Workshop Report

Engaging Trust & Safety and Human Rights Practitioners on Rights-Respecting Responses to Government Demands

With the growth of online platforms over the last two decades, the risks of harm to individuals, communities, societies, and democracies have evolved and become more complex. Over time, as platforms have become more embedded into people's lives, the challenges of advising platforms on possible risks and responding to harms in real-time and at scale have become ever more complicated, requiring complex trade-offs and decisions. Practitioners have developed a number of approaches – both within and external to online platforms – to assess and mitigate possible harms to people and protect fundamental rights like freedom of expression and privacy. In particular, the "Human Rights" field and the more nascent approach of "Trust & Safety" have both become key areas of expertise, and their practitioners are critical actors in mitigating harms and protecting rights online.

This past October, the Global Network Initiative (GNI) and the Trust and Safety Foundation (TSF) jointly brought T&S practitioners and human rights experts together in a virtual workshop to explore increased engagement across these fields.

Why bring together Trust & Safety and Human Rights?: aligned yet at times distinct fields

Trust and Safety (T&S) professionals identify and address risks and harms to users, primarily within companies and with a focus on developing and operationalizing policies and internal tools to monitor and enforce policies at scale. The T&S function emerged from within online platforms, yet historically has often been insufficiently prioritized as a critical function within these companies. While the field is still nascent, it is professionalizing rapidly and there is a burgeoning ecosystem of third-party vendors developing and providing T&S tools and services. Examples of T&S responsibilities can include content moderation, age verification, processing government requests, and more.

Human Rights or Digital Rights experts—both within and external to companies—have been researching the societal impacts of technology and advocating for the respect of fundamental human rights online, while leveraging established tools, frameworks, and practices developed in the context of international human rights law. As David Sullivan, Executive Director of the Digital Trust and Safety Partnership (DTSP), noted the human rights field is exogenous to the technology sector and benefits from seventy-five years of evolving debate, language, norms, frameworks, and implementation models. Over the last two decades, a sub-field of human rights has focused on addressing the duties of states and responsibilities of businesses with respect to the human rights impacts of private sector activities, as set out in the UN Guiding Principles on Business and Human Rights (UNGPs).[1] The Business and Human Rights approach offers a set of frameworks and vocabulary around concepts such as “salience”, “scale”, “scope”, and “remediability”. Examples of human rights functions within online platforms can include developing and advising on policies, processes, and critical decisions to identify and mitigate harms, such as human rights impact assessments, stakeholder engagement, and crisis protocols.

These two practitioner communities both focus on protecting online users from risk and harm, and therefore have many interests in common. There already is engagement and overlap between them, yet despite shared goals and objectives, there can be differences in the approach, guiding principles, and understanding of effectiveness between the communities. For example, broadly, the trust & safety approach tends to focus on operationalizing policies at scale, whereas the human rights approach focuses on frameworks and principles. There can also be gaps in collaboration and learning across these communities. As an example, there are two distinct predominant conferences for each community – TrustCon is oriented towards T&S practitioners and RightsCon is oriented towards Digital Rights practitioners. There are also distinct non-profit organizations emerging focused on T&S, which are not well integrated into the robust, global ecosystem of non-profits focused on human and digital rights.

Given this moment of rapid technological advances, political realignments, recent and upcoming laws and regulations, and the nascent professionalization of the T&S field, the Atlantic Council's Task Force for a Trustworthy Future highlighted in 2023 that "greater interoperability between T&S and human rights could serve to strengthen both fields, identify new pathways for achieving T&S goals, and improve T&S's ability to narrate its aims more clearly with a wider community of stakeholders."[2] The Task Force noted that the human rights space, as a more mature field, could offer fundamental insights to be incorporated "more intentionally into debates and innovation around T&S as that field emerges" and emphasized that "a rights-centric framework can help establish a foundation for normative debate and key trade-offs."[3]

This is particularly important as the term “trust and safety” does not (yet?) translate globally, whereas human rights is a globally-understood framing. So, there is untapped opportunity to better communicate across these fields collaboratively towards shared goals. Many civil society organizations, including those from “Majority World” countries, use human rights frameworks to identify and highlight the specific impacts that technologies have in their regions, particularly on traditionally marginalized communities. This expertise and insight are crucial for tech companies, given that T&S issues are of global importance and are often

[1] UN Guiding Principles on Business and Human Rights (2011). https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

[2] Duffy, Kat. (2023). *Scaling Trust on the Web: Comprehensive Report on the Task Force for a Trustworthy Future Web*. The Atlantic Council of the United States https://www.atlanticcouncil.org/in-depth-research-reports/report/scaling-trust/.

[3] Duffy, Kat. (2023). *Scaling Trust on the Web: Comprehensive Report on the Task Force for a Trustworthy Future Web*. The Atlantic Council of the United States https://www.atlanticcouncil.org/in-depth-research-reports/report/scaling-trust/.

disproportionately impactful in markets outside of the U.S. and Europe.[4] Yet, T&S functions emerged from predominantly American tech companies, often reflecting priorities and approaches that exhibit a Global North perspective. Another key critique has been that companies are not prioritizing the allocation of T&S resources to Global Majority contexts. Yet, T&S functions like content moderation are often outsourced to contract workers based in Global Majority contexts.

Bridging the Trust & Safety and Digital Rights Fields: A Conversation on Rights-Respecting Company Responses to Overly Broad Government Demands

Given these gaps and the opportunities of bridging them, GNI and TSF brought together T&S practitioners and human rights experts to explore increased engagement across these fields. We did this by collectively discussing an increasingly critical topic – overbroad government demands for user data and restrictions of content that online companies face, which adversely impacts user rights like freedom of expression and privacy.

Why focus on responding to overly broad government demands?

While many tech companies have already committed to the UNGPS and/or the GNI Principles[5], the landscape is rapidly shifting, requiring consistent engagement on how to put those commitments into practice. Fifteen years ago, most governments did not have specific authorities to make demands in this space; they relied on existing, non-digital-specific authorities to make demands for user data, content moderation, or network disruption especially regarding Internet-enabled services. Over the last decade, governmental powers have evolved to be more specific to internet related services and companies. As part of this trend, some governments have begun mandating that platforms assess risks to people, and even regulating content moderation. Various laws have been enacted—such as the Digital Services Act (DSA) in Europe and the UK's Online Safety Act—requiring platforms and service providers to conduct different types of procedures, such as risk assessments, human rights due diligence,[6] transparency reporting, and audits of the architecture, activity, and use of their services. These regulations are creating new incentives and driving new practices across the technology ecosystem,[7] requiring specific expertise and making participation from different internal teams and external stakeholders more relevant than ever. Additionally, many governments – particularly in more authoritarian or less human rights-respecting jurisdictions – have enhanced their authorities and developed more direct ways to access user information and/or restrict expression targeting different layers of the technology stack.

[4] Mukherjee, Sujata & Eissfeldt, Jan (2023). *Evaluating the Forces Shaping the Trust & Safety Industry*. Tech Policy Press. https://www.techpolicy.press/evaluating-the-forces-shaping-the-trust-safety-industry/

[5] GNI Principles. https://globalnetworkinitiative.org/gni-principles/

[6] Across the Stack Tool. https://eco.globalnetworkinitiative.org/wp-content/uploads/2022/11/Human-Rights-Due-Diligence-Across-the-Technology-Ecosystem_Ecosystem-Mapping_Oct2022.pdf

[7] Duffy, Kat. (2023). *Scaling Trust on the Web: Comprehensive Report on the Task Force for a Trustworthy Future Web*. The Atlantic Council of the United States.

Complicating matters further for companies – and therefore their users – 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 mitigate the impact of overbroad government demands or restrictions is especially limited in the context of conflict scenarios, public emergencies, and elections.

A Hypothetical "Tabletop Exercise" to Spur Discussions

During a virtual workshop, held under the Chatham House Rule, participants engaged in a hypothetical tabletop exercise, written by GNI and TSF to foster dialogue and creative thinking around company response to government demands and the role and approach by T&S and Human Rights teams when responding. The exercise focuses on social media companies and aims to illustrate some of the types of overbroad government restrictions and demands that tech companies may face, especially in countries that don't have strong human rights protections. The goal was to exchange experiences, explore types of decision-making processes and tradeoffs that might need to be made, and how company decisions might impact their users' rights. Additionally, we looked for ways to build connections between these two fields, so that practitioners can leverage each other's work and, ultimately, make online spaces safer while ensuring human rights are guaranteed. There were more than 30 participants, with experience within companies, civil society, and academia and from across multiple jurisdictions, including Brazil, Ireland, Sri Lanka, the United States, and Zimbabwe.

Lessons Learned on Pushing Back to Protect Rights

Discussion in the workshop highlighted several learnings. These include:

- **Continuing relevance of existing frameworks**: In the hypothetical situation, the GNI Framework offers more than a decade-and-a-half of relevant experience and guidance. In particular, as a first step in the process, participants noted that the company should require a formal written request from the government that is specific and references the applicable local law, to be able to ascertain whether the demands are lawful and also to document every step of the process.

- **Managing differing internal incentives:** It is complex to operationalize a human rights-centered approach. Different roles within the company likely would have different views and ideas about the demand, depending on their varying incentives. For example, there are trade-offs related to the size of the market, whether there is a risk of the service being removed from a jurisdiction, whether there are staff located in the jurisdiction who could be put at risk, etc.

- **Ensuring processes are informed by cross-functional teams:** It is important that the processes for responding to government demands are designed with human rights expertise from the beginning. Processes for reviewing and responding to government demands should make clear when a government demand might be overbroad, and if so, have a process to escalate those demands to senior decision-makers. If a request needs to be escalated, people with legal, human rights, and T&S operational experience – including senior decision-makers – should be included.

- **Monitoring processes to ensure intent is being applied:** While there's broad guidance and principles that apply, each situation is unique and requires case-by-case responses; there is not a "right" answer for all cases. This is why rights-respecting processes and controls to monitor those processes at the operational level are critical. Decision-makers should document everything and keep relevant evidence when making decisions, in order to capture the rationale behind them at each stage. This is important for many reasons, including that it can enable improvement of processes over time.

- **Navigating on the ground realities:** One critical observation was that, in the real-world, it is much more common for platforms to get demands that look something like the demand in the hypothetical scenario (e.g. for user data or content removal) than it was, say, 10 years ago – and that these demands are much more complex to navigate as many countries have complex regulatory frameworks and laws requiring in-country staff who could be put at risk from company push-back (often referred to as "hostage-taking laws").

- **Importance of dialogue:** Lastly, and critically, there are benefits to reaching out to trusted connections in the country in order to understand any similar precedents, how these prior situations may have been addressed by companies in the past, and any suggestions to take the most rights-respecting, yet feasible, course of action. Additionally, depending on circumstances, it can be productive to try to open conversations with the government to understand more and share specific concerns.

Continuing conversation

The workshop highlighted the importance of bringing trust and safety and human rights communities together, in particular to explore how to better operationalize human rights frameworks into trust & safety work, and in turn, to consider how human rights frameworks might be made more applicable to the day-to-day work of trust and safety teams.

GNI and TSF look forward to continuing these conversations. We are keen to hear your ideas on questions, approaches, and outputs to explore so we can foster productive conversations across the human rights and trust and safety fields.