Algorithmic Gatekeepers: The Human Rights Impacts of LLM Content Moderation

IV. Right to Freedom of Expression, Information and Opinion

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 Freedom of Expression

Legal basis

Article 19 ICCPR guarantees the right to freedom of expression, ensuring everyone can seek, receive, and impart information and ideas across borders and through any medium. States must avoid unjustified restrictions while also taking positive steps to foster free expression, including protecting individuals from undue interference by private entities. In 2011, the UN Human Rights Committee affirmed that freedom of expression fully applies to all electronic and internet-based communication, reinforcing its importance online.[1] In the digital age, this means ensuring procedural safeguards for online content moderation.

Article 19(3) ICCPR allows for certain restrictions, but only when they are provided by law and necessary to respect the rights or reputations of others or to protect national security, public order, public health, or morals. Any limitations must meet a strict test of legality, necessity, and proportionality, ensuring they are not used to suppress dissent or censor legitimate speech. The UN Human Rights Committee has emphasised that restrictions must not be overly broad or vague and should never be used to silence criticism of the government, political opposition, or marginalised voices.

Regarding automated content moderation, these systems often lack the transparency and accountability required to show that the conditions of the three-part test are fulfilled, raising concerns about overreach and unintended discrimination.Digital platforms, as private entities, are entitled to enforce community standards stricter than international human rights norms. However, their terms of service and moderation practices must align with core human rights principles, such as necessity, proportionality, and non-discrimination.[2] Striking this balance is critical to ensuring that the use of LLMs in content moderation supports, rather than undermines, freedom of expression while fostering safe and inclusive online spaces.

Over- and underenforcement of content policies

LLMs introduce opportunities for new forms of content moderation that do not involve outright takedown of content. Instead, interventions like labeling, contextualisation, or providing alternative perspectives can promote responsible online discourse and promote the freedom of expression. Improving LLM accuracy could potentially lead to more consistent, narrowly tailored moderation outcomes, avoiding the overly broad actions of less sophisticated systems. Such advancements could provide an alternative to existing machine learning methods and human moderators, offering greater consistency and fairness in content moderation. For example, the dating app company

[1] United Nations Human Rights Committee. (2011, September 12). General comment No. 34 on article 19: Freedoms of opinion and expression (CCPR/C/GC/34, paras. 12, 17, & 39). Retrieved from https://documents.un.org/doc/undoc/gen/g11/453/31/pdf/g1145331.pdf?OpenElement

[2] ARTICLE 19. (2023, August). Content moderation handbook. ARTICLE 19. Retrieved from https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf

Match Group is using AI to identify signals that a message may be inappropriate or excessively sexual.[3]

However, automated content moderation, especially by LLMs, can also negatively impact this right significantly. Actions such as automated content removal, downranking, account suspension, or account removal can suppress legitimate speech, particularly when decisions are opaque, arbitrary, or discriminatory. As digital platforms scale, even highly accurate content moderation models can produce significant volumes of over-action and under-action errors, leading to unintended consequences.[4]

For LLMs used in content moderation, accuracy remains a critical concern partly due to the inherent vagueness of language and the dynamic nature of cultural norms—an issue pertinent to automated content moderation in general. While LLMs can outperform traditional machine learning models and human moderators in some areas, accuracy rates for specific tasks remain concerning.[5] For instance, one LLM-based moderation system achieved a true-negative rate of 92.3% but struggled with a true-positive rate of only 43.1%, failing to effectively flag rule-violating content.[6] This performance gap highlights the need for independent audits of LLM-based moderation systems and for platforms to release and rigorously evaluate their internal audits to increase transparency and accountability.

Over-action—removing content that does not violate platform policies—has profound implications for freedom of expression. LLMs with insufficient precision may lead to overly broad takedowns of content, disproportionately silencing legitimate voices, especially those of marginalised communities. For example, posts discussing sensitive but lawful topics may be removed due to an inability to differentiate nuanced discussions from harmful content. These false positives can create a chilling effect for freedom of expression, discouraging users from sharing opinions and participating in public discourse online on some of the most sensitive and important topics for fear of being sanctioned.

Conversely, under-action—failing to remove content that violates policies or the law—also carries significant risks. Harmful content left unchecked can lead to incitement against targeted groups and prevent affected individuals from fully participating in public life online. For instance, false negatives, such as hate speech, abuse and harassment, or incitement to violence, can make users, particularly women or marginalised groups, feel unsafe online. This often results in self-censorship or users opting out of platforms entirely, eroding their ability to engage in digital spaces. Importantly, it also leads to harm and violence offline. Under-action ultimately undermines trust in moderation systems, further isolating vulnerable communities.

[3] Financial Times. (n.d.). Match enlists AI to nudge men into better behaviour on dating apps. Retrieved April 1, 2025, from https://www.ft.com/content/4e39d08b-41ef-41ea-abc0-952d06324484

[4] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 3). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[5] Huang, T. (2024). Content moderation by LLM: From accuracy to legitimacy. arXiv. https://arxiv.org/pdf/2409.03219

The researcher Tao Huang argued that accuracy alone is misleading, as it fails to distinguish between simple and complex cases or acknowledge trade-offs in achieving higher accuracy. Content moderation is fundamentally about building legitimacy, not just correctness. For straightforward cases, accuracy, speed, and transparency are key, while complex cases require reasoned justification and user participation. The primary goal of LLMs should be to enhance legitimacy rather than simply ensure correctness.

[6] Kolla, M., Chandrasekharan, E., Salunkhe, S., & Saha, K. (2024, May). LLM-Mod: Can large language models assist content moderation? University of Illinois Urbana-Champaign. Retrieved from https://koustuv.com/papers/CHI24_EA_LLM-Mod.pdf

Decisions about precision[7] and recall[8] thresholds in LLM moderation significantly affect the balance between protecting users and preserving freedom of expression. By improving precision and recall, LLMs could more effectively identify harmful content while minimising the risk of silencing legitimate voices. Indeed, high precision reduces false positives but risks letting harmful content remain online, while high recall minimises false negatives at the expense of mistakenly taking down legitimate content.[9] These trade-offs underscore the importance of stakeholder engagement, especially as automation increases the scale and impact of content moderation decisions. Without input from diverse communities, thresholds may fail to be narrowly tailored or justifiable, negatively impacting freedom of expression.

Disproportionate impact on marginalised groups

Biases within automated content moderation models frequently lead to unequal treatment of marginalised communities, especially those in the Global Majority. While applicable to automation in content moderation generally (and not specific to LLMs), these biases reflect a broader inability of classifiers to navigate the nuances of harmful speech and perpetuate systemic inequities, as elaborated under the section on the prohibition of non-discrimination.

For instance, automated systems often fail to detect or adequately address hate speech directed at Black individuals, while disproportionately penalising these same communities for perceived violations.[10] An internal investigation at Facebook revealed that some of the most harmful content left on the platform targeted Black people, whereas takedowns disproportionately involved hateful posts about White individuals.[11] Inadequate recognition of cultural and linguistic differences further amplifies content moderation challenges. Misinterpretation of culturally specific expressions or language nuances can lead to unjust takedowns or enforcement actions. A specific issue with LLMs is their ability to learn problematic word associations that reflect biases against specific groups.

Reclaimed language poses further challenges to automated content moderation.[12] Marginalised groups often reappropriate slurs or harmful terms as acts of empowerment, creating unique contexts for their use. As such, language that might be considered discriminatory in a broader context may instead form part of the dialect of a particular community and hold no offensive meaning within that space. However, automated systems struggle to distinguish between the reappropriated use of these terms and harmful intent, leading to false positives and overenforcement. For example,

[7] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 3). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/ "This metric measures the correctness of a model's positive predictions (flagged violations). It's calculated as the number of accurate flags divided by the total number of all flags (accurate and inaccurate). A high precision indicates the model is good at identifying actual violations." See definition in DTSP.

[8] Ibid., "This metric measures how well the model catches all of the actual violations. It's the proportion of actual violations that are correctly flagged by the model. This is a measure of the True Positive Rate (TPR)." See definition in DTSP.

[9] Ibid.,

[10] Chung, A. W. (2019, January 24). How automated tools discriminate against Black language. MIT Center for Civic Media. https://civic.mit.edu/index.html?p=2402;
Appelman, N. (2021, November 26). 'Race-blind' content moderation disadvantages Black users. Racism and Technology Center. https://racismandtechnology.center/2021/11/26/race-blind-content-moderation-disadvantages-black-users/

[11] Narayanan, A. (2024). Snake Oil: Why Can't AI Fix Social Media? Princeton University Press, p. 187.

[12] Using LLMs to Moderate Content.

researchers documented instances of over-moderation in LGBTQ discussion spaces.[13] The acceptability of reclaimed language ultimately depends on the speaker's identity and the context of its use. When these complexities are overlooked, marginalised users are disproportionately affected, with their posts wrongly flagged or removed. This not only violates their freedom of expression but also silences important cultural and social discussions.

Contextual (mis)understanding

Perhaps the primary advantage of integrating LLMs into content moderation lies in their potential to better understand context, enabling more accurate assessments of content against existing policies. Unlike earlier systems (e.g. machine learning), LLMs are designed to interpret nuances within text, providing hope for more sophisticated moderation of complex content categories. Traditional machine learning models can be reasonably accurate for detecting clearly defined content, such as child sexual abuse material,[14] and "content or conduct for which there are clear rules or legal parameters, and content that has a hash match in a hash database."[15] However, they work poorly when moderating more subjective categories like hate speech or terrorism, as these latter categories require enhanced contextual understanding.[16]

In situations where content moderation requires nuanced decision-making or the model has lower confidence, LLMs could serve as a supportive tool for human moderators. For instance, they could flag potentially violative content for human review, assess trends in enforcement, and help refine moderation policies. This collaborative approach could potentially improve the quality of decisions as well as enhance the training process by allowing human reviewers to label complex cases and feed this data back into the model, thereby improving its performance over time.[17]

"Algospeak," such as substituting "unalive" for "suicide" or "le$bean" for "lesbian," is sometimes used to bypass automated content moderation,[18] making it particularly challenging to moderate. Emerging technologies like LLMs (e.g., GPT-4o) show promise in detecting algospeak, with research indicating they can decode up to 98.5% of altered terms when provided with context. [19] However, applying these tools requires balancing the need to curb harmful content with the protection of legitimate expression, ensuring marginalised voices aren't inadvertently silenced. While malicious actors exploit it to spread hateful or harmful content undetected, algospeak also serves

[13] Algorithmic arbitrariness in content moderation. (2024). Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–12. Retrieved from https://facctconference.org/static/papers24/facct24-151.pdf.

[14] Technology Coalition. (2023, September 15). Update on voluntary detection of CSAM. Technology Coalition. Retrieved from https://www.technologycoalition.org/knowledge-hub/update-on-voluntary-detection-of-csam

[15] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 3). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[16] ARTICLE 19. (2023, August). Content moderation handbook (pp. 40-41). ARTICLE 19. Retrieved from https://www.article19.org/wp-content/uploads/2023/08/SM4P-Content-moderation-handbook-9-Aug-final.pdf

[17] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 3). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[18] Steen, E., Yurechko, K., & Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. Social Media + Society, 9(3). https://doi.org/10.1177/20563051231194586

[19] Fillies, J., Paschke. A. (2024). Simple LLM based Approach to Counter Algospeak. Proceedings of the WoAH Conference. Retrieved from https://aclanthology.org/2024.woah-1.10.pdf

as a vital tool for marginalised groups and civic space actors to express themselves freely. Indeed, these groups often use algospeak to protect their content from unjust takedowns, especially in contexts like documenting human rights abuses or when discussing politically sensitive topics. In these case, upload filters or hash databases can disproportionately flag their posts under policies like violent extremist or terrorist content. Effective moderation must prioritise nuanced understanding of the context to safeguard free expression while addressing moderation avoidance effectively.

Crisis and exceptional or unusual content

LLMs, including generative AI models, face significant challenges when encountering out-of-distribution (OOD) data—content that deviates from the patterns in their training datasets. As such, LLMs are not well-suited for crises, exceptions, or unusual situations, such as conflicts, due to their inability to handle OOD effectively. Crises and conflicts are not statistically common, meaning they are underrepresented in the datasets used to train LLMs. Since these models rely on patterns observed during training, they struggle to identify, interpret, or respond to content that falls outside these patterns. Behavior that might be innocuous in one context could be harmful in another, but LLMs, trained predominantly on generalised patterns, may fail to make these distinctions.[20] For example, the emergence of new slurs, cultural references, or context-specific terms during a conflict may go unrecognised or misunderstood.

In automated content moderation, this probabilistic design can thus lead to two critical issues. First, harmful or violative content might be missed entirely if the model cannot recognise it as such due to its absence in training data. In this case, harmful content would be overlooked, potentially inciting to or escalating violence in a conflict. Second, the model might misinterpret content (e.g. made in irony or sarcasm during a conflict), flagging benign material as harmful or vice versa, due to its reliance on generalised probabilities rather than specific contextual understanding. These errors undermine the effectiveness of moderation systems, particularly in dynamic environments where new forms of harmful content emerge regularly.

Moderating OOD often results in "hallucinations," where the model generates inaccurate or inappropriate responses. These errors are particularly problematic in content moderation settings, where accurate context and nuance are critical for identifying harmful behavior or content. In crisis situations, where misinformation and misinterpretation can have serious consequences, these hallucinations could exacerbate harm. "Hallucinations" are not mere bugs but inherent features of LLMs, stemming from their transformer architecture. LLMs operate on probabilistic principles, predicting the next word in a sequence based on patterns in their training data. This approach prioritises producing natural-sounding and contextually plausible text over ensuring factual accuracy. In essence, LLMs function as advanced autocomplete systems, generating outputs that align with the statistical likelihood of given sequences rather than the actual veracity of the information.[21]

During crises, the rapid evolution of language, behavior, and content makes it essential for moderation systems to adapt quickly. Generative AI models, typically trained on static datasets, lack the real-time adaptability required to address emerging trends effectively. While LLMs hold promise for adapting more rapidly than traditional

[20] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 34). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[21] Allen, D., Denkovski, O., & Giannaccini, F. (2024, October 14). Ensuring AI accountability: Auditing methods to mitigate the risks of large language models. Democracy Reporting International. Retrieved from https://democracy-reporting.org/en/office/EU/publications/ensuring-ai-accountability-auditing-methods-to-mitigate-the-risks-of-large-language-models#UsingLLMstoEvaluateLLMs%E2%80%99Outputs

machine learning systems in the future, their current lag can lead to delayed or inadequate responses to harmful content, exacerbating risks in high-stakes scenarios.

Government censorship

AI developers and deployers bear much of the responsibility for adverse impacts on freedom of expression of LLM moderation, but governments' role in leveraging LLMs introduces new challenges as well. Governments, particularly those with authoritarian practices, could potentially exploit LLMs' improved classification abilities to conduct censorship on an unprecedented scale, suppressing political dissent, silencing journalists and other civic space actors, and possibly criminalising them.[22]

LLMs could enable governments to censor content more efficiently than traditional machine learning systems. By using advanced contextual understanding, authoritarian regimes could automate the identification and removal of dissenting voices, making censorship harder to bypass. This would not only restrict freedom of expression but also diminish the diversity of available digital content. Past examples, such as the removal of Syrian war footage from YouTube,[23] highlight how automated content moderation practices disproportionately affect journalists and those documenting human rights violations. LLMs' broader, automated censorship capabilities could amplify these concerns. There is also ongoing research questioning the impact of existing censorship online, which limits the scope of available digitised data (e.g. Chinese internet[24]) on the quality of language and content within LLMs,[25] further entrenching State narratives and silencing dissent.

One emerging trend could be the potential for governments to produce machine-readable regulations that can be directly interpreted and enforced by LLMs. This innovation would allow regulators to write policies in a format that platforms could use as prompts to automate compliance. While this may streamline regulatory implementation, it also concentrates significant power in the hands of governments, potentially enabling them to dictate how platforms handle speech. In authoritarian contexts, this power could lead to broad enforcement of restrictive laws, further criminalising dissent and eroding freedom of expression. Moreover, such capabilities could reduce platform accountability, as companies might defer responsibility to governments for decisions about speech and content moderation.[26]

Concentration of power

[22] Edwards, E. (2023, August 21). Large language models will be great for censorship. LessWrong. Retrieved from https://www.lesswrong.com/posts/oqvsR2LmHWamyKDcj/large-language-models-will-be-great-for-censorship

[23] Kyle, P. A. (2024, July). Machine learning can undermine human rights: YouTube's struggle to moderate the Syrian crisis. Trust and Safety Foundation. Retrieved from https://trustandsafetyfoundation.org/blog/machine-learning-can-undermine-human-rights-youtubes-struggle-to-moderate-the-syrian-crisis/

[24] Yuan, L. (2024, June 4). As China's Internet Disappears, 'We Lose Parts of Our Collective Memory'. The New York Times. Retrieved from https://www.nytimes.com/2024/06/04/business/china-internet-censorship.html

[25] Ahmed, M., Knockel, J. (2024). Extended Abstract: The impact of Online Censorship on LLMs. Proceedings of the 2024 PET Symposium. Retrieved from https://www.petsymposium.org/foci/2024/foci-2024-0006.pdf

[26] Stanford Cyber Policy Center. (2024, April 23). The Future of Content Moderation and its Implications for Governance| S.Chakrabarti and D.Willner [Video, 34:00]. YouTube. https://www.youtube.com/watch?v=-JMq49FZ5qmY

In the near to medium term, it remains unlikely that organisations will entirely replace their systems with LLMs from a single provider. Most existing approaches continue to rely primarily on human evaluation of generated samples or adopt a hybrid human-AI evaluation process to assess sample sets prior to retraining the model. However, as human evaluation increasingly gives way to LLM-based moderation, the risk of widespread errors grows significantly. A single mistake within the system could rapidly scale, impacting millions of users and amplifying the consequences of faulty or biased decisions across entire platforms.

While LLMs can bring consistency to moderation, reliance on a few dominant LLMs and APIs raises concerns about power concentration and systemic failure. Disruptions—such as vendor downtime or errors—could potentially leave platforms unmoderated or poorly moderated at scale, allowing spam or harmful content to persist while legitimate content is mistakenly removed due to the fallibility of large-scale LLM moderation.[27]

Relatedly, the growing reliance on a handful of LLMs for content moderation introduces significant risks to diversity in online speech. Since there are only a limited number of LLM providers, most platforms fine-tune these foundational models to align with their unique content policies. However, the underlying values, biases, and assumptions baked into the baseline LLMs influence the deployment process, creating a ripple effect across platforms that use these models or their APIs. This convergence heightens the risk of systemic biases and homogenised speech. Values embedded in an LLM, such as overly restrictive rules on specific types of speech, can cascade down to platforms deploying these models, making it difficult for them to deviate from these judgments without significant effort or resources—at the very least, they need to ensure rigorous finetuning.

This dynamic leads to further centralisation of speech. Decisions made during the initial training of an LLM—such as how it addresses contentious issues like hate speech, misinformation, or political dissent—have far-reaching consequences for every platform that adopts or modifies the model. This impact is especially pronounced when smaller platforms deploy safety-fine-tuned models for content moderation (e.g., Llama Guard[28]). The models they deploy will reproduce the moderation decisions embedded within the foundational models. Unless they comprehensively fine-tune the model, platforms deploying these models can inherit and perpetuate those baseline decisions, even if their own content policies differ.

For example, if an LLM developer misclassifies groups such as pro-Palestine users as terrorists due to biases in training datasets or the models, pro-Palestinian voices will also be silenced on the platform that deploys the LLM internally, unless the model is meticulously fine-tuned for this specific purpose. In practice, it's unlikely that all biases and errors will be fully addressed at the fine-tuning level. Consequently, this could result in the suppression of marginalised voices and the reinforcement of

[27] Aliya Bhatia, personal communication, August 1, 2024.

[28] Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Testuggine, D., & Khabsa, M. (2023, December 7). Llama Guard: LLM-based input-output safeguard for human-AI conversations. Meta AI. https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/

harmful stereotypes.[29] What's more, governments or malicious actors could potentially exploit these weaknesses by influencing or manipulating LLM moderation systems to suppress dissenting voices or promote disinformation.

This centralisation limits the diversity of approaches to content moderation, as foundational LLMs establish a default framework that many platforms inadvertently replicate in ways that are not immediately apparent. The issue resides at the underlying model level, making it difficult to distinguish between problems inherent to the model itself and those introduced through fine-tuning by the deployer. Indeed, it allows a few models to heavily influence global speech moderation norms, reducing diversity in ideas and values online and entrenching some ideologies or cultural perspectives (often powerful voices) while marginalising others.

[29] Oversight Board. (2024, March 26). Oversight Board publishes policy advisory opinion on referring to designated dangerous individuals as "Shaheed". Oversight Board. https://www.oversightboard.com/news/oversight-board-publishes-policy-advisory-opinion-on-referring-to-designated-dangerous-individuals-as-shaheed/;
Wisniak, M., Moussa, R., & York, J. C. (2023). Submission to Policy Advisory Opinion 2023-01 [Policy submission]. European Center for Not-for-Profit Law (ECNL) & Electronic Frontier Foundation (EFF). https://ecnl.org/sites/default/files/2023-05/ECNL%20EFF%20Submission%20to%20Policy%20Advisory%20Opinion%202023.pdf

Right to Freedom of Information

Legal basis

Under article 19 ICCPR,[30] everyone has the right to seek and receive information and ideas through any media and regardless of frontiers. Access to information is the "right of the public to have access to information of public interest."[31] This right imposes obligations on states to ensure access to information, including government-held data, and to foster an open environment where individuals can freely exchange ideas.

There is a "limited scope of exceptions: reasons for the denial of access to information should be clearly and narrowly designed, bearing in mind the principles of legality, necessity and proportionality."[32] The same restrictions under Article 19(3) ICCPR apply as outlined under the right to freedom of expression.

Unreliability to counter misinformation

Some researchers argue that LLMs may hold some (limited) potential in improving misinformation detection when properly augmented. For instance, researchers proposed "MUSE," an LLM combining capabilities with access to real-time information retrieval and credibility evaluations.[33] By retrieving and cross-referencing evidence, MUSE could identify and explain inaccuracies in content, whether textual or multimodal, with references to support its findings.[34] Additionally, LLMs could potentially support efforts like automated Community Notes on social media, scaling up the review process to counter misinformation. [35]

That said, most researchers remain extremely cautious about relying on LLMs for misinformation detection due to their current limitations. As LLMs can only analyze the features present within the content itself, they are not "truth machines" capable of assessing the factual accuracy of a statement unless the information aligns with their training data.[36] For example, LLMs cannot label content as misinformation based on external verification, nor can they provide accurate outputs if the relevant information

[30] Office of the United Nations High Commissioner for Human Rights (OHCHR). (n.d.). International Covenant on Civil and Political Rights. OHCHR. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

[31] Office of the United Nations High Commissioner for Human Rights (OHCHR). (n.d.). Fact sheet No. 5: The right to freedom of expression. OHCHR. Retrieved from https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/Factsheet_5.pdf

[32] Ibid.

[33] Zhou, X., Sharma, A., Zhang, A. X., & Althoff, T. (2024). Correcting misinformation on social media with a large language model. arXiv. https://arxiv.org/abs/2403.11169.

[34] Ibid.

[35] Zhou, X., Sharma, A., Zhang, A. X., & Althoff, T. (2024). Correcting misinformation on social media with a large language model. arXiv. https://arxiv.org/abs/2403.11169.

[36] Stanford Cyber Policy Center. (2024, April 23). The Future of Content Moderation and its Implications for Governance| S.Chakrabarti and D.Willner [Video, Limitations Section]. YouTube. https://www.youtube.com/watch?v=JMq49FZ5qmY

is missing from their accessible data. This constraint is further compounded by their tendency to "hallucinate," leading to fabricated sources, false historical information, or even unsafe advice when prompted with questions that extend beyond their knowledge cutoff or training scope. Indeed, the DTSP itself cautioned that "the propensity of genAI models to sometimes hallucinate and exhibit unanticipated behavior, as well as the ways in which they can reproduce bias, can make genAI less reliable for knowledge retrieval and interpretation tasks and may also lead to classification errors."[37]

These issues raise significant concerns when seeking to address misinformation, particularly in high-stakes areas like public health or political discourse. LLM-powered systems like ChatGPT can invent false sources or output speculative and imprecise responses, which undermines their reliability.[38] They can also withhold critical information. For example, the Chinese AI startup DeepSeek refuses to provide information related to human rights abuses in Tiananmen or against Uyghurs in Xianjiang, exposing censorship on both the application and training level.[39] Without access to real-time updates or robust cross-referencing mechanisms, LLMs risk reinforcing inaccuracies rather than correcting them.

Efforts to create AI systems capable of evaluating the truth of statements in real time remain problematic. Narayanan and Kapoor, co-authors of AI Snake Oil and scholars from Princeton University, warn that even state-of-the-art models currently act more like "bullshit generators," generating plausible-sounding but often inaccurate outputs. Systems designed to detect misinformation could inadvertently lead to overreach, reinforcing dominant political or scientific narratives while suppressing dissenting voices, including legitimate critiques.[40]

Crises and exceptional or unusual events

LLM moderation carries significant risks when addressing misinformation related to sensitive or high-stakes events such as conflicts, crises, or elections. Notably, studies have revealed that these models often produce inaccurate or harmful outputs when tasked with moderating or providing information related to voting.[41] For instance, in the context of the 2024 European Parliament elections, LLM chatbots provided users with incorrect registration deadlines, misleading voting methods, and irrelevant resources.[42] Similarly, LLM responses ahead of the 2024 U.S. presidential election frequently violated voting rights by falsely claiming that no local polling stations existed within users' areas.[43] Such failures demonstrate the limitations of LLMs in upholding accuracy and neutrality in critical content moderation contexts, especially mis- and disinformation.

[37] Digital Trust and Safety Partnership. (2024, September). Best practices for AI and automation in trust and safety (p. 41). Retrieved from https://dtspartnership.org/best-practices-for-ai-and-automation-in-trust-and-safety/

[38] Allen, D., Denkovski, O., & Giannaccini, F. (2024, October 14). Ensuring AI accountability: Auditing methods to mitigate the risks of large language models. Democracy Reporting International. Retrieved from https://democracy-reporting.org/en/office/EU/publications/ensuring-ai-accountability-auditing-methods-to-mitigate-the-risks-of-large-language-models#UsingLLMstoEvaluateLLMs%E2%80%99Outputs

[39] Yang, Z. (2025, January 31). Deepseek's tool reveals how AI could be used for censorship. Wired. Retrieved from https://www.wired.com/story/deepseek-censorship/#main-content

[40] Narayanan, A. (2024). Snake Oil: Why Can't AI Fix Social Media? Princeton University Press, p. 197.

[41] Allen, D., Denkovski, O., & Giannaccini, F. (2024, October 14). Ensuring AI accountability: Auditing methods to mitigate the risks of large language models. Democracy Reporting International. Retrieved from https://democracy-reporting.org/en/office/EU/publications/ensuring-ai-accountability-auditing-methods-to-mitigate-the-risks-of-large-language-models#UsingLLMstoEvaluateLLMs%E2%80%99Outputs

[42] Ibid.

[43] Ibid.

In crisis situations, where content moderation requires real-time adaptability, contextual understanding, and high accuracy, generative AI models fall short, as with automated content moderation in general. Their reliance on historical patterns and probabilistic generation makes them ill-equipped to handle the dynamic and context-specific nature of crises and unusual events. Generative AI models may fail to distinguish between content that is harmful and content that is necessary for raising awareness or reporting on the crisis. For instance, graphic content shared to document human rights violations could be flagged as harmful and removed, hindering important advocacy and information-sharing efforts.

Efforts to reduce hallucinations and inaccuracies in LLM content moderation outputs often address surface-level issues but fail to tackle the underlying architectural challenges that give rise to these errors. In some cases, LLMs were fine-tuned to avoid responding to specific sensitive queries—such as those about European Parliament elections in English—yet continued to deliver incorrect answers in other languages.[44] This underscores a broader issue: LLMs tend to perform poorly in languages or contexts where their training data and oversight are less comprehensive or robust, leading to inconsistencies in content moderation outcomes.[45]

Moreover, LLMs' propensity to oversimplify complex topics and lean towards dominant narratives or common beliefs poses a subtle but significant risk to public understanding. When users or platforms rely on LLMs to curate news or moderate content, minor yet repeated inaccuracies can gradually distort public knowledge. This slow erosion of truth highlights a fundamental limitation of LLMs in content moderation, particularly when tasked with safeguarding accuracy in sensitive areas. While their factual accuracy may improve with successive updates and proper fine-tuning, the potential for error makes LLMs a notable source of misinformation and bias in content moderation efforts.

Government restriction to information

LLM moderation may inadvertently suppress public interest information due to biased datasets and false positives in moderation algorithms. Content flagged as misinformation or "fake news" is often removed at scale, sometimes based on government takedown requests. This automated approach risks silencing important information, especially when governments use these mechanisms to suppress dissent or remove newsworthy content. The disproportionate removal of content created by marginalised groups compounds this problem, as it further limits public access to diverse perspectives and critical information.

What's more, governments could potentially exploit prompt engineering or other techniques to directly moderate content using LLMs, resulting in fragmented access to information and the potential "splintering" of the internet (i.e the fragmentation of the global internet into separate, regionally controlled, limiting the free flow of information across borders). Such practices not only undermine the right to freedom of information but could lead to the erasure of content critical for public accountability in some countries or regions, such as evidence of human rights abuses. As mentioned above, the removal of Syrian war footage from YouTube underscores how automated

[44] Allen, D., Denkovski, O., & Giannaccini, F. (2024, October 14). Ensuring AI accountability: Auditing methods to mitigate the risks of large language models. Democracy Reporting International. Retrieved from https://democracy-reporting.org/en/office/EU/publications/ensuring-ai-accountability-auditing-methods-to-mitigate-the-risks-of-large-language-models#UsingLLMstoEvaluateLLMs%E2%80%99Outputs
[45] Ibid.

systems can inadvertently suppress crucial evidence of war crimes.[46] The removal of such material violates human rights by obstructing justice and impeding public awareness of atrocities.

LLM moderation can disproportionately affect journalists, whose work often challenges state narratives and exposes human rights abuses. Automated takedowns can limit their ability to report on sensitive topics, further diminishing the free flow of information necessary for informed public discourse.[47] Conversely, LLMs could potentially be weaponised by governments to generate state-sponsored propaganda or flood platforms with misinformation, creating "censorship through noise."[48] This tactic aims to overwhelm reliable journalistic information with irrelevant or misleading content, making it harder for users to discern trustworthy sources. Such erosion of meaningful discourse compromises the right to access accurate and unbiased information.

Personalised content and news feeds

LLM moderation may be used to enhance user access to relevant content through personalised news feeds. Proponents like Meta's Shirazyan and Sissons argue that LLMs and AI systems enable novel ways for individuals to access and share information, offering intuitive and conversational experiences. Users can ask personalised questions, explore complex topics naturally, and receive precise, contextually relevant answers. They claim that such interactions could democratise access to knowledge and skills, making information retrieval more inclusive and efficient.[49]

However, the downside of personalised news feeds is that they can limit exposure to diverse perspectives, posing challenges in ensuring that users have access to critical and varied sources of information. Moreover, the extent to which LLMs genuinely enhance the right to information is debatable. Without deliberate efforts by platforms to prevent barriers like limited search exploration or overly personalised feeds—and unless contextual understanding improves significantly—the benefits of these systems may not materialise. Existing algorithmic recommender systems already cater to user preferences, yet they often risk reinforcing echo chambers and limiting the discovery of new or diverse content.

[46] Kyle, A. P. (2024, July). How machine learning can undermine human rights: YouTube's struggle to moderate the Syrian crisis. Trust and Safety Foundation. Retrieved from https://trustandsafetyfoundation.org/blog/machine-learning-can-undermine-human-rights-youtubes-struggle-to-moderate-the-syrian-crisis/

[47] Radsch, C. (2023, February 27). The challenge of platform capture. Columbia Journalism Review. Retrieved from https://www.cjr.org/special_report/disrupting-journalism-how-platforms-have-upended-the-news-part-8.php

[48] Cybersecurity and Infrastructure Security Agency (CISA). (2024). Infographic on the tactics of disinformation. https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf

[49] O'Brien, M. (2020, October 21). AI's potential to advance human rights. Just Security. Retrieved from https://www.justsecurity.org/98097/ais-potential-to-advance-human-rights/n

Deepfakes and watermarking

Watermarking AI-generated content with the use of LLMs could positively impact the right to freedom of information by providing the source of content. However, current watermarking and detection tools are far from perfect, and their reliability is inconsistent and hard to measure.[50] Errors include mislabeling legitimate content—such as journalism—as AI-generated, while simultaneously failing to detect actual AI-generated content. The reliance on platforms to determine and apply such labels introduces inconsistencies and raises concerns about how information is curated and presented to the public.

One major risk is the "implied truth" effect, where people may begin to distrust all content, regardless of its source or authenticity, simply because of the presence or absence of an AI-generated label. This growing skepticism could lead to the widespread dismissal of credible information, undermining the public's ability to access trustworthy and accurate news. Such a scenario also compromises the right to freedom of information, as the lines between reliable sources and misinformation blur.[51]

Nevertheless, if watermarking systems can achieve higher levels of accuracy and transparency, they could potentially enhance freedom of information. A robust mechanism to label trustworthy or newsworthy content could help restore public confidence in digital media and ensure access to reliable information. However, until such systems are refined, watermarking may create confusion and limit the accessibility of credible content, thereby undermining the foundational principles of freedom of information.

[50] Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., & Goldstein, T. (2024). On the reliability of watermarks for large language models. arXiv. https://arxiv.org/abs/2306.04634

[51] Consultation with eliska pirkova

Right to Freedom of Opinion

Legal basis

The right to freedom of opinion is explicitly protected under Article 19 ICCPR. The UN Human Rights Committee stated that unlike freedom of expression, which can be subject to certain restrictions, freedom of opinion is absolute—no one may be penalised or subjected to interference for holding an opinion. This right ensures that individuals can form, hold, and change opinions without coercion, surveillance, or undue influence.[52]

The Human Rights Council Resolution on new and emerging digital technologies and human rights recognised that AI systems, including when used to support content moderation, "can entail serious risks to the protection, promotion and enjoyment of human rights, such as [...] freedom of opinion [...] in particular by embedding and exacerbating bias which potentially result in discrimination and inequality, and by intensifying threats from misinformation, disinformation and hate speech, which may lead to violence, including political violence [...]."[53] Safeguarding this right thus requires protecting individuals from manipulative information environments, algorithmic bias, and undue governmental or corporate interference that may shape or suppress the right to freedom of opinion.

In the EU, the AI Act prohibits a few practices that could be enabled by LLM moderation.[54] First, under Article 5(1)(a), the AI Act prohibits the harmful manipulation and deception, defined as "AI systems that deploy subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective or with the effect of distorting behaviour, causing or reasonably likely to cause significant harm."[55] Second, it prohibits under Article 5(1)(b) the harmful exploitation of vulnerabilities, i.e. "AI systems that exploit vulnerabilities due to age, disability or a specific social or economic situation, with the objective or with the effect of distorting behaviour, causing or reasonably likely to cause significant harm."[56]

The Council of Europe Convention on AI establishes that "Each Party shall adopt or maintain measures that seek to protect its democratic processes in the context of activities within the lifecycle of artificial intelligence systems, including individuals' fair access to and participation in public debate, as well as their ability to freely form opinions." (Article 5.2).[57]

[52] United Nations Human Rights Committee. (2011, September 12). General Comment No. 34 on Article 19: Freedoms of opinion and expression (CCPR/C/GC/34). United Nations. Retrieved from https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

[53] United Nations Human Rights Council. (2023). Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development (A/HRC/53/L.27/REV.1). https://docs.un.org/en/A/HRC/53/L.27/REV.1

[54] European Union. (2025). Article 5: Prohibited AI Practices. EU Artificial Intelligence Act. Retrieved from https://artificialintelligenceact.eu/article/5/

[55] Ibid.,

[56] European Union. (2025). Article 5: Prohibited AI Practices. EU Artificial Intelligence Act. Retrieved from https://artificialintelligenceact.eu/article/5/;
European Commission. (February 2025). Commission publishes guidelines on prohibited artificial intelligence (AI) practices as defined in the AI Act. Retrieved from https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act

[57] Council of Europe. (2024). Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe Treaty Series - No. 225). Retrieved from https://rm.coe.int/1680afae3c

Amplification

In their 2022 paper, Twitter researchers noted that "In summary, false positive bias on social media is a type of representational harm, where both content concerning marginalised communities (in the case of counterspeech or identity terms) or produced by marginalised communities (in the case of dialect bias or reclaimed speech) receives less amplification than other content."[58]

Indeed, LLMs carry the risk of amplifying already powerful voices, including adversarial actors or authoritarian leaders. While this risk already exists with machine learning systems, LLMs risk exacerbating and accelerating amplification. For example, adversarial actors are increasingly capable of overwhelming online platforms by leveraging generative AI tools as content creation aids, or LLM-based recommender systems could amplify some narratives at the expense of others. This dynamic would demote or suppress marginalised perspectives, potentially facilitating the spread of propaganda or state-driven narratives. When such voices dominate information ecosystems, they restrict users' access to alternative viewpoints, undermining their ability to critically evaluate and shape their own opinions. This not only skews public discourse but also infringes on the foundational right to freedom of opinion, which relies on the availability of diverse and unbiased information.

Moreover, "popularity bias" in recommender systems, where highly popular content overshadows less popular but potentially relevant content, poses challenges to the right to freedom of opinion. While Meta's human rights team claims that LLM-based recommenders demonstrate reduced popularity bias compared to traditional systems,[59] this improvement does not fully address the broader risks. By limiting exposure to diverse perspectives and amplifying dominant narratives at scale, such systems may constrain users' ability to access the full spectrum of information necessary for forming independent opinions.

[58] Yee, K., Redfield, O., Sheng, E., Eck, M., Schoenauer Sebag, A., & Belli, L. (2022). A keyword-based approach to understanding the overpenalization of marginalized groups by English marginal abuse models on Twitter (p. 3). arXiv. Retrieved from https://arxiv.org/pdf/2210.06351

[59] Shirazyan, S., & Sissons, M. (2024, August 2). *AI's potential to advance human rights? Striking the Right Balance*. *Just Security*. Retrieved from https://www.justsecurity.org/98097/ais-potential-to-advance-human-rights/

Polarisation

Generative AI and LLM moderation can amplify smear campaigns, disproportionately impacting activists and civil society groups. These tools allow for the rapid creation of content intended to discredit individuals or movements, exacerbating the reach, scale, and severity of harmful narratives. Detection technology for large-scale trolling operations, which often relies on behavioral rather than textual analysis, struggles to keep pace with such sophisticated AI-generated content.[60] This allows propaganda, including deepfakes, to spread unchecked on social media, creating confusion undermining trust in information.[61] As a result, people's right to freedom of opinion can be negatively impacted.

While there has been much debate on how traditional or machine learning search systems contribute to echo chambers and selective exposure, less is known about the specific risks posed by LLM-powered conversational systems. Recent research indicates that users interacting with LLMs often engage in biased information querying, with opinionated systems reinforcing pre-existing views.[62] This dynamic exacerbates the risks of polarisation and further entrenches individuals within ideological bubbles.

Political bias embedded in LLMs compounds these risks.[63] A 2023 study revealed that left-leaning users are more likely to receive favorable content about left-leaning figures and media outlets, while right-leaning users encounter similarly biased content aligned with their views. The authors found that personalisation mirrors the risks seen in traditional algorithmic systems, where demographic tailoring intensifies affective polarisation and can reinforce filter bubbles. [64]

[60] Ezzeddine, F., Ayoub, O., Giordano, S., Nogara, G., Sbeity, I., Ferrara, E., & Luceri, L. (2023, October 9). *Exposing influence campaigns in the age of LLMs: A behavioral-based AI approach to detecting state-sponsored trolls. EPJ Data Science*. Retrieved from https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-023-00423-4.

[61] Swenson, A., & Chan, K. (2024, March 14). *Election disinformation takes a big leap with AI being used to deceive worldwide. AP News*. Retrieved from https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd.

[62] Sharma, N., Liao, Q. V., & Xiao, Z. (2024, May 11). *Generative echo chamber? Effect of LLM-powered search systems on diverse information seeking. Proceedings of the ACM*. Retrieved from https://dl.acm.org/doi/10.1145/3613904.3642459.

[63] Bang, Y., Chen, D., Lee, N., & Fung, P. (2024, March 27). *Measuring political bias in large language models: What is said and how it is said. arXiv*. Retrieved from https://arxiv.org/html/2403.18932v1.

[64] Lazovich, T. (2023, October 31). *Filter bubbles and affective polarization in user-personalized large language model outputs. arXiv*, p. 1. Retrieved from https://arxiv.org/pdf/2311.14677.

Algorithmic Gatekeepers: IV. Right to Freedom of Expression, Information and Opinion

European Center for
Not-for-Profit Law