ROUNDTABLE SUMMARY NOTE

Generative AI Risks and UN Guiding Principles on Business and Human Rights

14 June 2023

Introduction

On 14 June 2023, the UN Human Rights B-Tech Project organized a multi-stakeholder roundtable in San Francisco, launching its Generative AI Initiative. The project seeks to demonstrate the ways in which the *UN Guiding Principles on Business and Human Rights* (UNGPs) can guide more effective understanding, mitigations and governance of the risks of generative artificial intelligence ("generative AI"). Over the course of 2023, the B-Tech Project will:

- Clarify the UNGPs' expectations for companies developing and launching generative AI products in order to achieve unified and more effective human rights risk management approaches across the tech industry;
- Raise awareness and facilitate exchange among key stakeholders and interdisciplinary experts to shape a comprehensive understanding about the role the UNGPs should play in governing generative AI responsibly;
- Inform the debate about policy options for managing human rights risks related to the development and launch of generative AI, through both mandatory and voluntary measures.

The roundtable both began this conversation and is serving to shape the near-term insights and recommendations for the project, due to be published in fall 2023. The meeting brought together participants from diverse stakeholder groups, including representatives from tech companies deploying and using generative AI, academia, civil society and the public policy field. A list of participants is provided at the end of this note.

This summary note highlights the following **five key take-aways** from the meeting.

1. **Place international human rights standards, in particular the UNGPs, at the centre of private sector and regulatory responses to the risks to people of generative AI products**. Participants were unanimous in their view of the unique value of the UNGPs in this context while also noting that this not yet widely understood beyond a relatively small community of company practitioners, civil society and academia.
2. **Curate an authoritative, regularly updated taxonomy and catalogue of human rights risks associated with generative AI**. It is hard to envisage coordinated action between States, business and civil society if there are highly diverse perspectives about the nature and source of risks connected to generative AI, including how these differ from those associated with existing "traditional" AI tools.
3. **Increase efforts to show and explain company practices for managing risks to people across the full lifecycle of generative AI development, deployment and use**. Better understanding of promising approaches as well as shortfalls (and the dynamics underlying these shortfalls) should ground advances in guidance, industry standards and best practices.
4. **Account for the complexity of the generative AI eco-system in governance responses to human rights risks.** Moreover, responses should be informed by extensive multi-disciplinary and multi-stakeholder cooperation.
5. **Convey a greater understanding of what robust human rights due diligence and remedy for harms looks like for specific generative AI technologies and use cases**. This will be key to focusing practitioner and public policy action on practical steps that can be taken now to advance responsible development and deployment of generative AI.

The meeting was held under the Chatham House Rule and moderated by Shift, the leading not-for-profit center of expertise on the implementation of the UNGPs. The roundtable was hosted by Swissnex in collaboration with the Swiss

Government, with additional support provided by the Government of Austria.

One: International human rights standards, in particular the UNGPs, should be placed at the centre of private sector and regulatory responses to the societal risk of generative AI products. Participants were unanimous in their view of the unique value of the UNGPs in this context while also noting that this not yet widely understood beyond a relatively small community of company practitioners, civil society and academia.

Participants at the Roundtable strongly supported the proposition that the UNGPs should be a prominent, even foundational standard of business action, investor attention and public policy responses to the risks of generative AI. Various participants reflected that as the international standard concerning business impacts on people, the UNGPs are an obvious fit with the global challenges and risks of generative AI – risks that are distinct from many past high risk technological innovations and are interwoven with market dynamics that, unless well-governed, will create externalities in the form of harms to the most vulnerable and exacerbated inequalities. In this same vein, many participants pointed to the helpful way in which the UNGPs affirm the distinct and complementary role of States and business actors in addressing business-related human rights risks.

In addition, company representatives and external advisers, emphasized the practical value of the UNGPs as a tool for credible risk management by business. These discussions reinforced that the UNGPs:

- Offer a proven risk-based framework that covers the design, development, deployment, use and misuse of products and services to which many global technology companies have already publicly committed.
- Set a high bar for what constitutes responsible corporate conduct such that attention to avoiding, mitigating and remediating impacts on human rights must be part of corporate governance, strategy, practices and culture.
- Provide a framework to think through the overlapping but differentiated responsibilities of enterprises across technology eco-systems, stacks and value chains based on the degree of a company's involvement in harms.
- Place considerable emphasis on risk management as a socio-technical process that can only succeed if centred on the experiences and perspectives of affected stakeholders, and often requires industry and multi-stakeholder action.
- Do not seek to holdback innovation but rather to define what responsible innovation looks like, with a clear baseline of not releasing technologies into the market without credible mitigations and safeguards in place.

Participants noted the considerable benefit of anchoring responses to the risks of generative AI in the UNGPs given that they are already a recognized global standard of conduct and are increasingly integrated into due diligence and reporting regulations (most notably in the context of the EUs Corporate Sustainability Due Diligence Directive and Corporate Sustainability Reporting Directive). At the same time, discussions surfaced a need for greater clarity about the similarities and differences between the UNGPs and prominent AI and technology design frameworks, principles and legislative developments such as the EU's AI Act and Digital Services Act, the U.S National Institute for Standards and Technology (NIST)'s AI Risk Management Framework, the OECD AI principles, and Value Sensitive Design Theory.

Most participants agreed that many (perhaps even most) business leaders, civil society organizations, academics and government officials focused on responsible conduct by technology companies and users of technology are not aware of, or poorly understand, international human rights standards and the UNGPs. Participants called for greater investment by the UN and others to socialise and firmly advocate for rights-based approaches, including by offering more clarity about the commonalities and distinctions between human rights and ethical or UN Sustainable Development Goals-oriented frameworks.

At the same time, participants agreed that not every issue at the root of generative AI-related risks can, or should, be tackled through the lens of UNGPs. For example, navigating geopolitical interests to reach meaningful multilateral commitment about how to govern technological advance will require a broader range of public policy tools and economic levers.

Two: An authoritative, regularly updated taxonomy and catalogue of human rights risks associated with generative AI is needed. It is hard to envisage coordinated action between States, business and civil society if there are highly diverse perspectives about the nature and source of risks connected to generative AI, including how these differ from those associated with existing "traditional" AI tools.

The roundtable surfaced a multitude of perspectives about the nature of risks to human rights that come with technological breakthroughs, commercialization and the broader use of generative AI. The types of impacts tabled by participants included:

- Impacts on individuals' human rights already associated with "traditional" AI and other digital technologies that could be "super charged" by generative AI. For example, privacy and data protection, algorithmic discrimination (bias), hate speech, online and offline harassment, disinformation, mental health, fraud, and labor rights.
- The use of generative AI in ways that undermine the institutions and societal norms that we rely on for human rights protection and realization (such as democratic institutions, elections, criminal justice, and media), or advance authoritarian governance due to the availability of more powerful tools of societal control, surveillance, and censorship.
- The potential for generative AI to reduce human autonomy through gradual impacts on "rights such as the right to information and freedom of opinion.
- The facilitation of violations of international humanitarian law as generative AI systems become used or abused for defence, security, intelligence and in conflict or humanitarian settings.

Many also noted considerable value in a more precise articulation of the sources of varied human rights risks, such as when they are:

- Intrinsic to generative AI technologies. A few participants raised the importance of understanding the inherent "capability overhang" of general-purpose AI tools that result in unknown and diverse use cases, making it hard to foresee all the potential risks. One speaker used the example of "scaffolded generative AI" to emphasize that evaluations of risk also need to include attention to the realistic ways that experts believe generative AI could have more advanced future capabilities that entail different, more serious or more unpredictable human rights risks than those associated with generative AI's current use cases and stage of technological development.
- Inherent or exacerbated by companies' business model choices and competitive strategies. A central point of emphasis here was the idea that many of the challenges in the sector as a whole connect to the speed of generative AI innovation and development, and business drivers in an ultra-competitive space as companies sought to be first movers in the market. This dynamic extends to smaller, start-up new entrants (and their venture capital investors) who have limited incentive or capacity to implement risk assessments.
- Extrinsic to technological capability and commercial practices. Examples included where risks arise from the contexts within which technologies are used, the vulnerability of specific impacted groups due to historical and current structural discrimination, or outlying malicious actors intentionally using generative AI tools to cause harm. Some participants added their belief that risks related to malicious actors are heightened by open-source development models in generative AI.

Understanding the nature of human rights harms and the various sources of those harms will be critical to identifying fit-for-purpose responses. What is clear is that under the UNGPs, companies across the generative AI value chain all need to start with the broadest possible view of human rights risks, explain clearly to stakeholders how they have prioritised issues on the severity of risks to people, and equally address sources of risks that are "in here" (e.g., in business models or technological choices) as well as "out there". States also need to discern which of the "smart-mix" of regulatory/policy measures, as well other State-based incentives and accountability mechanisms laid out by the UNGPs are necessary to maximize prevention of harms.

Three: More should be done to show and explain company practices for managing risks to people across the full lifecycle of generative AI development, deployment and use. Better understanding of promising approaches as well as

shortfalls (and the dynamics underlying these shortfalls) should ground advances in guidance, industry standards and best practices

The Roundtable discussions surfaced a range of promising company practices aimed at identifying and addressing the human rights risks of generative AI tools. For example, company practitioners and experts working with companies spoke about how they conduct product-specific human rights assessments, the results of which become inputs into business decisions about the development and deployment of technologies. One speaker shared the innovative practice of blending human rights assessments with scenario planning methodologies to grant companies the broadest possible lens on potential use cases of generative AI tools. Many, if not all, companies present also appear to engage in "Red Teaming" – the practice of engaging in adversarial technical testing and/or discussion to rigorously challenging plans, policies, design choices and assumptions concerning the risks of products and innovations. Some even do this in participatory ways, inviting external stakeholders, including civil society members, to catalogue potential use cases or test if they are able to "jail break" AI models.

Certain companies and leading academics (often in partnership) are already exploring how to advance transparency and public understanding of risks associated with AI products. One example presented is the use of "Model Cards", which are similar to a nutritional label for AI products that include information on an AI model's purpose, what data was used in training, what guardrails were applied, what bias assessments were conducted, and potentially the use cases considered and mitigations applied. Other examples mentioned were the labelling of synthetic data to enable end users and even regulators to better spot when they are interacting with AI-generated text, image, video or speech, as well as "watermarking", which embeds a digital watermark into digital media to flag synthetic audio and images.

Business participants – echoed by civil society, academia and government stakeholders present – also reflected on the need for increased investment and emphasis by companies on enhancing AI risk management processes, many of which are still nascent and need improvement. The key shortfalls and challenges raised in these discussions included: *fragmentation* into different and disconnected silos of work within companies (such as human rights, ethics and compliance functions) which results in diverse impact assessment frameworks and conclusions that complicate companies' efforts to identify and address shared critical issues effectively and efficiently; *speed* and the rapid pace of innovation and product development, which make it more difficult to pinpoint when and on what iteration of a technology to focus risk identification and mitigations; and *presumptions* that mitigations will be only technical in nature as opposed to being a function of how and under what conditions technologies are released and deployed.

Moreover, as is often the case in other industry contexts, certain key aspects of good risk management and business respect for human rights as set out by the UNGPs are either not robustly implemented or just not well understood and explained at an industry level or among external stakeholders. Examples include limits in the quality of companies ' engagement with affected stakeholders across the full life-cycle of product design, development, deployment and use; insufficient tracking and communicating about post-deployment risks/impacts and the effectiveness of mitigations; and very little attention to ensuring that those experiencing harms have access to remedy.

In sum, the Roundtable discussions signalled that there is a considerable body of company practice to assess and address risk to people connected to generative AI. However, there is currently no clear method through which industry-wide and cross-stakeholder understanding of these practices can be rapidly surfaced, constructively interrogated and improved. Without this, it will be difficult to shape standards, market incentives, and corporate conduct in ways that actually deliver better outcomes for affected stakeholders.

Four: The complexity of the generative AI eco-system must be accounted for in public policy and governance responses. Moreover, responses should be informed by extensive multi-disciplinary and multi-stakeholder cooperation.

Participants emphasized coupling UNGPs-aligned State and company action to address generative AI risks with a firm understanding of the generative AI ecosystem (sometimes referred to as the value chain or stack). A novel reality with generative AI technology is that, while private actors and capital dominate, the eco-system includes public sector, academic, and civil society actors, as well as individual developers. For example, much research in AI has been developed

within academia, and there is also a culture of open-source development. This decentralization and non-commercial innovation are arguably positive and exciting, but also carry risk that irresponsible or ill-meaning actors can use publicly available innovations to exploit and abuse vulnerable individuals at scale, quickly and often anonymously.

Cultivating AI expertise within governance bodies is key. This would ensure that regulators and enforcement agencies make informed decisions about risk management measures and to e.g., develop testing frameworks. The core message is that a too-narrow view of the private and public actors interacting across this eco-system will lead to missing opportunities to drive systems-wide changes in practice that result in better human rights outcomes at scale.

Often such generative AI innovations affect several areas, and as a result several regulatory authorities may need to be involved in their governance. Cooperation between firms, competition authorities, intellectual-property offices, national standardisation bodies, and data protection authorities, among others, is crucial. Another workstream of B-Tech has been setting out key criteria of how the "smart mix" of voluntary and mandatory measures can be applied for UNGPs-aligned tech regulation. One of the core messages is that the process-oriented nature of human rights due diligence under UNGPs can help to future-proof regulation, better than other types of regulation: for example, those that build on a list of non-permissible artefacts.

As a basis to prompt coherent and, where appropriate, collaborative human rights due diligence, tools that clarify the structure of digital technology value chains can be helpful. One such tool is the Across the Stack Tool: Understanding HRDD Under an Eco-System Lens published by the Global Network Initiative and BSR.

Five: It is critical to understand what robust human rights due diligence and remedy for harms look like for specific generative AI technologies and use cases. This will be key to focusing practitioner and public policy action on practical steps that can be taken now to advance responsible development and deployment of generative AI.

The Roundtable painted a rich and broad picture of the social and human rights opportunities and risks that accompany the development and deployment of generative AI technologies. Moreover, there is a clear conviction about the need to make international human rights standards, including the UNGPs, the normative reference point for State, business, investor and civil society action aimed at maximising those opportunities and minimizing risks to the most vulnerable stakeholders.

Participants reflected that work is needed to move from this high-level discussion to more specific examples and evidence of what it means in practice for robust corporate human rights due diligence and engagement to provide for or enable access to remedy. Some made the concrete suggestion that the B-Tech Project should lead and/or call on others to facilitate working sessions and publish short practical insights around a small but diverse set of specific generative AI applications and use case scenarios: hypothetical, real or a hybrid of the two.

Looking Ahead

This Roundtable was the first expert, multi-stakeholder meeting for UN B-Tech's generative AI project. Informed by this engagement, the B-Tech team will structure its iterative process of research and engagement, with the target of producing near-term conclusions and recommendations in late 2023.

Participants

Alex Walden, Google
Alyson Finley, US Dept. of State
Annie Ludwig-Lehman, US Dept. of State
Aviv Ovadya, Berkman Klein Center
Ben Pitler, UN Human Rights
Christian Mogensen, Denmark
Dennis Hirsch, Ohio State University
Eileen Donahoe, Stanford GDPi
Elonnai Hickok, GNI
Emily Bender, University of Washington
Eugenio Garcia, Brazil
Fatemeh Khatibloo, Salesforce
Giulia Geneletti, European Union
Hannah Darnton, BSR
Helen Harris, Amazon
Iain Levine, Meta
Isabel Ebert, UN Human Rights
Isabella Tomás, Austria
Jason Pielemeier, GNI
Josh Cohen, Apple
Kim Malfacini, OpenAI
Kip Wainscott, Snap
Lene Wendland, UN Human Rights
Lindsey Andersen, BSR
Margaret Taylor, Salesforce
Marie Alnwick, Canada
Mark Hodge, Shift
Mark Silverman, ICRC
Michael Kleinman, Amnesty Tech
Michael Yuan, Zoom
Miranda Sissons, Meta
Nathalie Stadelmann, UN Human Rights
Pamela Wood, Hewlett Packard Enterprise
Rashad Abelson, OECD
Remedios Gómez Arnau, Mexico
Sarah Luger, Orange
Sneha Shah, Amazon
Stephanie Seale, UN Human Rights
Steve Crown, Microsoft
Suba Jayasekaran, Shift
Ursula Rojas, Mexico
Vyoma Raman, Berkeley Human Rights Center
Yannick Heiniger, Switzerland