# Algorithmic Gatekeepers:
## The Human Rights Impacts of LLM Content Moderation

## Executive Summary

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


democracy
justice
equality
expression

# Executive summary

## Perennial challenges in content moderation

Content moderation plays a pivotal role in shaping digital discourse, determining what content is visible, removed, or amplified across online platforms. It relies on a mix of policy, automated tools, and human moderators to manage both the binary task of removal ("leave up/take down" model) and the more complex work of content curation through recommender systems. While moderation aims to reduce harm and foster safer online spaces, it also introduces new risks, particularly for marginalised communities. This report critically examines the growing use of automation in content moderation, focusing on large language models (LLMs), and explores how they can be used more responsibly to reduce harm and challenge dominant narratives.

Current content moderation methods face well-documented challenges. Human moderators are vital for contextual understanding, but human review is costly, difficult to scale, and often psychologically harmful. Automated moderation addresses scalability to some extent, but falls short in transparency, interpretability, and fairness, among other issues. Moderation at scale remains a seemingly impossible task, requiring deep policy expertise, diverse perspectives, iteration time, and significant computational resources. Machine learning-based content moderation systems are typically trained to predict what a human moderator would do, without the capacity to address contextual and nuanced content.[1]

## The promises and perils of LLM moderation

LLMs—powerful generative AI systems trained on vast textual datasets—are increasingly used in content moderation for tasks such as automated removal, flagging, behaviour detection, user notification, real-time and customised moderation, and recommender systems. While they offer scalability and adaptability, they also replicate many of the risks seen in earlier machine learning systems—exacerbating issues like systemic discrimination, censorship, and surveillance. Importantly, these systems often over- or under enforce content policies, disproportionately impacting marginalised groups. Legitimate speech may be silenced, while harmful content slips through, exposing already at-risk communities to hate and violence.

Despite their strengths in pattern recognition, LLMs lack true contextual understanding. They misinterpret nuance, such as irony, satire, or reclaimed slurs, leading to inconsistent or harmful moderation, especially in culturally sensitive contexts and in non-English languages. Their tendency to "hallucinate"—producing confident but incorrect outputs—raises concerns about their reliability in high-stakes decisions. These flaws can erode public trust and result in arbitrary or unjust censorship without proper safeguards.

1 Cyber Policy Center. (2024, April 23). The future of content moderation and its implications for governance [Video]. YouTube. https://www.youtube.com/watch?v=JMq49FZ5qmY

Like machine learning-based systems, LLMs often learn from biased training data, rooted in imperialist and colonial dynamics—and as a result, they can replicate and amplify already existing biases at scale. Research shows they're more likely to flag content from minority communities or activist groups unfairly. Unlike human moderators, who can be trained and held accountable for their decisions, LLMs make these decisions invisibly, without oversight. The problem is made worse by weak evaluation methods that don't fully account for differences across languages, cultures, or demographics.

## Concentration of power and "algorithmic monoculture"

One of the most pressing issues underlying these concerns is the concentration of power in content moderation, as a small number of foundational LLMs can dictate global online speech norms—creating a form of "algorithmic monoculture." Since most platforms fine-tune these foundational models rather than developing their own, decisions made at the training stage of LLMs cascade down across multiple platforms, shaping how content is moderated online. This centralisation of content moderation risks reinforcing systemic biases and ideological homogeneity, limiting diversity in the content we see online and stifling alternative views. For example, Palestine-related content has been disproportionately designated as terrorist content online.[2] Unless LLM deployers fully address the specific bias through fine-tuning, it propagates across every platform that relies on the model, leading to widespread silencing of pro-Palestine users. This dynamic exacerbates already systemic discrimination while simultaneously eroding pluralism in online spaces and reinforcing centralised control over speech.

## Rethinking innovation and labour

Overall, we should seriously question the premise that using LLMs will improve the working conditions of human moderators. First, underneath the promise of using LLMs for content moderation lies the not-so-hidden truth that the goal is ultimately to replace human moderators, with the end goal to cut costs for platforms. Second, if platforms truly prioritised moderators' well-being, they'd increase resources in hiring more moderators, improve their working conditions (including increased pay and benefits) and actively include them in designing systems that are meant to support them. Instead, journalists and researchers have exposed the harsh working conditions and low wages faced by workers—predominantly from the Global Majority—who label and prepare training data for AI systems, including automated content moderation.[3] Perhaps most importantly, we should resist the continuous urge to "innovate" at all costs, focusing on baseless promises over documented risks and harms.

---

2 Oversight Board. (2024, March 26). Oversight Board publishes policy advisory opinion on referring to designated dangerous individuals as "Shaheed". Retrieved from 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. European Center for Not-for-Profit Law & Electronic Frontier Foundation. Retrieved from https://ecnl.org/sites/default/files/2023-05/ECNL%20EFF%20Submission%20to%20Policy%20Advisory%20Opinion%202023.pdf.

3 Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. TIME. Retrieved from https://time.com/6247678/openai-chatgpt-kenya-workers/.

Given the above, meaningful improvements in content moderation require addressing platforms' underlying business models—known as "surveillance capitalism"—rather than relying solely on new, shiny AI systems like LLMs. That said, LLMs could potentially improve content moderation if developers and deployers put people at the center, enhancing their work through thoughtful, creative technological support. For instance, instead of using LLMs merely to replace human moderators or remove content, platforms could use LLMs to assist moderators by triaging simpler cases, summarising emerging trends, or offering second opinions, while leaving final decisions to people.

This approach should be accompanied by greater investment in the well-being and training of human moderators, ensuring that any AI deployment supports rather than undermines their role. A rights-respecting model would focus innovation on tools that empower both moderators and users—such as intuitive interfaces that let users customise their content preferences, or LLMs that help clarify platform rules—rather than using LLMs as a cost-cutting measure to eliminate human judgment. Transparency tools powered by LLMs could also notify users in real time when a post is flagged and explain how to appeal, improving both user experience and the remedy process. These kinds of human-centered, more accountable uses of LLMs stand in stark contrast to fully automating content removal—and they deserve deeper exploration.

# The human rights impacts of LLM moderation

**Privacy:** Using LLMs for content moderation—especially personalised moderation—raises concerns about consent and data collection. These models are often trained on indiscriminately scraped data and can infer sensitive information about users without their knowledge. This can enable forms of mass surveillance and profiling, particularly targeting marginalised groups and civil society actors, under the pretence of moderation.

**Freedom of expression:** LLMs could potentially help scale and improve moderation in under-resourced languages and offer more nuanced responses (e.g., labelling or downranking instead of takedowns). However, they frequently over- or under-enforce policies in practice. This means LLMs may wrongfully remove or restrict legitimate speech—often disproportionately affecting marginalised voices—or fail to flag harmful content, leaving users vulnerable to hate speech or incitement to violence. Because LLMs still struggle with context, especially in less-represented languages and dialects, and are prone to generating false outputs ("hallucinations"), they shouldn't be relied on when deciding what speech is permissible or not. These issues create an environment ripe for censorship and chilling effect: authorities or platforms could exploit LLM systems to quietly censor dissenting opinions or controversial discussion at scale, deterring users from freely expressing themselves.

**Freedom of information:** LLMs could enhance information access through translation, summarising, and personalised content delivery, particularly in multilingual contexts. But they are unreliable at detecting misinformation and often oversimplify or hallucinate facts. In crises or politically sensitive situations, they may suppress crucial public interest information. Moreover, governments can exploit these systems to limit access to information or flood platforms with propaganda—harms that often disproportionately affect journalists, activists, dissidents, and marginalised communities.

**Freedom of opinion:** When designed responsibly, LLMs could potentially broaden users' exposure to diverse perspectives and reduce the reach of harmful or manipulative content. However, without safeguards, LLM personalisation and biased filtering can reinforce dominant narratives and limit ideological diversity at scale. This risks subtly shaping users' views, particularly in authoritarian or politically tense environments, and undermines individuals' ability to form independent opinions.

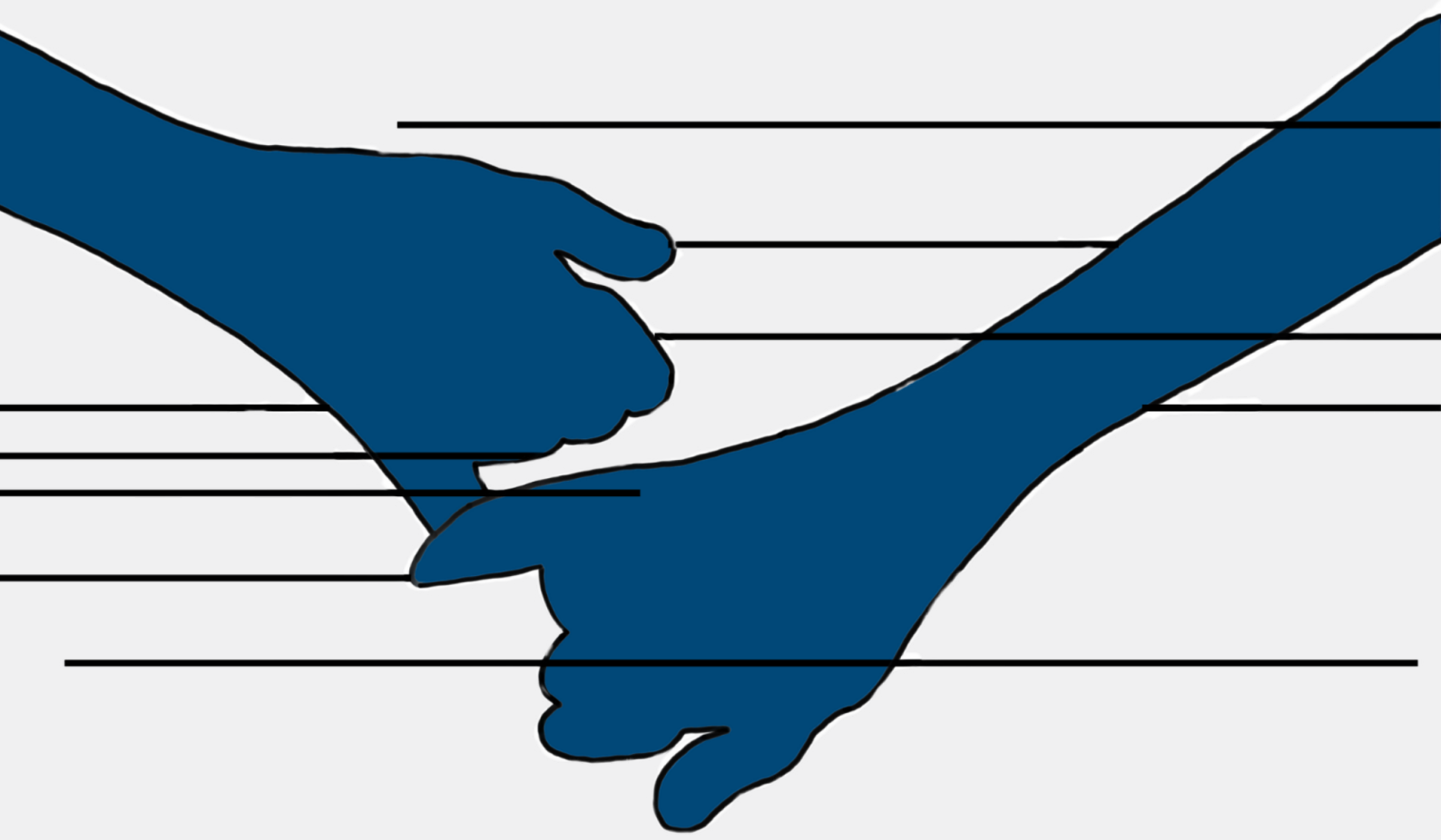
**Freedom of peaceful assembly:** LLMs may potentially help identify disinformation or threats of violence related to protests, improving safety around peaceful gatherings. But they can also misclassify protest-related content as extremist or violent due to poor contextual understanding and lack of representative data related to protests and contrarian views—especially in politically charged or multilingual settings. This can result in the suppression of organising efforts and documentation by activists, civil society, and human rights defenders, threatening democratic participation and peaceful assembly.

**Freedom of association:** LLMs could possibly improve safety for online organising by better detecting harassment or doxxing campaigns against activists. Yet, they may also mislabel content from advocacy groups, grassroots movements, or civil society organisations as harmful. Additionally, governments may weaponise LLM-based systems to surveil or target associations under the guise of moderation—either directly or by pressuring platforms to suppress content related to opposition groups.

Non-discrimination: When developed with inclusive datasets and stakeholder input, LLMs could potentially help detect and mitigate discriminatory content more consistently. However, most models are trained on data rooted in colonial and imperialist dynamics, leading to discriminatory outcomes, especially for communities in the Global Majority and marginalised groups. Debiasing efforts have shown limited effectiveness, and significant performance gaps persist between dominant (colonial) and underrepresented languages. Cultural nuance is often lost, and poor benchmarking prevents platforms from adequately identifying and addressing discriminatory impacts.

**Participation:** LLMs have the potential to support participatory design through tools such as customisable moderation AI agents. However, in practice, affected communities are largely excluded from shaping content moderation systems. Even methods like reinforcement learning from human feedback (RLHF) can embed biases when the annotators come from homogenous backgrounds and lack the diversity of communities' lived experiences, languages, cultures, and needs.

**Remedy:** LLMs could potentially enhance remedy mechanisms by generating clearer explanations for moderation decisions, speeding up appeals, and identifying systemic errors. But their lack of explainability and transparency creates barriers to remedy, too. Even LLM developers or deployers may not understand why a model made a specific content decision—let alone explain it to users. Furthermore, when platforms rely on third-party LLMs, accountability becomes fragmented, leaving users with limited recourse to contest or appeal moderation outcomes.



# From the ground up: community-led initiatives

Alternative approaches are emerging and can potentially offer a more inclusive path forward. Community-led initiatives in the Global Majority, such as Masakhane, Lelapa AI, and InkulaLM, focus on public-interest-driven AI development, culturally informed moderation, and decentralisation. These models—though smaller in scale—demonstrate comparable performance in tasks like translation and sentiment analysis, highlighting the potential for more rights-based, participatory LLM development that does not rely on monopolistic providers.

What's more, as new players emerge—such as CoPE (the COntent Policy Evaluator)[4] a small, special-purpose language model for content labelling, the Chinese LLM DeepSeek, Cohere's Aya, and faster inference-focused AI models like Groq—there is an opportunity to rethink the dependence on a handful of dominant AI firms. Community-driven natural language processing (NLP) projects, particularly from the Global Majority, further challenge the Silicon Valley AI model that prioritises data extraction, centralised control, and profit over rights and public interest. These initiatives illustrate that a decentralised, culturally aware approach to LLM development is both possible and necessary.

4 Zentropi AI. (2024, December 19). COPE-A 9B. Hugging Face. Retrieved from https://hugging-face.co/zentropi-ai/cope-a-9b.

# Preliminary recommendations and next steps

While perspectives may differ, this report primarily examines the risks and harms of LLM moderation and outlines emerging best practices for policymakers, LLM developers,and deployers. Looking ahead, we aim to refine these recommendations to make them more actionable, and expand guidance for policymakers working to regulate these systems. By centering human rights, we stress the need to focus on documented harms—not industry hype. As regulation is considered, it's critical to confront the systemic biases, centralisation, and lack of transparency in current LLM moderation.

This report contributes to a rights-based vision for AI content governance—one that protects civic freedoms and supports diverse, safe, and inclusive online spaces. Crucially, this work is just beginning. Ongoing multi stakeholder research, policy development, and participatory evaluation are needed to better understand and address the human rights impacts of LLM moderation and co-create more just and democratic models of AI content governance.

## 1. To policymakers

* No mandatory AI or LLM moderation: Avoid requiring or incentivising AI-driven moderation without human oversight. Automated systems should not replace human judgment, especially where rights are at stake.
* Mandate human oversight: Legally require platforms to integrate human-in-the-loop systems and provide accessible appeal mechanisms for LLM decisions. Oversight should include documentation, audits, and accountability measures.
* Enforce transparency and accountability: Mandate disclosure of how LLM systems function and how they're used, including accuracy rates and appeals enforcement. Require platforms to notify users about LLM moderation actions and enable appeals mechanisms.
* Mandate human rights impact assessments (HRIAs): Make HRIAs mandatory for LLM developers and platforms before deployment and throughout the LLM lifecycle. HRIAs should focus on marginalised groups, include external stakeholders, and be subject to public review.
* Meaningfully engage with external stakeholders: Establish mechanisms for civil society and affected communities to participate in AI policy. Require companies to engage external stakeholders in the development and use of LLM moderation.

## 2. To LLM developers and deployers

The report emphasises that both the developers of LLMs and the platforms that deploy them have a responsibility to respect human rights. Developers (often large AI firms) provide the foundation models: they decide how the model is trained, what data is used, and what safeguards are built in. Deployers (e.g. social media companies) then adapt the model for their internal needs and integrate it into their content moderation workflow. Governance and rights-based practices must address both sides.

### LLM developers

* Develop clear, human rights-based AUPs: LLM developers should create and enforce Acceptable Use Policies (AUPs) that include strong human rights commitments, define appropriate use of automation, require human oversight, enable appeals, and ensure transparency.
* Ensure accountability and continuous improvement: AUPs must include protections against bias and human rights risks, especially for marginalised groups, use measurable outcomes, be regularly updated, and be enforced across jurisdictions.

### LLM deployers

* Evaluate and test models thoroughly: Review vendor AUPs, assess model performance across diverse content and demographics, and conduct scenario testing with edge cases and culturally specific content.
* Ensure transparency and accountability: Request disclosure on training data and biases; negotiate contracts with performance standards, audit rights, and clear liability for harms.
* Build resilient moderation systems: Use multiple LLMs to avoid being locked into one model or developer, implement fallback procedures, and maintain scalable human moderation capacity.

### Shared responsibilities (developers + deployers)

Minimise data collection and embed privacy-by-design into product development and use.

* Ensure human oversight.
* Conduct meaningful human rights due diligence, including impact assessments.
* Carry out red-teaming and adversarial testing.
* Meaningfully engage external stakeholders in LLM design, development, and use.
* Ensure adequate language and cultural context.
* Refine LLMs through continuous feedback and retraining.
* Evaluate LLMs rigorously.
* Implement internal grievances mechanisms.
* Ensure meaningful transparency.

# Algorithmic Gatekeepers - Executive Summary

European Center for
Not-for-Profit Law