GLOBAL NETWORK INITIATIVE

TABLETOP EXERCISE

Trust & Safety, Human Rights, and Content Moderation

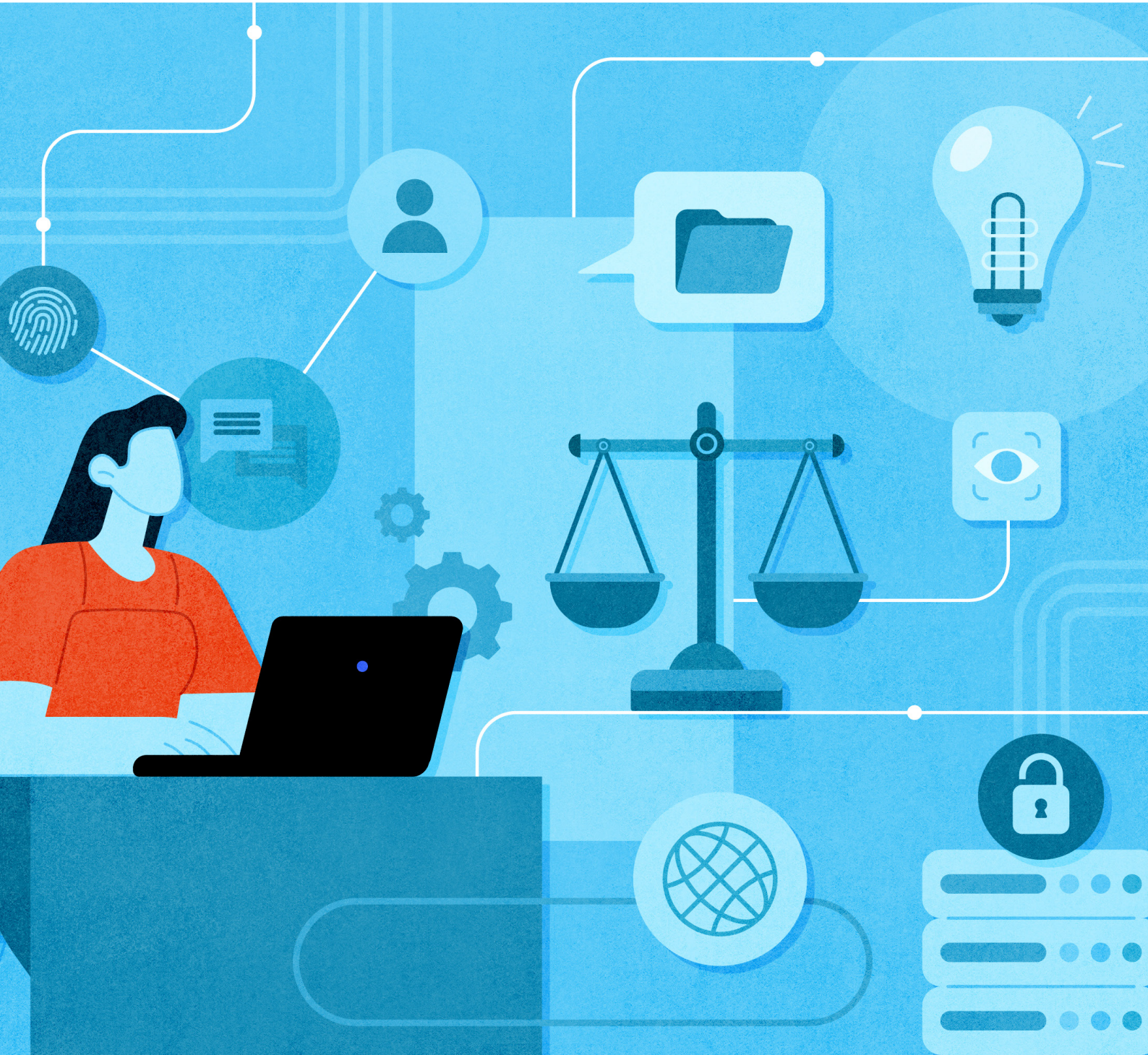

Table of Contents

About This Exercise

This is a fictional exercise designed so that relevant stakeholders – including technology company practitioners, civil society experts, and academics who work in areas such as platform policy, trust & safety, and digital rights – can work together and build on practices to address evolving government restrictions, demands, and legal requirements in rights-respecting ways. The exercise aims to illustrate some of the types of government mandates and requirements emerging in the context of content moderation, highlight factors that should be considered when complying with these, and illustrate how company decisions might impact human rights.

This exercise was developed as part of a workshop co-hosted by the Global Network Initiative (GNI) and the Digital Trust and Safety Partnership. It is part of a series of tabletop exercises produced by GNI that builds off of the "Across the Stack" tool, which GNI and Business for Social Responsibility (BSR) developed to explore how human rights due diligence considerations, including those around privacy and freedom of expression, intersect with different types of products and services across the tech stack.

Objective

This exercise seeks to explore trends related to government requirements relevant to content moderation. It seeks to spark discussion around the different human rights considerations involved in content moderation and discuss the opportunities and challenges that human rights and trust and safety communities face when developing and implementing policies and tools for content moderation.

Regulatory Overview and Context

In their efforts to address online harms, many governments have put in place, among other requirements, regulations and policies either requiring or encouraging companies to moderate content. This has included requirements for companies to use or have the capability to use automated tools for content moderation, requirements for language capabilities in moderation, the establishment and use of hash-databases, and structures such as trusted flaggers, as well as recommended mitigation measures specific to content moderation. Depending on the context, these requirements can be coupled with broad categories of prohibited content, short timelines for acting on content, and heavy-handed penalties for non-compliance. Civil society has also noted that less direct measures, such as mandatory risk assessments and recommended mitigation measures, can incentivize the over-moderation of content[1]. Alongside requirements for content moderation, there are also increasingly detailed requirements for transparency of moderation processes and decisions. Examples include India's Intermediary Liability guidelines, the EU Digital Services Act, the Australian and UK Online Safety Act, the Indonesian SAMAN system, and the proposed EU Regulation to Prevent and Combat Child Sexual Abuse[2].

[1] https://globalnetworkinitiative.org/wp-content/uploads/GNI-DTSP-Forum-Summary.pdf

[2] India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules https://www.meity.gov.in/writereaddata/files/Information%20Technology%20%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20%28updated%2006.04.2023%29-.pdf; EU Digital Services Act https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en, UK Online Safety Act https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer, EU Regulation to Prevent and Combat Child Sexual Abuse https://home-affairs.ec.europa.eu/policies/internal-security/protecting-children-sexual-abuse/legal-framework-protect-children_en

Overview of Content Moderation

Many online platforms need to moderate user-generated content at scale. User-generated content[3] could include any form of "organic" content, such as images, videos, text, and audio, that has been posted by users on a platform or service. Moderation efforts by companies can extend to both content and user accounts. Moderation can take different forms, including detection (flagging and labeling), removal (blocking, suspending, and removing), and curation (recommending, demoting or promoting, and ranking)[4].

The specifics of the company – including the business model, user base, scale, and types of content hosted – will impact the types of moderation undertaken. Examples of the approaches to content moderation include:

- A combination of automated content moderation and human review (primarily through outsourcing to trained teams of vendor moderators, and then escalating key cases to company staff).
- Enabling external stakeholders and users to give input to moderation, through programs such as "Trusted Flaggers", and "Community Notes"
- Enabling volunteers from the user community to moderate directly, such as Reddit and Wikipedia; in some cases, platforms may also engage employees to oversee community volunteers.
- Platforms that have private messaging features rely on user reporting and community moderators and have to determine what level of moderation is possible – particularly given encryption – and what level of moderation is an appropriate balancing of key goals and rights, such as user privacy and safety.

[3] According to the Trust and Safety Professional Association (TSPA) glossary, user-generated content can refer to "links, text, images, or videos created and/or shared by a user." https://www.tspa.org/curriculum/ts-curriculum/glossary/.
[4] This taxonomy of forms of moderation is from Udupa, Sahana, Elonnai Hickok, Antonis Maronikolakis, Hinrich Schuetze, Laura Csuka, Axel Wisiorek, Leah Nann. 2021. "AI, Extreme Speech and the Challenges of Online Content Moderation". AI4Dignity Project, https://doi.org/10.5282/ubm/epub.76087.

Tabletop Exercise: Hypothetical Scenario

Platform Background

WorldPlatform is a multinational social media platform based in a rights-protective jurisdiction with strong rule of law protections, and with operations in markets across the Americas, Europe, the Middle East, Africa, and Asia-Pacific. WorldPlatform enables users to post publicly or privately to their followers, as well as to send messages directly; these are not encrypted. WorldPlatform does not require users to display their real names, but the platform does require and verify a user's real name and birthdate upon registration.

WorldPlatform has centralized departments for Trust & Safety, Legal – including a sub-team of Human Rights experts, Public Policy, and Engineering. WorldPlatform also has a small office in each country it operates in, staffed by local leads, including one staff member who liaises with the government.

WorldPlatform has a number of processes through which human rights impacts are identified and addressed. WorldPlatform is a Global Network Initiative member as well as a member of the Digital Trust and Safety Partnership, meaning they have committed to GNI's Principles on Freedom of Expression and Privacy. In particular, WorldPlatform uses the United Nations Guiding Principles on Business and Human Rights to prioritize human rights concerns and impacts. When salient human rights impacts are identified, WorldPlatform considers a range of steps to better understand related risks and potential mitigations, including whether to conduct a human rights impact assessment – which could be a rapid, focused analysis or a slower, deep-dive exercise. Additionally, WorldPlatform publishes an annual transparency report that includes information about content removal decisions undertaken in response to government demands.

Your Role

You are on a **cross-functional team of senior managers at WorldPlatform**. The team includes leaders from the company's centralized Trust & Safety, Legal (including a Human Rights expert), Policy, and Engineering departments, as well as local representatives from key offices. The team can seek additional expertise as needed from colleagues or outside experts. Several emerging regulatory requirements and practices from jurisdictions you operate in have recently been brought to your team to address:

GOVERNMENT DEMAND SCENARIO

Jurisdiction 1 - Government-required automated content moderation

The government in this jurisdiction has recently imposed intermediary liability rules, which include requirements for online platforms to:

- Endeavor to deploy technology-based measures, including automated tools or other mechanisms to proactively identify information that has been identified as illegal under the rules.
- Implement mechanisms for appropriate human oversight of measures taken through automated means, including a periodic review of any automated tools deployed to evaluate the accuracy and fairness of such tools, the propensity of bias and discrimination in such tools, and their impact on privacy and security.

Jurisdiction 2 - Government-required systems for trusted flaggers, risk assessments, and risk mitigation measures

The government in this jurisdiction has recently passed a new Online Safety Act that puts in place several structures relevant to content moderation practices. This includes:

- A structure for government-appointed trusted flaggers whose flags online platforms must prioritize for review and action within 12 hours.
- A mandatory, annual risk assessment which must include consideration of how the design of recommender and content moderation systems influence risks on a platform.
- Requirements for online platforms to establish reasonable, proportionate, and effective risk mitigation measures. The Act notes that this could include adapting content moderation processes and testing and adapting algorithmic systems, including recommender systems.

Jurisdiction 3 - Government pressure around moderation

Instead of passing regulation, the government in this jurisdiction has exerted informal pressure on WorldPlatform to not moderate certain types of content- particularly content published by the ruling party- even if it has been proven to be inaccurate. This pressure has taken the form of phone calls, letters, and social media posts by the ruling party criticizing the company's

moderation decisions. WorldPlatform has the following practices in place:

- User reporting systems;
- Community moderators;
- Partnerships with fact-checkers.

Internally, there have been discussions to re-evaluate these measures and determine whether the company should rely more heavily on user-based content moderation systems.

Jurisdiction 4 - Government required hash-sharing databases

The government has recently passed a new Child Safety Act. Among other requirements, the Act establishes a hash database maintained by a domestic regulator and requires that the company scan their services for new violating content to report to the hash database and use the hashes stored in the database to report and remove previously identified violating content.

Discussion

Discussion questions:

- How would you respond to this type of government requirement or pressure? What technologies, policies, and processes would you consider or implement?
- What are the human rights impacts of different {automated, user-driven, risk-based} technologies, policies, and processes related to content moderation?
- What safeguards would Trust & Safety and Human Rights teams look to integrate when implementing these technologies, policies, and processes?
- What tools can the Trust & Safety community use to understand the impact of their work on human rights?
- How do the Trust & Safety and Human Rights communities complement each other when responding to these requirements?
- How can Human Rights Frameworks be applied in the Trust & Safety space?
- What areas are most important for these communities to work together?
- What is needed to create avenues for information sharing, collaboration, and shared understanding?

Aspects to keep in mind during the discussion include:

- How might you approach these questions from a Trust & Safety team perspective? From a Human Rights team perspective?
- What might be different if this request was directed at a different type of technology company (e.g. an open web platform, an encrypted messaging platform)
- What might be specific aspects to consider depending on the jurisdiction and global implications of the request (e.g. if sent in a Global Majority context)?
- How can Trust & Safety measures advance human rights in this context?

GLOBAL
NETWORK
INITIATIVE