GLOBAL NETWORK INITIATIVE

humane intelligence

B-Tech

ALGORITHMIC RISK ASSESSMENTS, AUDITS, & HUMAN RIGHTS

KEY TAKEAWAYS FROM A MULTI-STAKEHOLDER WORKSHOP

SEPTEMBER 2024

On 30 May 2024, the Global Network Initiative (GNI), the UN B-Tech Project, and Humane Intelligence co-organized a workshop on algorithmic risk assessments and human rights, and implications for audits, hosted at the EU Delegation to the UN in Geneva. The workshop set out to understand the risks that algorithms can pose to human rights, and the roles of assessments and audits in identifying and mitigating those risks. The event was held on the margins of the International Telecommunications Union's World Summit on the Information Society Forum and AI for Good Summit and over 60 experts from civil society, corporate, and government backgrounds attended.

This document summarizes key takeaways, which could be fruitful for further discussion by the participating organizations and the field at large. Hence, the discussed themes are not an exhaustive list, and are instead a summary of expert views regarding responsible practices for risk assessments.

1. The UNGPs are an existing established framework which can enable rights-respecting risk assessments for AI products.

Policy-makers in various countries across the world are debating or drafting regulation on AI. Several States are exploring or have already adopted regulations requiring tech companies to assess human rights risks. For example, recent EU legislation – such as the Digital Services Act and the Artificial Intelligence Act – and proposed approaches under development in Brazil, include mandatory risk assessments with regard to fundamental rights impact assessments. These approaches could affect the development of legislation in other jurisdictions.

Yet many regulatory initiatives do not incorporate the due diligence expectations laid out by the international standards of business conduct: specifically, the UN Guiding Principles on Business and Human Rights and the OECD Guidelines for Multinational Enterprises on Responsible Business Conduct. In particular, the UNGPs require policymakers to take effective measures to protect against the human rights risks associated with digital technologies. They also provide a framework for companies to assess and mitigate their possible impacts and risks to people.

The companies that develop, deploy, or maintain digital technologies are subject to an ever-increasing number of regulatory initiatives and processes at national, regional and international levels. Yet these new regulatory initiatives do not always align with international human rights standards. There is a risk that the regulatory initiatives, even if well intended, might create incoherence and misalignment with human rights frameworks. The UNGPs provide policymakers with an established framework for how to align regulatory initiatives and processes with human rights standards.

In addition, the UNGPs are already considered best practice in the tech sector. A growing number of companies use the framework of the UNGPs to inform the design, development, and deployment of digital technologies in many ways, including conducting rights-respecting risk assessments. These assessments increasingly include reviews of AI products and features.

The fact that some companies at the forefront of AI development are endorsing and implementing a rights-based approach to risk management points to human rights as a promising foundation for rights-respecting AI practices and risk assessment for algorithmic systems.

National, regional, international and industry-led initiatives focused on advancing responsible AI should use or align to the international standards of responsible business conduct. This means, in particular, integrating a true risk-based approach to identifying and taking action on impacts with a focus on re-centering on:

1. Using severity of risks to people to prioritize impacts for attention; and
2. Setting expectations of companies across the AI value chain commensurate with the nature of their involvement (causation, contribution or linkage) with human rights risks and impacts.

2. Legislators and companies alike must identify suitable methodologies to assess AI products and services with regard to human rights risks.

Triggered by the various regulatory and policy initiatives there is a need to clarify what constitutes responsible AI practices and the use of algorithmic systems, and more specifically how to assess the risks of harm to people and audit those assessments methodologically. How can stakeholders – including engineers – encourage comparable AI risk assessment and auditing benchmarks? What are appropriate methodologies for AI auditing? What data is needed to perform accountable AI audits?

AI risk assessments should center the protection of vulnerable populations, be conducted systematically across the full life cycle of an AI product or service, including algorithmic systems, and seek to reduce power asymmetries. This also includes heightened obligations for public parties in line with the State duty to protect human rights where public parties use AI and/or algorithmic systems. Processes should be built in a participatory manner to allow for meaningful stakeholder engagement.

Effective AI risk assessments should include a wide range of information collected from many stakeholders. For example, assessors should conduct interviews with employees – especially trust and safety teams, human evaluators of algorithmic systems, users, and non-users such as impacted communities. This kind of stakeholder engagement is crucial to understanding the decisions that go into product development and the needs of people affected by the relevant product or service.

In addition to collecting information directly from key stakeholders, AI risk assessments should rely on a variety of quantitative and qualitative data and metrics. Companies need to ensure they have appropriate guardrails for AI systems. For example, algorithmic content governance systems should be regularly reviewed, from the quality of training data, to model specifications, to the accuracy of model outputs; data from trusted flaggers could be a useful input to better understand changes in algorithmic systems used for content governance. If a system includes automated grievance or appeals review, those should also be similarly

reviewed; remedies, recourse, and appeals offerings should be verified for whether they actually meet the needs of people who have been harmed.

The data used to assess risks within AI systems and evaluate mitigations needs to be appropriately representative. This means it should both be evaluated for possible biases, and also might need to be specific to a certain location or group, in order to appropriately understand risks.

Participants agreed that the role of multi-disciplinary teams is essential in terms of expertise and skills-sets regarding human rights, data science, legal and product design. In this regard, collective coordination with parallel teams is essential: Engagement must involve not only policy teams and engineering teams, but also marketing and sales teams, because monetization is one piece of the puzzle related to risks. It is important to have conversations about human rights, choosing what to monetize, when to monetize, who to monetize from etc. Multidisciplinary capacity-building at company level is essential, so human rights approaches can be embedded at the stages of product design and development. Human rights must be integrated in all parts of the company like articulating, publishing, and designing templates and tools that could be used in intersectional ways, with multiple disciplines coming together to raise the profile of the discussion internally.

3. Expert stakeholders need to be involved in oversight and enforcement to ensure both are effective and rights-respecting.

Expert panelists discussed the role of enforcement and supervisory mechanisms, and how civil society and academia can most meaningfully engage around these processes. A human rights-based approach grounded in the UNGPs provides methodology and guidance to: identify and assess impacts to people and society; prioritize risks, determine appropriate action by individual companies, the industry, and broader ecosystem; and provide guidance on how to address tensions.

At the same time, several participants also flagged the need for clearer guidance by regulators to set expectations for what constitutes a high-quality risk assessment and audit.

The current status quo of corporate conduct on AI risk management is highly heterogeneous in nature. Companies take a wide range of approaches to assessing and addressing risks associated with algorithmic systems and/or AI. This includes a wide variety of methodologies and models for assessing risks and impacts to people, which are technical, non-technical, issue-based, jurisdiction-based, and at varying levels of depth.

When applied in practice, a human rights-based approach can fill gaps in understanding risks that other risk assessment approaches have not yet fully captured. Some of the existing responsible AI principles are aligned with human rights, while others lack human rights standards. More work has to be done to ensure that such principles are endorsing a human rights lens as companies seek to apply and operationalize human rights standards.

Human rights need to be included in risk assessments that are increasingly being integrated into existing AI product development processes across the product life cycle, heavily relying

on algorithmic systems. Since standalone assessments done at a particular moment in time are still prevalent, if not the norm, the extent to which companies are assessing their impacts on an ongoing basis and with a human rights lens is unclear. As a result, risk and impact assessment practices within AI companies vary greatly as companies tend to develop their own models for these assessments that are informed by public standards and best practices to various degrees.

Partly due to the high speed of technological innovation, questions of corporate responsibility and human rights risks in algorithmic systems have been outpaced, and the responsible AI field's ability to develop and communicate best practices is limited. As mentioned above, some of the assessment models companies have pursued for advanced AI, for example, have been ad hoc and their application experimental.

Effective enforcement and oversight mechanisms should verify that companies must ensure that even before the design phase, fundamental issues with the data used to train AI models are addressed. This also includes requiring companies to demonstrate human rights due diligence processes, which assess collection of data, makeup of training datasets, and presentation of data and feedback loops.

Going forward, experts stressed the necessity to move towards a level playing field enabled by policy coherence around the UNGPs and their interlinkages with the work of OECD AI expert group and its expert group on AI risk and accountability, OECD Guidelines for Multinational Enterprises on Responsible Business Conduct and its upcoming guidance on due diligence regarding AI. Since most technology companies are operating globally, standards need to represent global alignment, and such expectations need to be mirrored in enforcement and oversight of State regulations and policies.

What's next?

The workshop identified and reaffirmed a series of complex questions that should be further explored to ensure algorithmic systems respect people's fundamental rights, including the five laid out below. We look forward to addressing these questions collectively across the field.

1. How can the human rights and technical fields make it more implementable to adopt the UNGPs – as an existing established framework – to assess risks from AI products?
2. How can stakeholders – including engineers – encourage comparable AI risk assessment and auditing benchmarks?
3. What processes, resources, and structures can better enable civil society and academia to meaningfully engage around algorithmic risk assessments and mitigations?
4. How can we ensure companies appropriately demonstrate their human rights due diligence processes, including assessing their collection of data, makeup of training datasets, and presentation of data and feedback loops?
5. How can companies and external stakeholders use algorithmic risk assessments and audits to ensure accountability and catalyse change?