph 4, of the International Covenant on Civil and Political Rights (57th sess.). https://digitallibrary.un.org/record/221930?ln=en&v=pdf

2 The OHCHR Guidelines for States on the effective implementation of the right to participate in public affairs
United Nations. (2018, July 20). Guidelines for States on the effective implementation of the right to participate in public affairs. Office of the High Commissioner for Human Rights. https://www.ohchr.org/sites/default/files/Documents/Issues/PublicAffairs/GuidelinesRightParticipatePublicAffairs_web.pdf

3 Ibid.

4 United Nations. (2012). Guiding principles on business and human rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework. Office of the High Commissioner for Human Rights. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

5 OECD. (2018). OECD due diligence guidance for responsible business conduct. OECD Publishing. https://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-Business-Conduct.pdf

Opportunities for participatory development of LLMs

As noted in Business for Social Responsibility's (BSR) recent report on Human Rights Across the Generative AI Value Chain,[6] there is a growing recognition in the field of Responsible AI of the need to include affected stakeholders at every stage of the AI lifecycle— from design and development to deployment, oversight, and impact assessment.[7] Increasingly, research is focusing on participatory methods in AI,[8] including in foundation models,[9] and digital platforms are strengthening their stakeholder engagement processes (e.g. Discord is piloting ECNL's Framework for Meaningful Engagement as they develop machine learning-driven interventions for moderating content online[10]).

As outlined in the DTSP's report, LLMs may have the potential to streamline and enhance external stakeholder engagement. Currently, teams rely on complex combinations of user surveys, focus groups, and consultations with civil society and experts to address trust and safety concerns. These processes are difficult to coordinate and scale effectively. Generative AI could potentially support with synthesising and organising feedback from diverse stakeholders. For instance, LLMs could be used to summarise and cluster content, map themes, and identify gaps in stakeholder input. Relatedly, they could automate the creation of detailed reports that outline how external feedback has been integrated.[11] They could even be leveraged, potentially, to support the deliberation of various opinions and identify areas of consensus (and lack thereof).[12]

Furthermore, LLMs could potentially enable more accessibility in stakeholder engagement by developing multilingual content moderation tools, where diverse cultural and linguistic expertise is essential. NLP tools[13] and Cohere's PRISM.

---

6 Hoh, J. Y., Nigam, S., Andersen, L., & Darnton, H. (2025). Human rights across the generative AI value chain: Human rights assessment of the generative AI value chain and responsible AI practitioner guides. BSR. https://www.bsr.org/en/reports/human-rights-across-the-generative-ai-value-chain

7 BSR (Business for Social Responsibility). (2025). Conducting stakeholder engagement guide: 5 of the Responsible AI Practitioner Guides for taking a human rights-based approach to generative AI. BSR. https://www.bsr.org/files/BSR-Conducting-Stakeholder-Engagement.pdf

8 European Center for Not-for-Profit Law (ECNL) & Society Inside. (2023). Framework for meaningful engagement in human rights impact assessments for AI. European Center for Not-for-Profit Law. https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai ; Young, M., Akinrinade, I., Calderon, A., Lara Guzmán, R., & Onnekikami, T. (2023). Shaping AI systems by shifting power. Data & Society. https://datasociety.net/points/shaping-ai-systems-by-shifting-power/

9 Suresh, H., Tseng, E., Young, M., Gray, M. L., Pierson, E., & Levy, K. (2024). Participation in the age of foundation models. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil (pp. 1–13). ACM. https://doi.org/10.1145/3630106.3658992

10 European Center for Not-for-Profit Law (ECNL). (2023). ECNL and Discord are joining forces to pilot a framework for meaningful engagement. European Center for Not-for-Profit Law. https://ecnl.org/news/ecnl-and-discord-are-joining-forces-pilot-framework-meaningful-engagement

11 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 24) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

12 European Center for Not-for-Profit Law (ECNL). (2024). AI and public participation: Hope or hype? European Center for Not-for-Profit Law. https://ecnl.org/news/ai-public-participation-hope-or-hype

13 Mozilla Foundation. (n.d.). Common Voice. Retrieved February 27, 2025, from https://commonvoice.mozilla.org/en

Alignment[14] are relevant examples.[15] Such efforts could potentially lead to LLMs that are better informed by the lived experiences and linguistic nuances of the communities they aim to serve, mitigating the risks of cultural misrepresentation and linguistic bias.

Beyond enabling localised content moderation, generative AI may have the potential to deliver more nuanced, personalised moderation that aligns with individual user preferences and would possibly enable more external participation in content moderation. Currently, personalisation is limited to basic features such as keyword filters and sensitivity controls. With generative AI, content moderation systems could potentially adapt to specific user preferences in real-time, including considerations for intersectional identity features to better tailor moderation in conversations.[16] However, this approach also raises potential legal challenges in certain jurisdictions. The overarching risks and limitations relevant to personalisation, as outlined in above sections, also apply in this context.

Finally, generative AI could have the potential to facilitate external researchers' access to and comprehension of data.[17] Companies could leverage LLMs to streamline data-sharing processes, such as organising, refining, and annotating datasets, generating comprehensive documentation and guidelines, or offering an interactive chatbot-like tool to assist researchers in interpreting data.[18]

14 Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024, December 3). The PRISM Alignment Dataset: What participatory, representative, and individualised human feedback reveals about the subjective and multicultural alignment of large language models. Cohere. Retrieved from https://cohere.com/research/papers/the-prism-alignment-project-what-participatory-representative-and-individualised-human-feedback-reveals-about-the-subjective-and-multicultural-alignment-of-large-language-models-2024-04-24

15 Radiya-Dixit, E., & Bogen, M. (2024, October). Beyond English-centric AI: Lessons on community participation from non-English NLP groups (p. 2). Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2024/10/2024-10-18-AI-Gov-Lab-Beyond-English-Centric-AI-brief-final.pdf

16 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 23) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

17 Nicholas, G. (2022, September 8). Social media companies should give researchers access to more data. Other industries can show them how. Center for Democracy & Technology. https://cdt.org/insights/social-media-companies-should-give-researchers-access-to-more-data-other-industries-can-show-them-how/

18 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 32) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

Participation challenges of current reinforcement learning methods

Despite these opportunities, significant limitations and concerns remain. Current reinforcement learning methods used for moderating LLMs, such as "red-teaming," RLHF (Reinforcement Learning from Human Feedback),[19] and DPO (Direct Preference Optimisation), rely heavily on human feedback to identify unsafe or biased prompts and responses. Evaluators choose prompts which they deem as "unsafe," seeking to challenge the models through curated adversarial actions. However, these processes often lack inclusivity and diversity. For example, feedback on what constitutes a "safe" or "unsafe" response may vary depending on the evaluator's expertise, cultural context, or lived experience. Most of these evaluators are technical experts in AI, come from upper socio-economic backgrounds, and often embed Silicon Valley values in their processes and analyses.

Due to the underrepresentation of evaluators from marginalised groups and those with specific domain or regional knowledge, the way LLMs are moderated often excludes their perspectives, perpetuating existing biases. What's more, the data AI developers collect from red-teaming evaluators is also used for reinforcement learning, i.e. the main technique enabling LLMs to moderate themselves.[20] Whilst this is primarily a participation issue (or lack thereof), the result has downstream negative effects to other civic freedoms, as the biases embedded in the training process directly influence how the LLMs ultimately moderate content.

The challenges of meaningful external participation are further compounded in multilingual language model development. Initiatives like Meta's No Language Left Behind[21] project and Google's 1000 Languages Initiative[22] ostensibly aim to extend LLM capabilities to non-English languages; yet as mentioned in the section on non-discrimination, they often fall short by relying on machine-translated text, overlooking cultural contexts, and failing to engage deeply with local communities, including allocating adequate resources for such engagement. As CDT warns, the "one model, all languages" approach risks reinforcing the dominance of English and exacerbating linguistic inequities.[23] Without structures for robust and sustained community participation, these efforts may fail to effectively address the unique needs of non-English-speaking communities.

19 Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. arXiv. https://arxiv.org/abs/2305.18290; Zentropi AI. (2024, December 19). COPE-A 9B. Hugging Face. Retrieved from https://huggingface.co/zentropi-ai/cope-a-9b.

20 Roya Pakzad, personal communications, September 5, 2024.

21 Meta AI. (2023). No language left behind: Scaling AI for all languages. Meta AI. https://ai.meta.com/research/no-language-left-behind/

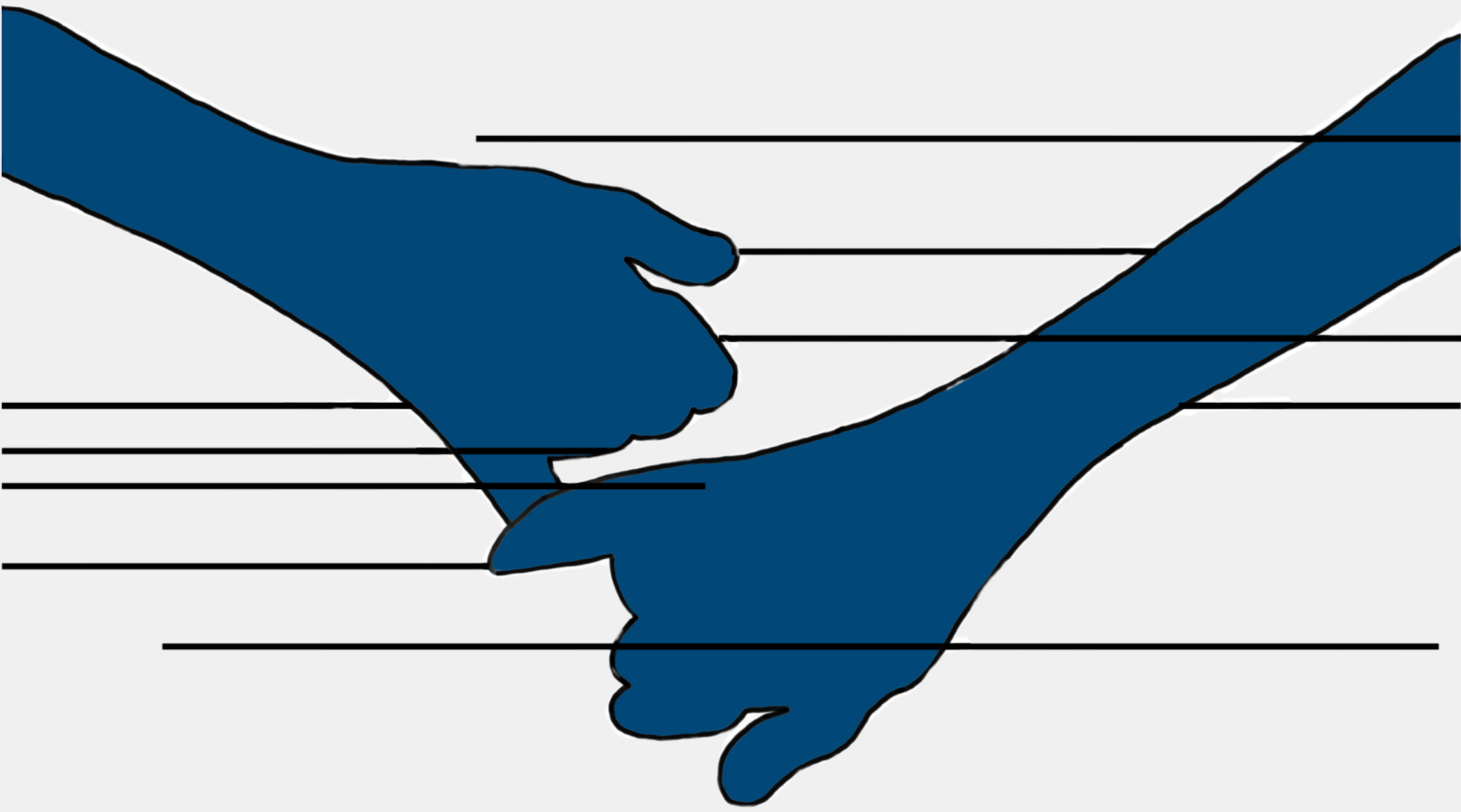
22 Dean, J. (2022, November 2). 3 ways AI is scaling helpful technologies worldwide. Google Blog. https://blog.google/technology/ai/ways-ai-is-scaling-helpful/

23 Radiya-Dixit, E., & Bogen, M. (2024, October). Beyond English-centric AI: A brief (p. 2). Center for Democracy & Technology (CDT). https://cdt.org/wp-content/uploads/2024/10/2024-10-18-AI-Gov-Lab-Beyond-English-Centric-AI-brief-final.pdf

Participatory evaluation of LLMs

Companies evaluate LLMs in various ways, including relying on user feedback through questions such as, “Did you have a good experience?” or “Was this helpful?” Current content reporting systems typically require users to select predefined categories to describe their concerns. This process could potentially be enhanced by integrating LLMs, enabling users to provide explanations through conversational interactions, such as answering, “Why do you find this problematic?”

However, such systems might frustrate users, too. As such, the reporting process itself would benefit from robust evaluation. Overall, evaluation methods for LLMs remain immature, with a notable distinction between systems-based (assessing performance in real-world applications such as social media content moderation, including user interactions and external integrations) and model-based evaluations (testing the model in isolation using commonly-known benchmarks and predefined tasks), particularly when third-party tools are integrated. Without equitable evaluation and benchmarking approaches that account for linguistic diversity and the needs of marginalised communities, selective performance reporting risks obscuring failures in real-world deployments.



Right to Remedy

Legal basis

Article 2(3) ICCPR grants individuals whose rights have been violated access to an effective remedy, including judicial, administrative, or other appropriate means.[24] While the ICESCR[25] does not explicitly outline a right to remedy, the UN Committee on Economic, Social and Cultural Rights affirmed that states must ensure effective mechanisms for individuals to claim their rights and seek redress for violations.[26]

Under the UNGPs, businesses have a responsibility to address human rights violations based on where they fall in the chain of causality. If they cause harm, they must take immediate steps to stop or prevent it. If they contribute to harm, they should not only cease their contribution but also use their influence to reduce any remaining impact as much as possible. When businesses are merely linked to human rights abuses through their operations, products, or services, their response depends on several factors. These include their ability to influence the responsible entity, the severity of the harm, the significance of the business relationship, and whether ending that relationship could lead to further human rights risks. In all cases, businesses should take proactive steps to prevent and address human rights violations in their value chains. Guiding Principles 30-32 outline companies' responsibility to implement effective operational-level grievance mechanisms.

24 United Nations. (1966). International Covenant on Civil and Political Rights. Adopted 16 December 1966 by General Assembly resolution 2200A (XXI). https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

25 Ibid.

26 Committee on Economic, Social and Cultural Rights. (1998). General comment no. 12: The right to adequate food (art. 11 of the Covenant). Refworld. https://www.refworld.org/legal/general/cescr/1998/en/53238

User notification

Generative AI and LLMs could potentially play a role in enhancing transparency and understanding for users affected by content takedowns and appeals. For example, LLMs could explain why certain content was flagged, removed, or recommended.

In the context of access to remedy, LLMs could perhaps support the first level of appeals within platforms by notifying users of decisions and allowing immediate appeals.[27] By providing more granular information, such as specifying why content was removed and offering actionable steps for correction, users could better understand the enforcement actions taken against their content—and how to appeal decisions where applicable. For instance, LLMs might highlight specific timestamps in a one-hour video that contain potential violations, enabling users to make targeted corrections and submit more informed appeals.[28] Unlike human moderators, who are often limited by time constraints, LLMs could generate instant explanations based on the text they processed.[29]

Furthermore, generative AI could potentially serve as an automated tool for managing user reports, appeals, and customer service requests. For example, it could be used to analyze the content of a query, support with sentiment analysis, cross-reference previous customer support interactions, or determine if the issue requires further action. It could then either provide an automated response or escalate the matter to a human representative. However, implementing an automated appeal process for an already automated decision could risk creating a self-perpetuating loop in the context of content moderation. Although LLMs could be useful for triaging appeals, they should not be responsible for making substantive decisions.

These capabilities could possibly support compliance with legal requirements for explainability, as LLMs could in theory record and explain the reasoning behind content moderation decisions, and store contextual information about their decisions. By automating parts of the process, generative AI could potentially make it more efficient to produce and present transparency reporting data in a way that is more accessible and understandable to a range of audiences. This could help meet the increasing disclosure requirements under emerging regulations, while also making transparency reports more user-friendly and accessible for diverse audiences.[30]

27 Eliska Pirkova, personal communication, September 10, 2024.
28 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 32) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/
29 Willner, D. (2024, April). Transcript: Dave Willner on moderating with AI. The Institute for Rebooting Social Media. https://www.techpolicy.press/transcript-dave-willner-on-moderating-with-ai-at-the-institute-for-rebooting-social-media/
30 Digital Trust & Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety. (p. 33) Digital Trust & Safety Partnership. https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

Challenges related to explainability and accuracy

The above opportunities are limited by the fact that explanations generated by LLMs may misrepresent the decisions of the model or the policies they enforce, raising concerns about their accuracy and overall reliability. This challenge is exacerbated by the issue of "hallucinations," leading to inaccurate statements, and preventing users from adequately seeking remedy. It also underscores the urgent need to further explore best practices for integrating LLMs into remedy processes, with a focus on user experience and alignment with regulatory frameworks like the EU Digital Services Act (DSA).

Moreover, the lack of explainability in LLMs, particularly in multilingual contexts, complicates access to remedy by obscuring how these models generate associations and decisions.[31] As platforms increasingly adopt third-party LLMs, such as ChatGPT or Llama, rather than developing models in-house, accountability becomes more diffuse. Previously, companies could address issues internally by directly modifying the model or training data themselves. However, reliance on external LLM providers introduces a new layer of complexity in content moderation, raising questions about who is responsible for addressing errors—the platform (i.e. LLM deployer) or the LLM provider.[32] Viewed through the lens of the UNGPs, it is essential to undertake a deeper analysis of how LLM developers and deployers may cause, contribute to, or be linked to adverse human rights impacts.

31 Aliya Bhatia, personal communication, August 1, 2024.
32 Roya Pakzad, personal communications, September 5, 2024.

Algorithmic Gatekeepers: VII. Right to Participate and Remedy

European Center for
Not-for-Profit Law