UNITED NATIONS HUMAN RIGHTS OFFICE OF THE HIGH COMMISSIONER

B-Tech

UN B-Tech Generative AI Human Rights Summit:

Advancing Rights-Based Governance and Business Practice

30 November 2023

SUMMARY NOTE

OVERVIEW

The UN B-Tech Generative AI Summit gathered stakeholders from business, civil society, the UN, government, and academia to explore the human rights implications of generative AI technology and the role of the UN Guiding Principles on Business and Human Rights (UNGPs) in ensuring that generative AI is developed and deployed responsibly. The Summit was grounded by B-Tech's recently released set of papers on generative AI, human rights and the UNGPs.

The foundational paper, "Advancing Responsible Development and Deployment of Generative AI," explores the intersection of business, states, generative AI, and the UNGPs. It is supported by two supplemental papers. "Responsible AI and Human Rights: An Overview of Company Practices" explores the state of implementing a human rights-based approach to responsible AI inside companies. While a second supplemental paper, "Taxonomy of Human Rights Risks Connected to Generative AI" lists the key human rights risks associated with generative AI as an example of how human rights can form the foundation for assessing risks to people and society.

About 100 participants attended the Summit from a variety of stakeholder groups: 22 from civil society, 17 from international organizations, 15 from academia, 16 from companies, 12 from states, 7 from business-oriented groups and consultancies, and 7 from investors.

Throughout the summit, participants learned about the key elements of a human rights-based approach to generative AI. They also gained insights into the practical integration of the UNGPs and a human rights-based approach to generative AI within companies and had an opportunity to discuss and brainstorm how to address key challenges.

Speakers noted many human rights risk linked to generative AI have also been associated with earlier iterations of AI, though in some cases generative AI has altered the scope of these risks. Speakers highlighted that many of the processes that companies have created to assess and address AI-related risks more broadly should be applied to generative AI. Moreover, focusing on generative AI specifically enabled speakers and participants to engage in more applied discussions.

SESSIONS

- Remarks from the UN High Commissioner for Human Rights, **Volker Türk**, and questions and interventions from **Peggy Hicks**, director of the Thematic Engagement, Special Procedures and Right to Development Division for the UN Human Rights Office.
- **Four panel sessions** then highlighted perspectives from a diverse range of speakers and discussed the value of the UNGPs and a human rights-based approach to generative AI from theory to practice.

***The sessions presented the **B-Tech papers foundational on human rights and generative AI**, explored the **value and practical use of the UNGPs** and existing human rights-based approaches to **identify, assess, and address the impacts of generative AI**, and established a foundation for productive discussions about how to **implement the UNGPs for general purpose AI** more broadly.*

HIGH COMMISSIONER'S OPENING REMARKS

- The High Commissioner called for attentive AI governance with a focus on people's rights, the importance of this work in the context of elections, the spread of disinformation, and unequal access to technology and resources around the world.
- Generative AI exemplifies a paradox—we need such technological advancements to address global challenges, but that same technology comes with significant risks to human rights. There needs to be a coherent and concerted effort by both governments and companies to address human rights risks linked to generative AI.
- Particularly in a year that will see over 70 elections around the world, the potential for generative AI to expand the scale and scope of civic mis/disinformation is a serious challenge and will require collaboration to address.
- We need more strategic foresight to think through longer-term risks and we need to make human rights due diligence more efficient.
- The UNGPs and OECD guidelines alone are not sufficient to address the challenges of generative AI because they do not cover the potential for misuse by state and criminal actors.
- Currently we see a cacophony of policy initiatives that have fragmented the space when a global approach is needed. Regulatory and multilateral frameworks that are coherent and aligned with international human rights norms are needed due to the cross-border nature of the impacts.

RESPONSIBLY ADVANCING GENERATIVE AI: THE VALUE OF THE UNGPS

Presentation of the B-Tech papers and exploring both the normative and practical value of grounding generative AI development in human rights.

Speakers: H.E. Claudia Fuentes Julio (Permanent Representative of Chile); Lene Wendland (OHCHR); Rashad Abelson (OECD); Alex Walden (Google); Benjamin Chekroun (Candriam); Isedua Oribhabor (AccessNow); Mark Hodge (Shift)

Insights: The primary challenge for states lies in adapting to the rapidly evolving technological landscape, as their existing structures struggle to keep pace. To address this, the development of AI should prioritize the welfare of people, uphold human rights, and ensure security, contributing to the overall well-being of individuals. Effective regulations and policies within this context necessitate inclusive discussions involving various stakeholders both internally and with the business sector to address the complex issues surrounding AI development.

1) The international human rights framework provides an existing, well-defined, and holistic architecture against which the state and other actors can evaluate the risks of AI. This is also the only existing architecture available at a practical, institutional level and a normative level that offers such a useful entry point related to generative AI governance.
2) The UNGPs outline a smart mix of measures like regulations, guidance and incentives supporting domestic and multilateral efforts to advance accountability. These expectations, set up in 2011, are still highly relevant and applicable, including to generative AI.
3) States must create regulatory mechanisms that keep pace with technological advancements, that provide effective guidance and capacity building to respect human rights, and that use authoritative corporate transparency mechanisms. States must make it the core of development of generative AI to build competence and expertise of relevant agencies and administrative advisory bodies to act in this space.
4) A major gap in upcoming AI-specific frameworks is that they only focus on some actors. Most of the standards focus only on AI developers and end users in the value chain, but do not cover the rest of the value chain: finance or investing in development of AI, hardware manufacturers, microchip manufacturers, digital infrastructure providers, data collectors, etc. Broader frameworks like the OECD guidelines and the UNGPs cover all actors in the value chain with an objective to bring those together in a coherent way taking into account national and regional frameworks in play.
5) States and companies planning to build operational guidance on responsible business conduct in AI and its value chain must: apply international HR instruments to this very

specific context; map all the actors in the value chain that bear this responsibility, and; conduct thorough due diligence in their operations.

6) AI is something that is the responsibility of all the industries (financial services, health, manufacturing, etc.) where technology is used. The UNGPs are a foundation for all the industries on how companies should think about these issues. They are the common framework that we all understand and agree upon.
7) We already know what the UNGPs' expectations vis-a-vis the duties of States and the responsibilities of companies, and that should be the starting point.
8) Human rights are the only way we can have a level playing field. Understanding the potential and already existing harms of generative AI requires robust human rights due diligence that paves the way to anticipate harms and develop tools to tackle them.
9) Effective stakeholder engagement is crucial and requires transparency from the beginning. Civil society and other stakeholders must be included in the development stage rather than in the deployment stage because feedback is really limited when they are brought in at a later stage. For expertise and feedback, companies must take an overarching practice of involving all actors, regional and global, within their consultations. All the conversations about regulatory and voluntary frameworks should involve the UNGPs as the foundation.
10) Transparency is the most crucial aspect that investors want to see in companies. They want to analyze companies based on the UNGPs and assess whether they follow rules and regulations or not. Companies have a wide spectrum of responsibilities, but the companies closest to the development of AI algorithms must realize how dangerous the outcomes or implications of their technology are and be transparent about it.

A Human Rights-Based Taxonomy for Generative AI

Presentation of the B-Tech taxonomy paper and discussion on the nature of generative AI risks.

Speakers: Angela Müller (AlgorithmWatch); Alex Warofka (Meta); Lindsey Andersen (BSR)

Insights: The responsible AI field has created new risk / harm taxonomies for generative AI, most of which are not linked to the human rights framework. This is partially due to a lack of knowledge about the international human rights framework, the UNGPs, and how they can be utilized.

1) B-Tech's Taxonomy paper is a step toward addressing that issue by showing how human rights can form the foundation for thinking about risks to people and society associated with generative AI upon which other approaches can be added.
2) There are many practical reasons to use human rights as a foundation: they are universal, there are existing state and corporate responsibilities, they provide an established list of

impacts to assess risks against with definitional clarity and specificity, and they are adaptable to new contexts.

3) With generative AI, the responsible AI field has been dominated by a focus on existential risks at the expense of focusing on existing and near-term harms. Looking at human rights risks is one way of addressing this.
4) Generative AI can have a huge range of rights impacts beyond the obvious privacy/data protection and intellectual property risks. This includes impacts to free expression, access to justice, non-discrimination, labor rights, and many others.
5) Many generative AI related risks are new permutations of known AI risks (e.g. biased outputs) and it's important we don't lose sight of other risks associated with greater automation / integration of AI across society amongst the generative AI hype. Key new risks include the scale of potential harm due to the availability of tools, and the challenge of minority language performance. These issues are cross-industry and will require multi-stakeholder collaboration to address.
6) Although developers have a role to play in addressing human rights risks, many of the most serious risks are in the use of generative AI systems. Deployers, including non-tech companies that are increasingly adopting AI into their operations, need to also ensure they assess and address risks.

APPLYING THE UNGPS TO GENERATIVE AI IN PRACTICE

Discussion on how companies are applying the UNGPs to generative AI and the challenges they face in doing so.

Speakers: Pamela Wood (Hewlett Packard Enterprise); Ramsha Jahangir (GNI); Gayatri Khandhadai (BHRRC); Chris Sharrok (Microsoft); Hannah Darnton (BSR)

Insights: Companies take a wide range of approaches to assessing and addressing risks associated with generative AI. This includes a wide variety of risk/impact assessment models– both technical and non-technical / issue 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 figured out.

1) As companies with responsible AI principles steadily work to operationalize them, risk assessments are increasingly being integrated into existing AI product development processes across the product life cycle. Some are based on human rights, and some are not. Standalone assessments done at a particular moment in time are also still prevalent, and it's unclear the extent to which companies are assessing their impacts on an ongoing basis.

2) Companies tend to develop their own models for risk / impact assessments that are informed by public standards and best practices to various degrees. Because the rapid evolution of generative AI outpaced the responsible AI field's ability to develop and communicate best practices, some of the assessment models companies have pursued for generative AI have been ad hoc and experimental.
3) Best practices for development, testing and deployment of responsible AI at a company level includes a multi-disciplinary approach bringing together researchers, engineers, and policy experts. This should put the UNGPs at the heart of the conversation and ensure that human rights principles are considered, used, and applied throughout the company.
4) 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 trade-offs / tensions.
5) Explainability and Transparency are two of the core principles where partnerships, processes and products must be assessed for risk assessment. It is key to emphasize collective assessments where it is required for partners to disclose AI that they are developing, as well as any AI they are sourcing or using internally for solutions.
6) Companies must ensure that even before the design phase, fundamental issues with the data that is used to train AI models are addressed. There are concerns regarding privacy, consent and ethics of the data collected to train these models. Therefore, the due diligence process must assess collection of data, makeup of training datasets, presentation of data and feedback loops.
7) Ensuring collective coordination with parallel teams is essential to better articulation of policies. Engagement must involve not only policy teams and engineering teams, but also marketing and sales teams, because monetization is the seed of most of the problems. It is important to have conversations about human rights, choosing what to monetize, when to monetize, who to monetize from etc.
8) Multidisciplinary capacity building at company level is essential for the intersection where engineers are made to think about human rights. 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.

BREAKOUT SESSIONS

Two breakout sessions allowed participants to explore and think through potential solutions to the challenges in taking a human rights-based approach to generative AI. The sessions covered a wide range of insights from all participants focusing on prioritization, responsibility, risk assessment and effective mitigation, leverage, stakeholder engagement and remedy for rightsholders.

PRIORITIZATION

Key Questions:

- *How can companies prioritize taking action based on severity of potential harm to people?*
- *What does meaningful human rights due diligence look like on the timescale of rapidly changing/evolving technologies like generative AI?*

- Assessing the severity of certain harms is challenging. For example, when a model produces an output that reflects harmful stereotypes about a vulnerable group, the potential harm can range and often becomes more severe with scale. This is especially true with general purpose models vs. a specific application. It's important for companies to start by addressing the most severe harms, not the easy fixes, and to be transparent about their approach to addressing risks.
- Prioritization is challenging with such a fast-moving technology. By the time you've completed an assessment, things have already changed. We need to rethink the notion of HRIAs and shift toward an ongoing HRDD model. It's easy for companies to end up reactively "putting out fires" rather than proactively anticipating and addressing risks before harm occurs—especially "non-tech" companies who are responding to rather than leading developments in generative AI.
- Ongoing, integrated HRDD is key. We should take inspiration from other risk assessment approaches, such as cybersecurity approaches where risk spotting happens continually. The staff developing and managing the deployment of generative AI products also need to be trained to identify relevant risks. Required documentation of risk identification throughout the product development lifecycle can drive more effective mitigation, and products can have built- in review / feedback mechanisms. Companies will also need to appropriately employ staff and resource specialized teams to think about risks. These teams need guidance about what HRDD looks like at different points in the lifecycle and help translating human rights language to technical audiences.
- Diverse perspectives are key for identifying difficult to predict risks. Generative AI is in its infancy and many of the risks aren't known yet. The pace of evolution also means there are constantly new use cases. Understanding risks is personal and relational and based on context. The wider the backgrounds / contexts in the room, the more comprehensive the risk assessment will be. Ongoing communication with stakeholders is key.
- Tracking of mitigation measures needs to improve. Companies need to monitor the implementation of mitigation measures related to different risks to see what works and then adjust. If you can't come up with any mitigation measures and the risks are severe, then it's important to consider whether the product should be available / used at all.
- Data enrichment labor rights risks are underexplored. There is a human rights management gap on this issue inside of companies, but there are many lessons learned from the broader supply chain labor rights field that can be drawn upon. Cross-industry collaboration on this issue will also be important.

Responsibility

Key Questions:

- *What are the roles and responsibilities of technology companies developing generative AI, non-technology companies deploying generative AI, regulators, and investors?*
- *How should these overlap and where are they distinct?*

- At a company level, a structural process approach should be used to define what review processes should be put in place for looking at new products and product features that use or implement AI as a technology. At an operational level, trust and safety teams must understand what the relevant human rights obligations are and how they can be embedded in processes to educate users.
- Downstream due diligence is extremely important when it comes to these sectors and end use. The tech sector is a great example of why we look at the whole value chain. Partnerships, processes, and products– all 3 must be assessed. Partners should disclose the AI that they are developing, as well as any AI they are sourcing or using internally for solutions.
- Companies must ensure that human rights principles are considered, used, and applied throughout the company when the technology is built, tested, and deployed. This can be done through various means, including capacity building programmes, human rights resources, transparency notes, guidance for engineers, impact assessment templates etc.
- There are limits to single company efforts to address risks to developing and deploying generative AI tools. As a single company, it is hard to address risks across the entire value chain. As a deployer of generative AI technology, you don't necessarily know all the risks posed by the technology as a whole or by specific foundation models. Developers of foundation models do not know all the risks either, and so it is important to involve actors from across the entire value chain and build structures across the value chain that don't involve only a single actor.

Risk Assessment and Effective Mitigation

Key Questions:

- *What are the quality criteria for a generative AI risk assessment?*
- *How can we achieve effective mitigation (both technical and non-technical) in addressing harms associated with generative AI?*
- *What are the limitations of mitigations being taken today?*

- Governance of risk mitigation measures is crucial: it requires effective monitoring systems, lists of high-risk areas, testing of crisis protocols and rapid human rights impact assessment and due diligence measures where necessary.

- Existing risk assessment structures can be used to integrate considerations, including with stakeholder feedback.
- Aiming for high quality risk mitigation procedures is key, paired with stakeholder feedback. While financial and/or geopolitical factors might threaten to overrule risk mitigation measures, it is vital that structural safeguards are cemented into corporate governance structures to be applied under all circumstances for safe AI systems.
- Business cannot outsource its responsibility for risk mitigation to third parties: ensuring in-house buy-in is key.
- Identified gaps in risk mitigation measures should be closed in prioritization by the severity of risk to people: the nature of the risks (salient vs. emerging risks) will determine which measures are needed and how they should be rolled out (centralized vs. rapid human rights impact assessment).
- To continuously improve risk assessment and mitigation measures, allowing scrutiny of assessments is key, e.g., through red teaming exercises with competing in-house and adversarial teams.
- Measures need to consider cumulative impacts on human rights over time, e.g., when different AI products are interacting with each other.
- Experiments with models for robustness of AI systems, not only explainability and transparency, are important.
- An iterative process of prioritization of addressing risks to people needs to take into account which strategies can be applied to fix and determine the effect of an AI model; for certain models, speed of mitigating measures is key, whereas constant moderation and assessment of risk might be necessary for large language models.
- Different perspectives and expertise are necessary across collaborative governance structures to provide for agility in management responses to emerging risks to people

LEVERAGE

Key Questions:

- *What does leverage look like in the context of generative AI?*
- *What leverage do different actors have - civil society, business, policymaker, etc. and what are the limits?*

- Some states may be less willing to fulfill their human rights obligations, but every state has ratified some human rights conventions and they are all willing to have at least a certain openness. The international human rights framework and mechanisms can be used to bring these countries into these discussions. In these states, companies can use leverage where regulations are not a top priority.
- Controls should not be concentrated in one jurisdiction, especially if we take the value chain approach. In the AI sector, so many of the foundational models are located in

various countries. We have seen the power of market access and leveraging the value chain in the garment sector, which has led countries like China, UAE, India to adopt the OECD guidelines even when they are not members.

- One of the big challenges is that the majority of states are not moving forward due to lack of capacity. This divide between technologically advanced nations and others who lack the infrastructure and capacity will only grow. Therefore, we need alignment between different actors on the issue of capacity building on new technological advancements, and we need to use that as a leverage to bring the state to the table.

Stakeholder Engagement

Key Questions:

- *What should meaningful and sustainable stakeholder engagement look like for generative AI?*
- *What are the challenges in achieving this and how can they be addressed?*

- Key stakeholders include governments and regulators, tech companies, international organizations, civil society organizations, academia, and investors.
- The challenges for meaningful engagement with companies on generative AI vary among stakeholders. Common ones include limited time and resources (particularly for civil society organizations), limitations in influence (e.g., governments in smaller states), and the rapid evolution and development of the technology in question (e.g., for governments looking to regulate, or international organizations setting standards).
- For effective stakeholder engagement, companies must look for avenues to connect with local NGOs working in contexts which are difficult to navigate. Initiatives like GNI provide those avenues.
- Civil society is a great resource for multi stakeholder engagement but the burden of monitoring how a product or service is used should not fall upon already under-resourced and overworked civil society organizations. Companies must work towards developing a partnership and ensuring that this partnership leads to something.
- Multi-stakeholder engagement requires cultivation of relationships over time. This ensures that you do not have the same conversations over and over, and that communities do not have to flag every instance or violation over and over. Effective multi stakeholder engagement ensures that we learn over time and apply learnings from past experiences to similar scenarios.

Remedy for Rightsholders

Key Questions:

- What does remedy for generative AI-related adverse human rights impacts look like and
- What is the role of different actors and what are the challenges?

- The key question on where to seek remedy and who you go to depends on the type of harm. One key challenge is figuring out who is responsible for the provision of remedy in the context of a general-purpose technology. It is very challenging in the AI space because it is not clear who is in charge or who is responsible: the developer, the user or the deployer.
- The four B-Tech papers on remedy and technology can be a helpful baseline resource. Remedy for tech-related harms is especially challenging and has long been underexplored. Many of the known challenges for effective remedy in the tech sector also apply to generative AI — e.g., assigning attribution for harm and identifying who is responsible for providing remedy, what remedy can look like in practice, etc.
- Because the harms of generative AI are cross-society, an ecosystem approach is needed, with different actors in consultations involving non-state actors as well. Practical guidance and resources are needed about how different actors in the generative AI ecosystem can provide access to remedy.
- There are existing legal channels for remedy, as well as channels for remedy in emerging tech regulation. We need to figure out how the provision of remedy for generative AI-related harms could further align with existing legal processes.
- Human rights due diligence must connect back to the question of accountability. It must come back to the question of remedy and redressing of grievances and must be conducted accordingly.
- In practice, any remedy must have an apology as a starting point. It should lead to something and that something should be communicated to people who are affected or suffering. It must be followed by clear and accurate communication about what went wrong, who was responsible and how it can be fixed.

The *B-Tech project* is the UN Human Rights flagship initiative that aims to prevent, address, and remedy the vast array of actual or potential human rights harms related to digital technologies, by providing an authoritative and broadly accepted roadmap for applying the *UN Guiding Principles on Business and Human Rights (UNGPs)* to the design, development, and use of digital technologies.

For additional information about the B-Tech Project, including past and upcoming activities, please visit the B-Tech Portal or contact : ohchr-b-techproject@un.org