GLOBAL NETWORK INITIATIVE

Content Moderation, Human Rights Due Diligence, and Government Demands

NOVEMBER 2025

Table of Contents

Introduction

The Global Network Initiative (GNI) emphasizes shared learning as a core component of our mission. By tapping into the collective knowledge and practical experiences of our diverse members, GNI is uniquely positioned to address emerging challenges related to freedom of expression, privacy, and the tech sector.

As part of our shared learning activities, and in conjunction with our Human Rights Due Diligence Working Group, GNI held a **series of three calls** for our members from May to October 2025 on the **intersections of online content moderation, government restrictions and demands, possible impacts to freedom of expression and privacy, and related human rights due diligence practices.** This series was identified as critical in light of the amount of content being exchanged on social media, and the increasing pressure from governments to moderate that content.

The series featured calls on the following topics:

- **Community moderation practices:** a broad overview of various models of community moderation, possible vectors for government demands and restrictions that could impact the rights to freedom of expression and privacy within those models, and what related human rights due diligence could look like.

- **"Hash databases" and other types of signal sharing:** how hash databases are used for the moderation of terrorist content, child sexual abuse material, and other types of harmful content; the ways government demands and restrictions related to those databases could impact the rights to freedom of expression and privacy; and what related human rights due diligence does and could look like, including current practices and opportunities to strengthen transparency, accountability, and safeguards.

- **AI systems and automated content filtering:** the freedom of expression and privacy impacts of using AI in content moderation – such as in automated filtering and more novel forms of moderation – with a focus on how government regulations and interventions can incentivize and influence these products and services in ways that impact these rights.

GNI published three posts on our website synthesizing learnings from each of these calls. The following report collates lightly edited versions of those posts. We hope the learnings from these discussions can help practitioners across sectors better understand the possible human rights implications of these practices and improve efforts to respect human rights.

SECTION 1

Exploring Human Rights Due Diligence in Community Moderation Models

This post was published on GNI's website on August 25, 2025.

As social media platforms increasingly turn to community moderation to manage content at scale, governments around the world are stepping up efforts to regulate online content through various regulatory approaches, ranging from mandating risk assessments and audits to issuing legal demands that regulate how content is governed and made accessible to users. In this evolving and complex landscape, understanding the human rights implications of community-based models has become critical.

In May 2025, the Global Network Initiative (GNI) held the first in a new series of learning calls for our members focused on the intersection of content moderation, human rights due diligence, and the human rights impacts of government demands and restrictions. This series aims to deepen collective understanding of how different models of content governance can impact the rights to freedom of expression and privacy, particularly as governments around the world become more active in regulating online spaces.

The first call provided a foundational look at how human rights due diligence applies to the design and implementation of community moderation systems. Participants discussed the roles and responsibilities of various stakeholders, including platforms, governments, and civil society, in ensuring this due diligence. The discussion also highlighted potential government demands and restrictions that could affect rights to freedom of expression and privacy within community moderation models, and considered what human rights due diligence might look like to address these specific risks. This piece provides an overview of key takeaways from this first learning call.

Understanding community-based models

Unlike traditional platform-driven content moderation, where companies set and enforce rules centrally, community moderation distributes governance – to varying extents – to users themselves. This can take many forms: decentralized rulemaking, community voting systems, or hybrid models where platform infrastructure supports but does not control content decisions. In some cases, community input is used as a weight in the algorithmic decision-making process, while in others, human moderators can directly decide on what content is shared and how. Some communities are structured around volunteer-led governance, where policies are created and enforced from the ground up. Others operate on interest-based participation, where communities create their own rules and rely on both user feedback and voting systems to elevate or suppress content.

These models aim to support pluralistic dialogue, but they can also reflect dominant community norms and values, which may inadvertently silence minority voices. For example, moderation rules that rely heavily on written sources or formal expertise can exclude forms of knowledge not traditionally recognized by Western or academic institutions, raising concerns around epistemic injustice and cultural exclusion.

Recognizing this gap, Sarah Gilbert of Cornell University's Citizens and Technology Lab encouraged participants to view moderation through an intersectional lens, recognizing that moderation systems, even when designed to protect freedom of expression, can replicate power imbalances. Moderation can act both as a form of oppression and resistance, depending on how it's structured and who holds decision-making power. Applying human rights frameworks to community moderation requires not only centering freedom to *speak*, but also freedom from *harm*. This involves accounting for different levels of power: the systemic power of platforms, the social dynamics within communities, and the interpersonal relationships between users and moderation teams.

Human rights due diligence in community-led systems

A key theme of the discussion was the need to integrate human rights due diligence (HRDD) into content governance, including with respect to decentralized systems. While some platforms retain minimal involvement in content decisions, they still bear responsibility for supporting inclusive, rights-respecting environments, especially when their tools and structures shape how communities operate.

Due diligence in this context may look different from how it applies to platforms and products with more top-down moderation models. For example, due diligence could involve examining how best to equip community moderators with tools and training, ensure transparency in

decision-making, promote accessibility for underrepresented groups, and provide appropriate review mechanisms. There was also recognition that some models are evolving to be more inclusive by adopting intersectional approaches that account for differences in power and lived experience.

Government demands and the limits of decentralization

The discussion highlighted concerns about how government demands and legal restrictions may interact with community-led moderation. Even with regard to decentralized models, platforms' policies and technical systems can be influenced by government pressure through legal requests or regulatory frameworks. Participants raised questions about the potential challenges these demands pose to the autonomy and privacy of users involved in community moderation, and the importance of considering human rights frameworks when responding to such requests. Transparency and accountability were noted as important principles to guide platform responses, though specific approaches are still evolving.

Some moderation models aim to reduce centralized editorial control by using collaborative labeling systems that attach notes or context to content without removing or downranking it. Importantly, participants noted that current approaches to collaborative labeling systems treat legal demands for content related to labels in the same way as other user-generated content, without special exceptions or carve-outs. These systems depend on broad user participation and seek to reduce bias by requiring consensus across diverse perspectives, even if, in most cases, applying a note or label to content does not automatically lead to its removal, demotion, or monetization penalties. These tools are still in early stages, raising important questions about transparency, accountability, and scalability.

The role of AI

Participants highlighted growing challenges for community moderators in identifying and evaluating AI-generated content. Concerns were raised about how AI can crowd out human contributions, complicate sourcing and verification, and potentially demotivate volunteers. One GNI company noted its ongoing work to develop clearer policies and tooling, including an upcoming Human Rights Impact Assessment on AI. Meta noted that it deploys "AI info" labels on content that has been identified as generated or modified by AI, and that contributors in its Community Notes pilot can add context to misleading AI-generated posts.

The discussion also touched on the growing role of AI-generated content and its implications for moderation. As AI becomes more widespread, distinguishing between real content becomes

more complex, posing challenges for community moderators, who may be overwhelmed by the volume and nuance required to make accurate assessments.

Volunteers on platforms reliant on community moderation report mounting burdens due to the volume and complexity of AI-related content. Mistakenly flagging genuine human content as AI-generated (false positives) poses risks to trust and volunteer engagement, as well as freedom of expression and access to information.

Some community-moderated platforms serve as training data sources for AI models, contributing to significant AI-generated content that can place a heavy burden on volunteer moderators. This situation raises broader concerns about the costs and impact of AI on public interest technologies, which will require further discussion moving forward.

Some platforms now label AI-generated content or require users to disclose the use of AI tools. But there is ongoing debate about how such content should be treated: Should AI content be judged differently? Should it be removed, flagged, or simply contextualized? Some platforms require AI-generated content to be clearly labeled and linked to sources, with community members reviewing these labels before they're applied. The focus is on evaluating content based on its substance, not on whether it was created by AI, to avoid bias or false assumptions. When AI-generated content is detected, it is transparently flagged to inform users. Participants stressed the need for transparency, clarity, and consistency to build trust, both within communities and with the broader public.

Continuing the conversation

As community-based moderation becomes a more prominent model for governing online spaces, ensuring that these systems uphold human rights standards will be critical. Ongoing dialogue is needed to better understand how human rights due diligence can be effectively integrated into evolving models of community moderation, particularly amid growing regulatory demands and the expanding use of AI. Upcoming sessions in this series will continue to explore the challenges, responsibilities, and practical steps for human rights due diligence on hash databases and automated filtering.

Resources and references:

- J. Nathan Matias and Sarah Gilbert, Citizens and Technology (CAT) Lab at Cornell University, “Freedoms of Assembly and Association in Digital Technologies,” October 2024.
- Asterisk Magazine, “The Making of Community Notes,” November 2024.
- Meta, “Our Approach to Labeling AI-Generated Content and Manipulated Media,” April 2024.
- Meta, “Introducing Community Notes- Adding Context to Posts.”
- E. Glen Weyl et al., “Prosocial Media,” March 2025.

SECTION 2

Hash Databases, Due Diligence, and the Boundaries of Government Oversight

This post was published on GNI's website on October 27, 2025.

As online platforms increasingly rely on shared technical systems to detect and remove harmful content, the human rights implications of those systems are becoming clearer, and more urgent. Tools such as hash and signal databases play a vital role in addressing serious online harms, including terrorist and violent extremist content and child sexual abuse material (CSAM). Yet as governments expand regulatory oversight and pressure platforms to act faster and more broadly, these mechanisms can also create new risks for freedom of expression, privacy, and due process.

In October, the Global Network Initiative (GNI) convened the second session in its learning series examining the role of human rights due diligence in content moderation. This series aims to deepen collective understanding of how different models of content governance can impact the rights to freedom of expression and privacy, particularly as governments around the world become more active in regulating online spaces.

Hash databases – and broadly signal sharing – is becoming a staple in industry efforts to combat terrorist content, violent extremism, and CSAM. But their widespread use raises complex questions about fairness, accountability, and rights protection.

The October call focused on the use of hashed and signal databases in content moderation, exploring how these technologies are evolving, what risks they introduce, and how meaningful safeguards and oversight can ensure their operation aligns with international human rights standards. This piece provides an overview of key takeaways from the learning call.

The promise and peril of shared hashes and signals

At their core, hash databases allow platforms to share unique digital fingerprints (hashes) of known harmful content, without exposing private user data. In theory, this enables more consistent and efficient removal of illicit content. Yet the industry is swiftly moving into signal sharing, which can encompass URLs, usernames, and other identifiers beyond simple image or video hashes. This expansion promises greater coverage of harmful activity, such as human trafficking or large-scale scams, but it also deepens the human rights stakes.

Signal-sharing systems heighten risks to privacy, raise the potential for misidentification, and expand the range of content that could be swept into enforcement actions. Because definitions of "terrorism" or "violent extremism" are often inconsistent across jurisdictions, there is real danger of scope creep, where signals used for one purpose gradually extend to others less closely justified. Moreover, once content is removed, it may be irrecoverable, even when it has documentary or evidentiary value.

These risks are not evenly distributed. Users from historically marginalized or surveilled communities – such as Muslims, LGBTQ+ individuals, or minority language groups – are more likely to suffer from wrongful enforcement. Smaller platforms, particularly those without robust moderation capacity, face a distinct challenge: they have less capacity to review materials related to shared signals, which could exacerbate the risks.

Embedding stronger due diligence and oversight

Hash and signal-sharing initiatives such as GIFCT and the Tech Coalition have taken steps to embed human rights safeguards into their work. They have conducted human rights due diligence, engaged civil society through advisory bodies, and committed to transparency and accountability through regular reporting and independent oversight. These measures aim to ensure narrow, legitimate use and prevent misuse, providing a foundation for continued improvement as the field evolves.

Throughout the call, participants emphasized that human rights due diligence must not be an occasional exercise but a continuous, embedded practice. Companies and database operators should proactively assess risks, engage civil society experts, and build in auditability, transparency, and pathways for remedy.

Meaningful safeguards might include limiting data shared or stored, preventing bulk downloads, automatically flagging anomalous behavior, and requiring human review for high-risk signals. Oversight mechanisms should be independent, with stakeholder participation from affected communities, researchers, and human rights bodies. Transparency reporting must go beyond

aggregate numbers, offering insight into how decisions are made, how errors are handled, and under what conditions government requests are accepted or rejected.

Equally vital is creating redress mechanisms: individuals should be able to challenge wrongful takedowns or account suspensions and have their content re-evaluated. Feedback loops, where flagged signals can be contested and revised, help databases evolve and self-correct over time.

Participants emphasized that moderation practices must be grounded in international human rights law, particularly Articles 19 and 20 of the International Covenant on Civil and Political Rights (ICCPR), which require that any restriction on expression meet the tests of legitimacy, legality, necessity, and proportionality. These principles are essential to ensuring that moderation systems and hash databases do not impose arbitrary or excessive restrictions on speech. Yet in practice, the absence of clear definitions, consistent standards, or meaningful due process makes these tests difficult to uphold.

The discussion also touched on the growing interest in applying signal-sharing models beyond the technology sector. Some financial services institutions, for example, have begun exploring ways to contribute signals related to online fraud or child exploitation. While cross-sector collaboration can strengthen responses to complex harms such as sextortion, it also introduces new challenges. Many of these industries lack the established frameworks and human rights experience that have developed within the technology sector through initiatives such as GNI and GIFCT.

One example is the Irish Central Bank's role as a trusted flagger under the EU's Digital Services Act. As a trusted flagger, the Central Bank of Ireland can report potentially illegal content (e.g. fraudulent services or scams) that must be given priority by platforms, and acted upon without undue delay. This means that, in contexts like signal sharing, when such entities supply signals tied to illegal content in their area of expertise, those signals carry more weight, faster processing, greater legal obligation by platforms, but also bring risks if signals are not well-curated, definitions are fuzzy, or oversight is insufficient. Strong collaboration among companies, civil society, and governments can help ensure that new entrants learn from existing best practices and avoid repeating early mistakes. At the same time, the diversity of sectors, languages, and operational cultures involved means that developing shared standards for transparency and accountability will require careful coordination and sustained dialogue.

Government pressure and overreach

Speakers warned of increasing government interest in these databases, whether through regulation or direct access requests. While cooperation between industry, government, and civil society can play a constructive role in addressing online harms, government involvement must be carefully managed to prevent politicization or censorship. Overly broad or unclear legal

definitions risk transforming tools designed for safety into instruments of repression. Several participants stressed that clear governance frameworks and procedural safeguards are necessary to prevent such outcomes.

The discussion underscored that these challenges do not have easy answers. As one participant reflected, while there are real human rights risks associated with hash and signal sharing, there are – likely larger – human rights risks in not using them at all, particularly when it comes to protecting children and preventing violence. The goal, therefore, is not to reject these tools but to ensure they are implemented responsibly, transparently, and with meaningful oversight. Without proper safeguards, databases created to counter violent extremism or child exploitation could be repurposed for censorship or surveillance, undermining the principles of legality, necessity, and proportionality that international human rights law requires.

The challenge is compounded by the fact that these systems often operate in highly automated or semi-automated modes, leaving little room for human review or accountability. In turn, users whose content or accounts are affected may receive no explanation or recourse.

Toward a rights-respecting path forward

The discussion underscored a persistent tension between the serious human rights risks posed by hash and signal systems and the reality that refusing to use such tools is not a straightforward alternative, especially as platforms face mounting pressure to address child abuse, terrorism, and other grave harms. The path forward lies not in rejecting these tools, but in building them responsibly, transparently, and with accountability baked in.

This learning call formed part of GNI's ongoing effort to deepen understanding of how human rights principles can be effectively applied to evolving content moderation practices. GNI will continue to convene members and experts to identify practical ways companies can strengthen transparency, accountability, oversight, and remedy across the stack.

Resources and references:

- BSR Human Rights Assessment of GIFCT
- Technology Coalition Human Rights Impact Assessment
- GIFCT Definitional Frameworks and PrinciplesFrameworks Microsite
- Oversight Board Case: Protest-Related Cartoon in Colombia
- EFF Report: *Caught in the Net*
- ROOST Open Source Tools for Trust and Safety

SECTION 3

Navigating AI Moderation and the Risks to Free Expression

Online platforms are increasingly integrating AI and AI-based tools into content moderation, creating risks to fundamental rights, especially freedom of expression and privacy. As automated systems scale, they also concentrate power over what is seen, who can speak, and how platforms respond to government pressure.

In October, the Global Network Initiative (GNI) convened a learning call exploring how AI is being used in content moderation, how government interventions intersect with these tools, and what human rights due diligence (HRDD) looks like in this shifting landscape.

The call was part of a broader series that aims to deepen collective understanding of how different models of content governance can impact the rights to freedom of expression and privacy, particularly as governments around the world become more active in regulating online spaces. As with all GNI activities, these discussions were held under GNI's policies, including our antitrust compliance policy and code of conduct.

Moderation at scale

A central theme was scale. Automation has become essential in content moderation, helping platforms meet – at times legally mandated – response times and reducing the psychological toll on human moderators. Yet this reliance brings significant rights concerns. Keyword filters and hash-matching tools routinely take down lawful speech, including documentation of abuses, because they cannot distinguish between harmful content and reporting. In one case, automated translation errors led to benign Arabic phrases being read as violent commands. These incidents underscore the sometimes imprecise nature of current systems. While these tools are significantly improved from a few years ago, they still lack the ability to fully grasp nuance or intent.

The conversation considered the effectiveness of AI tools across different problem areas. These systems tend to perform relatively well when content can be judged from the image or text itself, such as in the detection of nudity. They perform far less effectively in cases that depend on external context, such as misinformation or identity verification, where accuracy requires understanding beyond the content itself.

Participants noted that while they come with risks, large language models do introduce new opportunities as well. Trust and safety teams are expanding the use of LLMs in operations, as the tools have improved enough to assist in shaping policy, managing appeals, and drafting user notices, not just assisting not with flagging content. These systems can be updated quickly when policies or standards change, making moderation more responsive and adaptable.

Yet, many challenges and risks remain. The Center for Democracy and Technology (CDT)'s recent paper, Lost in Translation: Large Language Models in Non-English Content Analysis, offers a deeper look into some of these challenges. The study found that large language models (LLMs) used for moderation perform unevenly across languages because most are trained primarily on English data. The problem compounds when crises erupt in regions where datasets are sparse and don't adequately represent the new and evolving contexts and linguistic patterns. Systems that function adequately in English are much less accurate when analyzing speech from other linguistic or cultural communities.

Additionally, in this new tooling landscape, governments could seek to influence moderation more indirectly and opaquely, by providing standards or expectations that are then encoded into these systems. The same flexibility that allows platforms to refine moderation practices could also facilitate more direct state involvement.

Regulating automated moderation

Several participants warned that AI tools could become bargaining instruments in negotiations between platforms and governments. Companies might deploy or withhold certain moderation features to avoid regulatory penalties or to shape future oversight. Others cautioned against the euphemistic language often used to describe moderation, noting that, in practice, it can function as a form of censorship.

The discussion turned to the broader regulatory environment, where laws increasingly incentivize the use of AI while providing little guidance on how to govern its responsible use. For instance, laws in multiple jurisdictions require platforms to proactively take steps to mitigate content-related harms or risks, which in practice would likely involve increased reliance on automated approaches. While these laws may include language underscoring the importance of preserving freedom of expression, the focus of their compliance and enforcement provisions tends to prioritize content moderation. This "risk-based" orientation focuses on harm mitigation rather

than on promoting pluralism or safeguarding freedom of expression. Without clearer guidance, platforms may end up over-removing content or implementing rigid systems that suppress lawful speech.

Participants reflected on how these systems are transforming not just moderation but the structure of online visibility itself. Freedom of expression online is no longer simply about the ability to post, it's also about the ability to be seen. With algorithmic curation, recommendation, and demotion at play, "freedom of reach" becomes as consequential as freedom of speech. Practices like "shadow banning" or "algorithmic throttling" can silence users without explicit removals, creating invisible layers of moderation that escape oversight.

A participant noted that, as a result of these approaches, the Internet is shifting from an "open by default" to a "closed by default" environment with an increasing share of online interactions taking place between humans and bots. These bots are trained both to function as chat assistants and as autocomplete systems. They described this dynamic as a form of value creation suggesting that the entire online space should now be understood as operating within this emerging paradigm of the future.

As governments increasingly demand action against "harmful" or "illegal" content, platforms face mounting pressure to embed compliance into their moderation infrastructure. The danger lies in allowing regulatory demands to dictate algorithmic design. AI, once intended as a tool for scale and safety, risks becoming a conduit for indirect state control. At the same time, platforms may use their AI capabilities as leverage in negotiations with governments, offering or withholding moderation features to shape regulation in their favor.

Role of human rights due diligence

Across the discussion, participants emphasized that effective human rights due diligence (HRDD) must evolve alongside AI itself. HRDD cannot be a one-off assessment; it has to be built into product design, testing, and deployment. Continuous monitoring, independent evaluation, transparency about error rates, and meaningful user recourse are essential ways that platforms can ensure that their content moderation efforts avoid creating unintended consequences. Open channels for engagement with key stakeholders, such as civil society organizations and researchers, can also help platforms understand the impacts of automated tools and find ways to adjust them. Platforms should make moderation datasets and model evaluations public wherever possible, while regulators must focus on overseeing systems rather than dictating content outcomes.

This learning call contributed to GNI's ongoing work to explore how human rights principles can be applied in the context of AI and content moderation. GNI will continue bringing together members and experts to identify practical approaches for enhancing transparency, accountability,

oversight, and remedies across platforms, ensuring that automated systems are designed and operated in ways that respect fundamental rights.

Resources and references:

- Dave Willner and Samidh Chakrabarti, "Using LLMs for Policy-Driven Content Classification," January 29, 2024.

- Agustina Del Campo, Nicolas Zara, and Ramiro Álvarez-Ugarte from the Center for Studies on Freedom of Expression at the University of Palermo, "Are Risks the New Rights? The Perils of Risk-based Approaches to Speech Regulation," March 2025.

- Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon, "Algorithmic Arbitrariness in Content Moderation," June 2024.

Acknowledgements

ABOUT THIS REPORT

These calls featured a range of speakers who brought a range of expertise from backgrounds within academia, civil society, human rights consultancies, tech companies, and trust & safety organizations.

As with all GNI activities, the discussions that informed the initial blog posts and this subsequent report were held under GNI's policies, including our antitrust compliance policy and code of conduct. Please note that the views expressed in this report do not reflect the positions or opinions of the organizations or representatives who participated in the related learning calls.

This report was funded through a grant from the Ministry of Foreign Affairs of the Government of the Netherlands. The work was conducted independently by GNI.

The report was written by GNI's senior fellow, Ramsha Jahangir.

The report was designed by Meher Rajpal.

ABOUT GNI

The Global Network Initiative (GNI) is the leading multistakeholder forum for accountability, shared learning, and collective advocacy on government and company policies and practices at the intersection of technology and human rights. We set a global standard for responsible company decision-making to promote and advance freedom of expression and privacy rights across the technology ecosystem, in particular when addressing overly broad government requests and restrictions.

GLOBAL
NETWORK
INITIATIVE