Tabletop Exercise

Rights-Respecting Responses to Government Demands

About This Exercise

This is a fictional exercise designed so relevant stakeholders – including social media company practitioners, civil society experts, and academics who work on areas like platform policy, trust & safety, and digital rights – can work together and build on practices to address government restrictions and demands in rights-respecting ways. The exercise aims to illustrate some of the types of overbroad government restrictions and demands that Open Web, Social Media, and Online Messaging Applications companies may face, trade-offs these companies need to consider, and how company decisions might impact user rights. This document opens with a contextual overview of the current sectoral and regulatory landscape and is followed by the hypothetical "tabletop exercise".

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

Overview of the Current Landscape

Industry landscape

Broadly, Open Web, Social Media, and Messaging App companies host and facilitate user-generated content for public, semi-public, or private transmission.[1] Many of these companies operate across multiple jurisdictions and markets. Open Web, Social Media, and Messaging App companies have a variety of different business models, product architectures, product features, and policies as well as approaches to enforcement. These companies can have comparable services and are often regulated by similar sets of laws, yet their services are unique:

- Company business models can be can be for-profit (e.g. Meta) or non-profit (e.g. Wikipedia);
- their product architectures can be based on proprietary systems (e.g. TikTok) or open-source systems (e.g. Signal);
- their features can have varying levels of encryption; and
- their policies can set different thresholds for acceptable user behavior (e.g. platforms designed for kids versus for adults-only)

[1] This classification of companies is from the "Across the Stack" tool.

The nature of these companies requires them to operate at and moderate user-generated content at scale. User-generated content[2] 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 content as well as 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).[3]

The specifics of the company – including the business model, scale, and types of content– impacts the types of moderation undertaken. For example:

- companies typically use 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).
- some companies enable external stakeholders and users to give input to moderation, through programs such as "Trusted Flaggers", which is required under the EU's Digital Services Act, and X's "Community Notes"
- some platforms enable 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 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.

Given this, the risks to user rights from overbroad government demands for content removal and restriction and access to user data will be distinct, as are the options each company has to respond.

Regulatory landscape

Open Web, Social Media, and Messaging App Companies are increasingly subject to governmental initiatives that place the responsibility on tech intermediaries to moderate user content proactively or in response to a request pursuant to domestic law and/or civil liability, and/or court proceedings. These are often referred to as "intermediary liability" laws.[4]

Key trends that have emerged in intermediary liability law and policy include:

- broad categories of content for removal as well as criminalization of content such as disinformation.
- short time frames for content removal,
- heavy penalties for failure to comply with governmental requests for content removal
- requirements for local offices to be established

[2] 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/.

[3] 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.

[4] For more on these trends, see GNI. 2020. "Content Regulation and Human Rights: Policy Brief," https://globalnetworkinitiative.org/wp-content/uploads/GNI-Content-Regulation-HR-Policy-Brief.pdf.

- proactive monitoring requirements
- traceability requirements.

Examples include India's Intermediary Liability guidelines, Germany's NetzDG (Network Enforcement Act), Brazil's Fake News Bill 2630, and Sri Lanka's Online Safety Act.[5]

The recently expressed willingness by some governments to impose intermediary liability for user content stands in contrast to earlier regulatory approaches, such as Section 230 of the Communications Decency Act in the United States and the e-Commerce Directive in the EU, which create clear safe harbors for tech intermediaries against liability for user-generated content.

Across jurisdictions, regulators are also starting to pursue legislation focused on platform accountability such as seen in the Digital Services Act and the UK Online Act. These often include requirements such as risk assessments, transparency reporting, and access to data for research purposes.

Additionally, social media companies may be subject to other relevant regulatory frameworks that impose obligations on how they treat user data and require them to respond to government requests for user information or removal of content. This can include data protection provisions, Criminal Codes, and legislation relating to cyber security or information technology. GNI has mapped governments' legal authorities to intercept communications, obtain access to communications data, or restrict the content of communications in more than 50 countries in the Country Legal Frameworks Resource.[6]

Government demand landscape

One consistent theme throughout GNI's history is the continuous evolution and expansion of government efforts to conduct surveillance and censor content. Fifteen years ago, most governments had little or no clear authority to make demands for user data or censorship, especially regarding Internet-enabled services. Since then, governments have increasingly considered and found ways to regulate content and conduct – both through formal legal demands and through less direct, non-legal approaches, such as pressuring intermediaries in relation to setting and enforcing policies.[7]

Complicating matters further for companies is the fact that governments are experimenting with this expanded toolkit at a time when global geopolitical developments are emboldening authoritarian and autocratic governments, while simultaneously leading some democratic governments to pull their punches and shy away from visibly defending companies or confronting those who make inappropriate demands of them. Companies' ability to resist or

[5] 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; Germany's NetzDG https://perma.cc/9W8E-GSWM; Brazil's Fake News Bill 2630 https://www25.senado.leg.br/web/atividade/materias/-/materia/141944; and Sri Lanka's Online Safety Act https://www.parliament.lk/uploads/acts/gbills/english/6311.pdf.

[6] GNI. 2024. "Country Legal Frameworks Resource", https://clfr.globalnetworkinitiative.org/.

[7] GNI. 2020. "Content Regulation and Human Rights: Policy Brief," https://globalnetworkinitiative.org/wp-content/uploads/GNI-Content-Regulation-HR-Policy-Brief.pdf.

mitigate the impact of overbroad government demands or restrictions is especially limited in the context of conflict scenarios, public emergencies, and elections.

As illustrated over time in GNI's assessments[8], one reasonable outcome of risk assessment and responsible company decision making in these challenging contexts is to avoid entering or to consider exiting certain challenging jurisdictions. But if responsible tech companies avoid these contexts, users are left even more vulnerable.

These real-world scenarios have provoked discussions within GNI and beyond about what responsible entry, remain, and exit look like for technology companies; topics that GNI continues to explore through shared learning like these hypothetical scenario discussions.

Tabletop Exercise: Hypothetical Scenario

Country background

The **Republic of Genovia** is a medium-sized country in the "Majority World". It will hold presidential elections next year. The incumbent **President** is campaigning for a third term and her party holds a majority in the legislature. The leading opposition candidate is a **Senator**, a longstanding representative of a minority ethnic group. The Senator has been gaining popularity as a vocal critic of the current administration's increasingly authoritarian actions.

A few years ago an unidentified driver tried to crash a van into a crowded square near the Presidential residence; the driver was killed. Independent reports have been unable to verify the attacker's affiliation, but the President suggested that the driver was affiliated with her challenger, the Senator. Since the attack, public demonstrations in the capital have become more frequent, escalating into violence. The government has taken action against some opposition supporters and journalists, citing threats to public order.

Platform background

WorldPlatform is a multinational online social media platform based in a rights-protective jurisdiction with strong rule of law. 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 a user's real name and birthdate upon registration. WorldPlatform regularly receives requests and demands from governments around the world to share data about specific users (including location data, private content, and messages), as well as to take down specific content, types of content, or accounts.

WorldPlatform has a large user-base in Genovia; it is the largest social network used by Genovians. In addition to WorldPlatform, there are two popular, smaller regional competitors.

[8] GNI. "Public Assessment Reports," https://globalnetworkinitiative.org/what-we-do/foster-accountability/assessment-reports/.

WorldPlatform has centralized departments for Trust & Safety, Legal – including a sub-team of Human Rights experts, Public Policy, and Engineering. In Genovia, WorldPlatform has a small office, 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, meaning they have committed to GNI's Principles on Freedom of Expression and Privacy. In particular, WorldPlatform uses United Nations Guiding Principles on Business and Human Rights guidance on severity, to prioritize human rights concerns and impacts. When salient human rights impacts are identified, WorldPlatform considers a range of steps, including conducting human rights impact assessments – from rapid to deep-dive – and subsequent mitigations. Additionally, WorldPlatform publishes an annual transparency report.

WorldPlatform is closely tracking the situation in the Republic of Genovia and internally has identified it as a 'crisis', activating the company's internal crisis protocols.

Legal Background

The **Constitution of the Republic of Genovia** guarantees citizens the right to free expression and protects the privacy of citizens and their homes, correspondence, telephone conversations and telegraphic communications. However, the Constitution also permits laws that abrogate these protections if they are reasonably justifiable in a democratic society in the interest of defense, public safety, public order, public morality or public health.

The relevant communications authority oversees the regulation of online intermediaries and administers its intermediary guidelines, which include:

- A court and/or authorized government agency may issue an order for the removal of content that is prohibited by law.
- The intermediary must remove or disable access to the information within 24 hours of receiving the order.
- The intermediary must retain user registration information for a period of 165 days and must provide the information within 72 hours to authorized government agencies on receipt of an order for the purposes of identity verification or the prevention, detection, investigation, or prosecution of an offense.
- The intermediary must appoint a Compliance Officer who is a resident of the Republic of Genovia and can be held liable for noncompliance with government orders and directions in such circumstances as well as a 24x7 contact person to coordinate with law enforcement.
- If the intermediary provides services in the nature of messaging, they must enable the identification of the first originator of the information and provide the same on the receipt of a judicial order.

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.

Government Demand Scenario

There is another public demonstration, which results in the hospitalization of one of the President's supporters and the arrest of an opposition supporter. In the wake of this, opposition supporters marched again, nonviolently, blocking streets as they moved toward the Presidential residence. The demonstration was partially organized on WorldPlatform, where users are posting information about the demonstrations and where to gather; in particular, three posts by one user got significant engagement.

A **government representative** calls the local WorldPlatform office via a publicly available general number and demands that the company:

1. Remove all content related to organizing the protest as it is an "incitement to violence" and
2. Share data with the government about the user whose three posts about organizing the protest went viral.

The call is answered at WorldPlatform's local office. The **local WorldPlatform staffer** requests a written Government Request Form – per human rights best practices and WorldPlatform's publicly available policies. But, the government official says there's no time for a written request. Per the policy, the local staff member then escalates the request to your cross-functional team.

Cross-Functional Team Discussion Questions

During this tabletop, aspects to keep in mind 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 online platform (e.g. an open web platform or 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)?

Questions to explore during this tabletop exercise include:

- Process
  - How is the request escalated?
  - Who is consulted?
  - Is more information needed?
  - What frameworks can be used to guide decision-making?

- Decision
    - What options are available and in what configuration?
    - On what grounds is the decision made?
    - What are the risks and trade-offs associated with this decision for the company (including employees) and users of the platform (specifically rights impacted)?
    - What steps can be taken to mitigate negative impacts to users' safety online and offline?