B-Tech

UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

Advancing Responsible Development and Deployment of Generative AI

The value proposition of the *UN Guiding Principles on Business and Human Rights*[1]

Headlines and Recommendations from UN B-Tech Foundational Paper

November 2023

Context

Technological breakthroughs in the field of generative AI, coupled with the unprecedented speed and scale of uptake of new consumer tools and enterprise-facing applications have captured the public imagination. Aspirations of leveraging artificial intelligence to dramatically improve our lives suddenly seem much less fictional: whether helping individuals to reach new heights in creativity and productivity, bolstering industrial development, or uncovering solutions to shared challenges in the realms of healthcare and climate change.

And yet, it also seems more likely than ever that these same tools will be designed and used (or abused) in ways that erode individual freedoms, undercut livelihoods, reinforce inequalities, and undermine norms and institutions designed to uphold democratic values and protect human rights. In fact, evidence of adverse impacts on people from generative AI tools — whether stemming from in-built characteristics of these tools or from their misuse — are already being reported: for example increasing technology-enabled gender-based violence, the amplification of discriminatory racial and ethnic stereotypes, the supercharging of online disinformation campaigns or the creation of child sexual abuse material at scale.

Governments, civil society, academics, technologists, investors and business executives have all called for regulation to govern the design and deployment of generative AI systems to protect against harms and maximize their benefits. However, these initiatives have tended not to 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*.

About the B-Tech Generative AI Project

The B-Tech Generative AI project was established to raise awareness and facilitate exchange among key stakeholders and interdisciplinary experts and shape a comprehensive understanding about the role the UNGPs can play in governing generative AI responsibly. The Project aims to do this by:

- Clarifying the expectations under the UNGPs for companies developing and deploying Generative AI technologies and products in order to achieve common and more effective human rights risk management approaches across the industry.
- Spotlighting the growth and maturation of existing company responsible AI approaches, as well as academic research and civil society advocacy that have all laid important foundations for addressing the risks to human rights associated with generative AI.
- Informing the debate about policy options for managing human rights risks related to the development and deployment of generative AI, including through mandatory and voluntary measures.
- Complementing parallel efforts to embed the international standards of business conduct into AI governance such as the work being led by the OECD[2].

There are, of course, limits to what frameworks focused on responsible business conduct and corporate accountability can tackle. They are not a panacea. Many issues will require other tools, laws, enforcement regimes and multi-lateral solutions. That said, advancing responsible business conduct, as well as being valuable in its own right, can serve as one powerful tool in minimizing the likelihood of the most egregious deployments of generative AI proliferating.

[1] The UNGPs are the global authoritative standard for preventing and addressing business impacts on people, unanimously endorsed by the Human Rights Council in 2011. The UNGPs sparked an unprecedented regulatory dynamic for issue-specific and overarching due diligence legislation; civil society in campaigns, complaints and litigation; companies, and more recently investors, building and implementing good practice principles, codes and guidance aligned to the UNGPs; and reporting standards, see also: An Introduction to the UN Guiding Principles in the Age of Technology, a B-Tech foundational paper.

[2] The OECD is working to apply and adapt international standards on responsible business conduct to actors in the AI value chain. This work is being led by a multistakeholder Network of Experts, which includes the UN B-Tech Project, and is overseen by government delegates in the OECD Working Party on Responsible Business Conduct and the OECD Working Party on AI Governance. The project is systematically building towards the development of concrete and practical recommendations for AI actors under an overarching due diligence framework by first mapping out and consolidating recommendations, terminology, and risk scopes from existing AI-specific and generic risk management frameworks (e.g. the OECD Due Diligence Guidance for Responsible Business Conduct, the NIST AI Risk Management Framework, the G7 Code of Conduct for the Development of Advanced AI Systems, IEEE 7000 series, ISO 31000, and ISO/IEC 23894).

[3] This articulation is based on the depiction of a typical AI value chain, proposed by the OECD's Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI. By way of example: 1) Suppliers of AI knowledge and resources can include; Content creators; Data providers and data annotators; Investors; Digital infrastructure providers; Hardware manufacturers. 2) Actors in the AI lifecycle can include companies, States, research institutions involved in Planning & design of the system; Collecting & processing of data; Building & using the model; Verifying & validating the model Deploying the system, regardless of the distribution channel (including the distribution of open-source software); and Operating & monitoring the system; 3) Users/operators of the AI system can include Businesses, including financial institutions and businesses in the 'real' economy (e.g., manufacturing, purchases, and flows of goods and services); Individuals or other actors using AI for personal use, commercial, or research, activity; and States.

[4] According to the Ada Lovelace Institute, foundation models are "a form of AI designed to produce a wide and general variety of outputs, capable of a range of tasks and applications, such as text, image or audio generation (…) notable examples are OpenAI's GPT-3 and GPT-4, foundation models that underpin the conversational tool ChatGPT. Following the launch of large language model (LLM) interfaces (…) foundation models are more widely accessible than ever".

Headlines and Recommendations

HEADLINE ONE

Impacts on internationally agreed human rights should be the focus of State and company action to advance the responsible development and deployment of generative AI technologies

Key Messages:

- Human rights provide an existing, well-defined, and holistic set of outcomes against which States, companies, and other actors evaluate the risks related to generative AI.
- The international human rights framework has a developed architecture of international, regional and national institutions and processes which facilitate consideration of these issues and, in some instances, monitor and even enforce implementation.
- Focusing on international human rights has the merit of reinforcing existing, well defined State obligations and corporate responsibilities.

Recommendations:

To catalyse greater attention to applying a human rights lens to developing and deploying generative AI, B-Tech has developed a Taxonomy of Generative AI Human Rights Harms. The taxonomy shows clear connections between "risk examples" connected to generative AI across nine categories of internationally agreed human rights:

- Freedom from Physical and Psychological Harm
- Right to Equality Before the Law and Protection against Discrimination
- Right to Privacy
- Right to Own Property
- Freedom of Thought, Religion, Conscience and Opinion
- Freedom of Expression and Access to Information
- Right to Work and to Gain a Living
- Rights of the Child
- Rights to Culture, Art and Science

HEADLINE TWO

The UNGPs offer guidance on how to establish the multi-layered architecture of governance needed to address the conduct of private sector actors across the full generative AI value chain. This includes companies that are suppliers of AI knowledge and resources, actors in the AI system lifecycle, and users/operators of an AI system[3]

Key Message:

States should implement a "smart-mix" of regulation, guidance, incentives, and transparency requirements – all supported by policy coherence in domestic and multi-lateral efforts - to advance corporate responsibility and accountability for human rights harms.

Recommendations:

- States should enforce laws that are aimed at, or have the effect of, requiring companies developing and deploying generative AI technology to respect human rights, periodically assess the adequacy of such laws and address any gaps.
- States should provide effective guidance and associated capacity building to business enterprises on how to respect human rights when developing or deploying generative AI.
- Authoritative corporate transparency regimes from the corporate responsibility and accountability field should be used to complement technology specific transparency requirements.

	– States — especially those States home to market-leading companies at the core of developing AI systems — should build the competence and capability of relevant agencies, administrative supervisory bodies and officials. – States should pursue multi-lateral action focused on the protection and respect of human rights: to spread best practices between States minimize the risks of States pursuing their own interests at the expense of building dignity and respect into the heart of generative ai development and deployment. – States — whether part of national, regional or international initiatives — should establish and sustain stakeholder engagement with companies, civil society and especially affected stakeholders to learn about risks, impacts and challenges/opportunities to advance meaningful generative AI risk assessment and mitigations.
Key Message: Regional, national, international and industry-led initiatives focused on advancing responsible generative AI should use or align to the international standards of business conduct. This means, in particular, integrating a true risk-based approach to identifying and taking action on impacts that: a) Uses severity of risks to people to prioritize impacts for attention; and b) sets expectations of companies across the generative AI value chain commensurate with the nature of their involvement (causation, contribution or linkage) with human rights risks and impacts	**Recommendations:** – Reaffirm and ground policies in States' existing duty to protect and businesses' *Corporate Responsibility to Respect Human Rights* as laid out by the UNGPs and OECD Guidelines. – Integrate risk-based prioritization based on severity of risks to people as well as the cause, contribution, linkage "Involvement Framework" into legislative texts, technical standards and guidance. – Establish multi-stakeholder dialogue to deepen appreciation of what a full value chain approach to addressing human rights risks means in practice.
Key Message: Greater urgency is needed towards ensuring effective judicial and non-judicial access to remedy for individuals whose human rights are harmed by the development or deployment of generative AI.	**Recommendations:** – All stakeholders should collaborate to establish processes for understanding the experience and perspectives of impacted or at-risk individuals or groups about what meaningful remedy for generative AI harms means in practice. – States should ensure access to judicial remedies where individuals may have been harmed by the development or deployment of generative AI technologies. – States, companies, civil society experts and affected stakeholders (or legitimate representatives) should work together on how to establish non-judicial routes through which people may seek remedies for specific human rights related harms connected to generative AI.

HEADLINE THREE

Implementation of thorough human rights due diligence by companies developing foundation models[4] will provide an important basis for risk management across the generative AI value chain. Clear and regularly updated guidance on what constitutes best practice is required, building on company practice and informed by civil society and relevant experts. Emphasis should be placed on key practices, which are currently under-emphasized in regulatory proposals and technical standards.

Practice 1: Boards and executives identifying the extent to which the company's business model and strategy carry inherent human rights risks, and taking action to address this.	**Proposed Next Steps:** Multi-stakeholder deliberations, and case studies of good practices focused on: – Boards identifying as part of initial business model design and strategy – and in any changes to these - the inherent human rights risks that flow from these and ensuring that the company has systems and plans to address these. – Senior leaders establishing and implementing commitments to release or scale the capability of foundation models in a responsible manner, including evaluating which situations might merit adopting an approach akin to the "precautionary principle". – The best ways to establish and sustain corporate cultures that reward the identification of risks and adverse impacts, including by ensuring that individuals feel able to raise concerns without fear of retribution. – Ensure that the company has in place the right competence, resources and processes to hear, and act on, the perspectives of especially affected or at-risk stakeholders.
Practice 2: Embedding human rights risk assessment – focused on all human rights with any necessary prioritization being based on severity - into the product-oriented working methods and cultures of technology companies developing foundation models.	**Proposed Next Steps:** Multi-stakeholder deliberations, and case studies of good practices focused on: – Identifying the most impactful moments at which a company should assess the actual and potential human rights impacts that it could become connected to due to the development or deployment of its foundation model. – Creating tools, assessment methodologies and training that support an evaluation of impacts based on the full range of internationally agreed human rights and prioritization of impacts for attention based on their scale, scope and irremediability. – Mechanisms to allow external stakeholders to understand, appreciate and inform the quality of human rights risk identification and prioritization practices.
Practice 3: Evaluating "technical" mitigations with a focus on people in situations of vulnerability or marginalization.	**Proposed Next Steps:** Multi-stakeholder deliberations, and case studies of good practices focused on: – The extent to which existing quantitative methods used by companies to evaluate mitigations can feasibly and responsibly be leveraged to offer insight into differential risks to distinct vulnerable groups. – Identifying how qualitative methods can offer feedback loops from affected stakeholders about the effectiveness, and indeed risks of technical mitigations for groups in situations of vulnerability. – Innovating collaborations that bring academics and civil society into the evaluation of effectiveness of mitigations but without compromising their independence and safety, or legitimate commercial interests of companies.

Practice 4: Creatively building and using leverage to address "residual risks" and enable remedy for harms.	**Proposed Next Steps:** Multi-stakeholder deliberations, and case studies of good practices focused on: – Responsible use policies, terms of use in contracts, guidance and enforcement with initial points of emphasis on understanding the impact of these policies and practices i.e., in what ways do they make a difference and what can be improved; and how to monitor third party practices without violating the rights of data subjects. – Know Your Customer assessment and follow-up with initial points of emphasis on: how to use indicators capable of evaluating customers' commitment and competence to manage risk and impacts from their own use of the company's foundation model and products; and strategies and tactics for building and using leverage when a customer is considered to be high risk from a human rights perspective. – Collective action with peer competitors (including smaller market entrants), value chain companies, civil society and international organizations with initial points of emphasis on ensuring that civil society and perspectives of affected stakeholders have an equal seat at the table; and targets and accountability measures that go beyond pledges and principles to focus resources on delivering results. – Leverage for remedy including providing proactive support to strengthen customers' redress mechanisms; identifying where industry-level mechanisms at the deployment level might be necessary; and "enabling remedy" in specific instances of harm.
Practice 5: Engagement with affected stakeholders and civil society experts across the full cycle of human rights due diligence, and as part of enabling remedy for harms	**Proposed Next Steps:** Multi-stakeholder deliberations, and case studies of good practices focused on: – Companies developing foundation models establishing the necessary internal commitment, capacity and culture to engage with affected stakeholders and civil society representatives across all phases of the AI development life cycle. – The meaningful integration of affected stakeholder perspectives within industry-led responsible generative AI collaboration, with particular attention to removing logistical barriers to participation, diversity among participants and investing in the technical capacity of communities to engage. – Companies developing foundation models using "leverage for engagement" by taking a proactive role in advocating for more formalized mechanisms, and possibly funding options, for at-risk stakeholders to convene and advocate for their rights with relevant actors across the generative AI value chain

Looking Ahead

The insights and recommendations laid out in this paper and supporting supplements from the first phase of the B-Tech Generative AI project have been released to support multi-stakeholder dialogue and collaboration that advances UNGPs-consistent public policy, regulation and business practice. The findings, and responses to them, will inform B-Tech ongoing work on generative AI in 2024.

UN Human Rights invites engagement from all stakeholders as we move into the second phase of this B-Tech initiative. Please contact us if you would like to engage with our work, including if you have recommendations for practical tools, case studies and guidance that will advance company, investor and State implementation of the *UN Guiding Principles on Business and Human Rights* in the context of Generative AI development and deployment

Ohchr-b-techproject@un.org

Acknowledgements

The UN B-Tech team expresses thanks to all the experts and stakeholders that provided input into this foundational paper such as representatives from the OECD Centre for Responsible Business Conduct, the Global Network Initiative, BSR and Shift. The team is especially appreciative to Mark Hodge, Vice President of Shift, the lead author of this paper.