Algorithmic Gatekeepers:
The Human Rights Impacts of LLM Content Moderation

V. Right to Freedom of Peaceful Assembly and Association

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.

Freedom of Peaceful Assembly

Legal basis

Under Article 21 ICCPR, "The right of peaceful assembly shall be recognised. No restrictions may be placed on the exercise of this right other than those imposed in conformity with the law and which are necessary in a democratic society in the interests of national security or public safety, public order, the protection of public health or morals or the protection of the rights and freedoms of others." [1]

Freedom of peaceful assembly encompasses the ability to organise and participate in peaceful gatherings, including meetings, sit-ins, strikes, rallies, events, or protests. Everyone has the right of peaceful assembly: citizens and non-citizens alike.[2] This right is typically exercised to express opinions, demonstrate, or advocate for specific causes or issues and is focused on temporary gatherings or events. It protects the act of gathering itself in solidarity with others, regardless of how or what content is expressed during the assembly.[3] Crucially, freedom of peaceful assembly applies to both offline and online spaces (or a combination thereof),[4] reflecting the growing role of digital platforms in modern civic engagement.[5] Examples include a protest march, a peaceful rally, or a virtual meeting organised to discuss a particular issue.

Over- and under enforcement of content policies

The right to peaceful assembly is impacted by LLM moderation in two critical ways: first, online assemblies can be suppressed and second, barriers to organising offline protests can emerge. Online spaces play a crucial role in facilitating assemblies, allowing individuals to hold meetings, organise protests, and share strategies for advocacy. If LLMs' accuracy improves, LLM moderation may positively impact peaceful assembly by ensuring better access to information and fostering safe and open communication for organising non-violent protests. However, as demonstrated throughout this report, limited accuracy can result in the suppression of legitimate organising discussions, highlighting the ongoing challenge of balancing precision with freedom of expression.

For example, in contexts where individuals mobilise against oppressive regimes—especially in low-resource languages—they may use emotionally charged phrases such

1 United Nations Human Rights Office. (n.d.). International Covenant on Civil and Political Rights. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.
2 United Nations Human Rights Committee. (2020). General comment No. 37: Article 21 (Right of peaceful assembly), para. 5. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-37-article-21-right-peaceful
3 Ibid., para. 1.
4 Ibid., para. 6.
5 European Center for Not-for-Profit Law. (2022). A guide to digitally mediated assemblies and how to monitor them. Retrieved from https://ecnl.org/handbook/guide-digitally-mediated-assemblies-and-how-monitor-them#Generalprovisionsregardingtherighttopeacefulassembly.

as "destroy them" or "annihilate the enemy" to voice dissent. While these expressions may not constitute actual harm, AI classifiers might interpret them as violent, leading to the automatic removal of such content. [6]

Furthermore, even if a single member of a group posts content that may constitute incitement to violence and violate platform policies and/or Article 20 ICCPR, this does not justify suppressing the content of the entire group.[7] For example, platforms such as Facebook were criticised for taking down information about a protest as soon as there were any messages calling for violence, arguably suppressing entire conservative political movements.[8] This misclassification thwarts freedom of assembly, silences political organising, and hinders democratic movements.

These challenges are exacerbated when moderating content with multilingual and culturally diverse contexts. LLMs, partly due to biased pretraining data, often lack the linguistic and cultural nuance needed to accurately moderate content related to assembly. These biases are often political, religious, and language-dependent, leading to discrepancies in how different contexts are handled. LLM moderation can thus result in the disproportionate removal of online protests, further marginalising already underrepresented groups.

For example, the suppression of activists advocating for Palestinian rights[9] illustrates how the right to protest online can be stifled, limiting the ability to assemble, share grievances, or build collective movements. Moreover, moderating discussions related to Palestine differs whether the content is in Arabic versus Hebrew.[10] The same applies to conversations around Taiwan, with different moderation outcomes whether the content is in Mandarin versus in English.

Finally, there is also a risk that LLM moderation fails to remove non-peaceful or violent assembly due to mislabeling or insufficient oversight. Platforms have often been criticised for leaving up messages inciting violence during protests, with hindsight revealing these as mistakes (i.e. false negatives). Meta's infamous involvement in enabling ethnic cleansing against the Rohingya in Myanmar is just one of many similar cases.[11]

---

6 Maung Maung, B. (2023). When conflict goes online: How trust & safety systems fall short in handling crises in the global majority. Tech Global Institute. Retrieved from https://techglobalinstitute.com/research/when-conflict-goes-online-how-trust-safety-systems-fall-short-in-handling-crises-in-the-global-majority/.

7 United Nations Human Rights Committee. (2020). General comment No. 37: Article 21 (Right of peaceful assembly), para. 50. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-37-article-21-right-peaceful(applicable to online assemblies as well):
"In accordance with article 20 of the Covenant, peaceful assemblies may not be used for propaganda for war (art. 20 (1)), or for advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence (art. 20 (2)). As far as possible, action should be taken in such cases against the individual perpetrators, rather than against the assembly as a whole. Participation in assemblies whose dominant message falls within the scope of article 20 must be addressed in conformity with the requirements for restrictions set out in articles 19 and 21."

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

9 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.

10 Kawash, A. (2024, February). AI and racism. The Arab Center for the Advancement of Social Media. Retrieved from https://7amleh.org/storage/AI%20&%20Racism/7amleh%20-AI%20english1-1.pdf.

11 Amnesty International. (2023, August 25). Myanmar: Time for Meta to pay reparations to Rohingya for role in ethnic cleansing. Amnesty International. https://www.amnesty.org/en/latest/news/2023/08/myanmar-time-for-meta-to-pay-reparations-to-rohingya-for-role-in-ethnic-cleansing/

Counter-speech, minority views, and exceptional events

Automated content moderation systems typically rely on statistical patterns, prioritising majoritarian views as the norm while treating everything else as edge cases. This approach disproportionately affects protests, which are inherently contrarian and challenge dominant power structures. Protesters and their relevant content are at greater risk of being discarded in datasets, particularly regarding fragile democracies or conflict zones, where protests are often organised in hiding and/or do not conform to typical statistical patterns. This could be potentially mitigated when fine-tuning the LLM, since labelled examples could be used from the deployers' own platforms and datasets.

The challenges posed by LLM moderation are further exacerbated by static training data that fails to adapt to evolving social norms. As noted by Bender, Gebru, McMillan-Major and Mitchell in their seminal paper "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," social movements use language strategically to disrupt dominant narratives and highlight underrepresented perspectives.[12] However, LLMs risk leading to "value-lock," perpetuating outdated or exclusionary understandings.[13] Poorly documented movements or those ignored by media often go unrepresented in training data, leading to a disproportionate alignment with existing power regimes.[14] Media coverage, which tends to amplify dramatic or violent events over peaceful protests, further skews these models, misrepresenting social movements and suppressing critical voices.[15]

In conflict-affected areas, these risks are heightened, and content moderation errors can silence life-dependent online assemblies. As such, they hinder affected communities' ability to organise, mobilize, and sustain movements, ultimately limiting democratic progress. According to the UNGPs, businesses operating in such contexts bear an enhanced responsibility to respect human rights and international humanitarian law.[16] The likelihood and severity of online-to-offline harm are significantly greater in these areas, necessitating stricter thresholds for addressing hate speech, incitement to violence, and other conflict drivers.[17] However, this often leads to increased over-enforcement, which can be problematic during times of crisis when access to information is especially critical. In any case, balancing these considerations presents a trade-off.[18] Businesses must exercise heightened due diligence to avoid complicity in human rights abuses and ensure that content moderation systems do not exacerbate harm or suppress vital dissenting voices.[19]

12 Bender, E. M., McMillan-Major, A., Gebru, T., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, p. 614. https://doi.org/10.1145/3442188.3445922
13 Ibid.
14 Ibid.
15 Bender, E. M., McMillan-Major, A., Gebru, T., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, p. 614. Retrieved from https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.
16 United Nations. (2011). Guiding principles on business and human rights: Implementing the United Nations 'Protect, Respect and Remedy' framework. United Nations Human Rights Office. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf
17 BSR. (2021). A human rights-based approach to content governance. (p.6) Business for Social Responsibility. https://www.bsr.org/reports/A_Human_Rights-Based_Approach_to_Content_Governance.pdf
18 Ibid.
19 Allison-Hope, D., Andersen, L., & Lovatt, J. (2021, March). A human rights-based approach to content governance, p. 6. BSR. Retrieved from https://www.bsr.org/reports/A_Human_Rights-Based_Approach_to_Content_Governance.pdf.

Government surveillance and suppression of collective organising

While social media monitoring for identifying "persons of interest" or potential threats to government authority is not new, the integration of LLMs could amplify its scale and effectiveness, raising critical questions about the impact to civic space. Government agencies, especially law enforcement, could weaponise these tools to scan social media content with the goal to identify and suppress collective organising, including protests, and/or limit the visibility of online assemblies and organising efforts.[20]

This surveillance capability can have a chilling effect on freedom of assembly, deterring users from organising or participating in online activism, or to join existing organising groups. Many may fear being profiled as activists or protesters by platforms using LLMs, with the resulting information potentially shared with law enforcement, border control, or immigration authorities. This risk is especially acute for migrants and refugees, who might face heightened scrutiny or deportation due to their activism and organising.

Restricting visibility through recommender systems

While LLM-based recommendation systems may provide more in-depth and comprehensive recommendations by analyzing both content and user behaviour, they also have significant limitations. Instead of directly removing protest content, suppression may occur by not recommending, demoting or even "shadow banning" activist content (i.e. making it de facto invisible). LLM-based recommender systems may exacerbate these risks, including by misclassifying content related to assemblies as violent or inciting offline violence.

For instance, LLMs typically favour mainstream content while side lining non-traditional or undervalued perspectives due to biased training data.[21] This could severely impact marginalised groups' collective organising efforts, as their content may be unjustly demoted. These dynamics risk limiting visibility for critical or dissenting voices, including protest-related content.

Furthermore, questions remain about whether LLM-based systems can accurately summarise or interpret exchanges (e.g., through sentiment analysis) to assess whether online organising is considered violent or peaceful. Such features could further shape how content is recommended, suppressed, or framed, potentially amplifying biases or redefining user engagement with certain topics.

---

20 Dyson, I., Milner, Y., & Griffiths, H. (2024, April 30). Documents reveal how DC police surveil social media profiles and protest. Brennan Center for Justice. Retrieved from https://www.brennancenter.org/our-work/analysis-opinion/documents-reveal-how-dc-police-surveil-social-media-profiles-and-protest#:~:text=DC%20police%20also%20rely%20on,%23RefuseFacism%2C%20and%20%23Anticapitalist.

21 Anonymous. (2024, June 16; modified July 2, 2024). How trustworthy is AI? A deep dive into the bias in LLM-based recommendations. OpenReview. Retrieved from https://openreview.net/forum?id=rrzw1t7LHc.

Concentration of power

A small number of companies control how LLMs moderate content, deciding what counts as peaceful assembly online and which groups are allowed to organise on their platforms, shaping the rules of online assembly. Yet developing LLMs requires massive resources, such as large datasets and powerful computing systems, making it difficult for new companies to compete.[22]

Decisions about acceptable online organising can have a strong influence over public opinion and politics. Social media platforms have historically been leveraged by affected communities to challenge regimes and powerful entities. Online organisation efforts, from the Arab Spring to the #MeToo movement, were critical for offline assembly. The flip side of the coin is that platforms yield incredible power in deciding which entities or governments can be challenged and what content related to collective organising, including protests, is allowed. Recent changes to platform policies, such as Meta's actions a few days before Trump was sworn into presidency,[23] highlight how platforms can be influenced by powerful governments and may readily align with their interests when pressured. This dynamic extends beyond cultural debates, as platforms are key battlegrounds for political struggles and conflicts.[24]

Moreover, LLM's decisions on acceptable online assembly impact not only their own platforms but also smaller platforms, which rely on foundation models for moderating content on their platforms (see above section on freedom of expression and information). As such, internal content moderation policies of an LLM significantly influence its ability to moderate externally. For example, if an LLM such as ChatGPT determines that specific content related to organising or assembly is unacceptable, this decision will cascade to the moderation practices of platforms that deploy ChatGPT for their internal content moderation, effectively standardising how users can organise online across platforms and narrowing the diversity of acceptable discourse around assembly.

As seen in previous sections, deployers can mitigate this risk by fine-tuning the foundation model in accordance with their own content policies. However, any elements of the model that are not fine-tuned will, by default, reflect moderation decisions made at the foundation level, potentially leading to outcomes that do not align with the deployer's specific guidelines or enforcement choices.

22 Google Cloud Skills Boost. (n.d.). Video: 3:05. https://www.cloudskillsboost.google/course_templates/539/video/466324

23 Kaplan, J. (2025, January 7). More speech and fewer mistakes. Meta. Retrieved from https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/.

24 Narayanan, A. (2024). Snake Oil: Why Can't AI Fix Social Media? Princeton University Press, p. 211.

Freedom of Association

Legal basis

The right to freedom of association is protected under Article 22 ICCPR[25] and "involves the right of individuals to interact and organise among themselves to collectively express, promote, pursue and defend common interests."[26]

Freedom of association protects the right to form, join, and maintain organisations, both online and offline, for lawful purposes. It enables individuals to organise collectively for long-term goals—political, economic, cultural, or social—and encompasses both formal and informal groups, such as trade unions, political parties, or advocacy organisations.

Closely tied to freedom of expression, freedom of association goes beyond collective speech by emphasising the ability to collaborate and create stable organisations that are essential for civic participation and advocacy. Unlike the spontaneous nature of assemblies, associations are more enduring and structured, providing a foundation for sustained action and societal impact.

Undue take downs for violent extremist and terrorist content

When groups use online platforms to organise—whether for virtual meetings or to plan offline protests—LLM moderation may unintentionally disrupt their activities. Due to biased or overly broad criteria, these systems can misclassify peaceful content as harmful or inappropriate, limiting individuals' ability to form and sustain groups, a fundamental aspect of freedom of association.

Distinguishing between lawful and unlawful activities is inherently challenging. Violent or criminal acts are statistically rare, making them difficult for both humans and AI to predict accurately.[27] As noted by the former UN Special Rapporteur on Freedom of Expression, the line between permissible and impermissible activities is often blurry, further complicating moderation decisions.[28] This complexity makes it unlikely that

25 United Nations Human Rights Office. (n.d.). International Covenant on Civil and Political Rights. Retrieved from https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

26 United Nations Human Rights Office. (n.d.). Freedom of assembly and association. Retrieved from https://www.ohchr.org/en/topic/freedom-assembly-and-association.

27 Panduranga, H., Mella, E., & Pablo. (2022, February 9). Federal government social media surveillance explained. Brennan Center for Justice. Retrieved from https://www.brennancenter.org/our-work/research-reports/federal-government-social-media-surveillance-explained.

28 Recommender systems add another layer to this issue. These algorithms shape what content users see, which groups they are encouraged to join, and what communities they interact with. For marginalized groups, this can help build connections and solidarity. However, when these systems reinforce biases, they can trap users in echo chambers or silence dissenting voices.
The definition of non-peaceful activity adds another challenge. It is not a neutral or universally agreed concept but depends on political and cultural contexts. What one government calls non-peaceful dissent might be seen elsewhere as a legitimate protest. LLMs, trained on data shaped by dominant perspectives, often struggle with these nuances. A phrase that is harmless in one language could be flagged as threatening in another, putting already marginalized groups at greater risk of scrutiny and suppression. https://www.unodc.org/e4j/zh/terrorism/module-13/key-issues/freedom-of-association.html

LLMs can consistently recognise the nuances of legitimate associations.

Misclassifying online associations can have serious consequences. Legitimate groups may be unfairly removed or suppressed online, as seen in cases where counter-terrorism measures disproportionately impacted minority communities who were unjustly designated as violent extremist or terrorist.[29] For example, if English-language data broadly categorizes Antifa as a terrorist organisation, peaceful groups associated with the term could face unwarranted suppression. Similarly, a non-violent organisation might be flagged incorrectly due to a lack of cultural or linguistic context, directly infringing on their freedom of association. Legitimate content of Muslim and Arabic-speaking communities is often disproportionately removed as a result.

On the other hand, moderation systems also fail to detect genuinely harmful or violent associations due to mislabelling or insufficient oversight, enabling violence online and offline (see above sections on freedom of expression and peaceful assembly). Both overreach and under reach in content moderation disrupt the delicate balance between protecting legitimate associations and addressing unlawful activity.

Over enforcement of marginalised groups' content

Improved accuracy in LLMs could offer new opportunities for marginalised groups to organise and express themselves without fear of unjust moderation. Proponents claim that these systems might even outperform human moderators and manual content moderation, which is typically prone to errors and inconsistencies.

For groups such as LGBTQIA+, racialised persons including religious minorities, women and non-binary persons, migrants, political dissidents, activists, or journalists, among others, social media platforms offer opportunities for connection and advocacy, yet remains fraught with challenges of bias and suppression. These groups, as statistical minorities, are underrepresented in the training data of LLMs—systems rooted in colonial and imperialist dynamics as explained in the section on non-discrimination. As with traditional machine learning, this underrepresentation translates into higher risks of misclassification, leading to silencing their perspectives. As explained above, another issue with LLMs is their ability to learn problematic word associations that reflect biases against specific groups.

Recommender systems, both machine learning and LLM-driven, add another layer to this dynamic. These algorithms determine how users see content, what groups users are encouraged to join, and what communities they are exposed to. For marginalised groups, this could foster vital connections and facilitate solidarity-building. However, when these systems amplify existing biases, they risk isolating users in echo chambers or erasing dissenting voices entirely.

Furthermore, defining non-peaceful activity is challenging. Indeed, it's neither a neutral nor universally agreed concept and depends on political and cultural contexts. What one government calls non-peaceful dissent might be seen elsewhere as a legitimate protest or act of resistance. LLMs, trained on data shaped by dominant perspectives, often struggle with these nuances. A phrase that is harmless in one language could be flagged as threatening in another, putting already marginalised groups at greater risk of scrutiny and suppression.

29 European Center for Not-for-Profit Law. (2022, November). CT and tech: Mapping the impact of biometric surveillance and social media platforms on civic freedoms. Retrieved from https://ecnl.org/publications/ct-and-tech-mapping-impact-biometric-surveillance-and-social-media-platforms-civic.

As seen in above sections, centralised power among a small number of companies controlling LLMs exacerbates these risks. These companies increasingly define the parameters of legitimate associations and acceptable speech, wielding significant influence over who can organise and whose voices are heard. This consolidation raises profound concerns about accountability and the potential for these systems to reinforce existing power structures. How these systems are designed and governed will have far-reaching implications, not only for marginalised voices but for the health of democratic participation and civic discourse in the digital age.

Algorithmic Gatekeepers: V. Right to Freedom of Peaceful Assembly and Association

European Center for
Not-for-Profit Law