Advancing Responsible Development and Deployment of Generative AI

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

A UN B-Tech Foundational Paper

November 2023

B-Tech

UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

Introduction

The *UN Guiding Principles on Business and Human Rights*[1] (UNGPs) — the global authoritative standard for preventing and addressing business impacts on people — can add considerable value to efforts aimed at achieving responsible development and deployment of generative artificial intelligence[2] (generative AI) foundation models, applications and products. Based on six months of research and consultation, the UN Human Rights B-Tech Project has identified three broad headlines and associated practical recommendations for how lawmakers, standard setters, businesses and civil society can leverage the UNGPs to foster governance and business practices capable of tackling human rights impacts and risks of generative AI.

HEADLINE ONE (see page 5)

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.

Rights-based approaches focus attention on specific harms to people's dignity and equality. They also provide agreed norms for assessing and addressing impacts, along with a shared language that can facilitate understanding and engagement across diverse stakeholder groups. To catalyse greater attention to applying a human rights lens to developing and deploying generative AI, B-Tech has developed a Taxonomy of Human Rights Risks Connected to Generative AI.

HEADLINE TWO (see page 7)

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]. A UNGPs-informed approach emphasizes that:

- 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.

[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 defines generative AI as "a form of AI model specifically intended to produce new digital material as an output (including text, images, audio, video, software code), including when such AI models are used in applications and their user interfaces. These are typically constructed as machine learning systems that have been trained on massive amounts of data. They work by predicting words, pixels, waveforms, data points, etc. that would resemble the models' training data, often in response to prompt."

[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.

- Regional, national, international and industry-led initiatives focused on advancing responsible generative AI should 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.
- Greater urgency is needed to ensure effective judicial and non-judicial access to remedy for individuals whose human rights are harmed by the development or deployment of generative AI.

HEADLINE THREE (see page 17)

Implementation of thorough human rights due diligence by companies developing advanced 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 the following 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 carries inherent human rights risks, and taking action to address this.

Practice 2 Embedding human rights risk assessment into the working methods and cultures typical of the product-oriented technology organizations developing foundation models.

Practice 3 Evaluating "technical" mitigations with a focus on people in situations of vulnerability or marginalization.

Practice 4 Creatively building and using leverage to address residual risks and enable remedy for harms.

Practice 5 Engagement with affected stakeholders and human rights experts across all phases of human rights due diligence.

For each Headline this paper recommends specific near-term actions that States, companies and other stakeholders should pursue. These key messages and recommendations are summarised in the Appendix.

When developing these propositions for ways forward, attention has been given to spotlighting existing practice and initiatives, the perspectives from diverse stakeholders about how generative AI technologies are currently built and function, and the rapidly evolving landscape of users, use cases, risks and actual harms connected to these technologies.

These recommendations from the first phase of 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.

[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".

Background

The sudden, widespread availability of generative AI tools has reshaped the digital landscape and animated a global debate about the opportunities and risks this technological development creates. However, early discussion and initiatives on generative AI regulation and technical standards have not sufficiently taken into account international standards and practices from the field of corporate responsibility and accountability. To address this gap, in May 2023 UN B-Tech launched its Generative AI Project to identify how the UN Guiding Principles on Business and Human Rights can guide more effective understanding, mitigations and governance of the risks associated with the development and deployment of generative AI.

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.

Attention to the near-term opportunities and risks of current generative AI models are taking place within a wider debate about the future promise and threats to humanity of Artificial General Intelligence[5] (AGI). Some see the potential of AGI to augment and propel the human experience to new, currently unimaginable heights. Others have voiced concern that the achievement of AGI will usher in a dystopian future in which AGI works to undermine human existence. Regardless of one's position on these matters, addressing current risks and harms should be a priority. Focusing on what is in front of us also has the merit of being a way to iterate guardrails that can protect against present but also future, as yet unknown, harms.

[5] "Artificial general intelligence (AGI) is the representation of generalized human cognitive abilities in software so that, faced with an unfamiliar task, the AGI system could find a solution. The intention of an AGI system is to perform any task that a human being is capable of. Definitions of AGI vary because experts from different fields define human intelligence from different perspectives. Computer scientists often define human intelligence in terms of being able to achieve goals. Psychologists, on the other hand, often define general intelligence in terms of adaptability or survival." Source: Tech Target

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. This has added even more urgency to an already high number of regulatory and other AI initiatives[6], in some cases resulting in amendments to regulatory proposals[7] and voluntary initiatives.

While many of these initiatives seek to advance the governance, assessment and management of risks to society by private sector actors developing and deploying generative AI technologies, few have incorporated 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* (OECD Guidelines). This misses the opportunity to benefit from several decades of policy developments, regulatory convergence, business practice, multi-stakeholder collaboration and civil society advocacy about how to (and how not to) advance responsible corporate conduct across complex global value chains, including in the technology sector itself.

Against this backdrop, B-Tech launched its Generative AI Project 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[8].

Building on OHCHR's existing work on tech and human rights and on B-Tech's other workstreams, the B-Tech Generative AI Project is being implemented through an iterative process of research and engagement. This has included exploratory interviews with company practitioners, civil society, technical experts, and other key stakeholders, as well as workshops and other convenings involving multiple stakeholders.

This paper lays out the findings from the first phase of the Project, implemented between June and November 2023.

---

[6] OECD AI Policy Observatory lists over 1000 AI policy initiatives from 69 countries, territories and the EU.

[7] See, for example Generative AI: A closer look at the EU AI Act.

[8] 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).

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.

International human rights comprise a list of basic rights that are universally recognised as necessary for a person to live a life of equality and dignity. They have developed— and will develop—based on debate, cooperation and consensus building between countries from different regions and people from many different groups, cultures and ethical perspectives. Human rights are not, and may never be, uniformly protected, upheld and respected around the world, but a rights-based approach to advancing the responsible development and deployment of generative AI brings the global legitimacy and pragmatism that no other set of standards or ethical frameworks can claim. Human Rights also offer an intentionally aspirational roadmap and moral compass grounded in our shared humanity to help guide decision-making.

A rights-based approach to advancing the responsible development and deployment of generative AI provides[9]:

- **Agreed norms for assessing and addressing impacts:** 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. This offers a common basis for evaluating the nature of risk, i.e., whose lives may be adversely impacted by the proliferation of generative AI, and in which specific circumstances and in what specific ways. A human rights lens can also prompt attention to categories of impacts on people that may be otherwise missed such as impacts on political participation, access to public services, freedom of assembly, the right to a fair trial, the right to physical and mental health and freedom to form and hold opinions, for example.
- **An architecture for convening, deliberation, and enforcement:** The international human rights framework has developed a relatively elaborate architecture of regional and international institutions and processes (e.g., courts, specialized agencies, intergovernmental bodies, designated experts) which can be used both to facilitate consideration of these issues and, in some instances, to monitor and even enforce implementation of any resulting outputs. But the international human rights system is not the only human rights regime. Regional human rights systems, national courts, local civil society actors and human rights defenders — all actors paying increasing attention to human rights in the digital economy — also play an important, if uneven, role.
- **Standards that already apply to both States and corporations:** Focusing on upholding international human rights has the merit of reinforcing that State action related to generative AI should not run counter to States' human rights commitments. This includes the requirement that States should refrain from interfering with or curtailing the enjoyment of human rights when deploying generative AI to deliver State functions (such as education, healthcare, social security or defence) or when regulating generative AI development and use by others. At the same time, the international human rights framework was never *exclusively* framed around States. The Universal Declaration of Human Rights, which is the foundation of the contemporary international human rights, directs "every individual and *every organ of society*" to strive and to secure

[9] See also Jason Pielemeier, Global Network Initiative: The Advantages of Applying the International Human Rights Framework to Artificial Intelligence which informed these key messages.

universal and effective recognition *and observance*" of human rights (emphasis added). And the UNGPs outline and clarify the respective *duties of States to protect*[10] *individuals from business-related* human rights harms, and the related, distinct and independent *responsibilities of corporations to respect*[11] human rights in the course of their activities.

The idea that impacts on human rights should sit at the core of governing technologies, including artificial intelligence, is gaining traction. Government policy and regulatory initiatives focused on the societal risks of artificial intelligence already focus, to varying degrees, on human rights impacts[12]. Some the most well-known technology companies have in place public commitments and processes focused on operating with respect for human rights and many have invested considerable attention to "fairness and bias" of artificial intelligence and machine learning models, demonstrating one way in which principles of equality and non-discrimination already have some purchase in the field. At the international level, the UN Secretary-General has called for guardrails to ensure AI governance is grounded in human rights, transparency, and accountability.

Nonetheless, a large number of international and national policy initiatives aiming to address the risks of generative AI entirely omit reference to international human rights standards. At best, this omission creates an unnecessary normative vacuum in which a global patchwork of laws, industry standards and business practices based on similar sounding aims — e.g., ethics, fairness, openness, transparency — are vaguely or differently defined without being tethered to the important question of real impacts on real people in real places. At worst, vague definitions of responsible conduct and judgments of what constitutes acceptable risk when developing and deploying generative AI systems may be determined solely by short-term geopolitical interests and market incentives that externalize impacts on people.

To catalyse greater application of a human rights lens to developing and deploying generative AI, B-Tech has developed a Taxonomy of Human Rights Risks Connected to Generative AI. The Taxonomy outlines numerous "risk examples" connected to generative AI across nine categories of internationally agreed human rights. While the Taxonomy does not attempt to comprehensively list all potential human rights harms, it does offer an examination of some of the main ways in which human rights are currently at risk from generative AI development and deployment, as well as risks that are likely to materialize in the medium-term future. While many of these human rights impacts may have been associated with earlier forms of AI, they have been or risk being exacerbated by the particularities of generative AI.

---

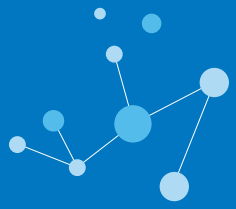
[10] According to the UNGPs, States should take "*appropriate steps to prevent, investigate, punish and redress human rights abuse through effective policies, legislation, regulations and adjudication*" (UNGP1), and that States "*should consider a smart mix of measures—national and international, mandatory and voluntary—to foster business respect for human rights.*" (UNGP3). For more information about the UNGPs calls for States to apply a "smart mix" of measures and ensure "policy coherence", see: B-tech foundational paper Bridging Governance Gaps in the Age of Technology — Key Characteristics of the State Duty to Protect

[11] The Corporate Responsibility to Respect Human rights requires all business enterprises to: 1) **avoid causing or contributing** to adverse human rights impacts through their own activities, and address such impacts when they occur; and 2) **seek to prevent or mitigate adverse human rights impacts** that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts. See also B-Tech foundational papers Key Characteristics of Business Respect for Human Rights and Designing and implementing effective company-based grievance mechanisms

[12] For example, the **Draft EU AI ACT**: Art. 35 "seeks to ensure a high level of protection for fundamental rights and aims to address various sources of risks through a clearly defined risk-based approach"; the **Draft Brazilian AI Bill**: Art. *7* " grants persons affected by AI systems the following rights vis-à-vis "providers" and "users" of AI systems, regardless of the risk-classification of the AI system" and lists Right to information about their interactions with an AI system; Right to an explanation, Right to challenge decisions or predictions, Right to human intervention, Right to non-discrimination and the correction of discriminatory bias, and Right to privacy and the protection of personal data; the **U.S Blueprint for an AI Bill of Rights** is set out "to help guide the design, use, and deployment of automated systems to protect the rights of the American public in the age of artificial intelligence. (…), these principles are a blueprint for building and deploying automated systems that are aligned with democratic values and protect civil rights, civil liberties, and privacy."

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.



There are multiple forces shaping the evolution of generative AI technologies and the ways in which they are used. These include the ambition by some technologists to pursue AGI, the economic potential of these technologies, diverse industries investing in generative AI-enabled efficiency gains and innovations, research organisations exploring new solutions to climate and other shared challenges, and diverse interests — some benign, some malicious — of States and individuals.

What is undeniable is that private enterprises sit at the core of technological breakthroughs and the modes through which generative AI will permeate our lives. Many of these companies also have highly specialized understandings of generative AI's functioning and the considerable financial, human and other resources needed to remain at the forefront of innovations. Leaders from across academia, government, civil society and business have called for States to regulate the practices of companies developing and/or deploying generative AI foundation models, applications and products. In sum, the challenge of addressing the adverse impacts of generative AI is in large part a challenge of establishing a robust, principled and pragmatic governance framework of corporate responsibility and accountability for those impacts.

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. For example, addressing States deploying generative AI technologies in ways that violate the human rights of their own citizens, political parties flooding social media with AI-generated disinformation about opposition candidates, or criminal actions by individuals (such as use of synthetic voice to commit fraud) will not be eradicated through a sole focus on corporate conduct. That said, advancing responsible business conduct, in addition to being valuable in its own right, can serve as one powerful avenue to minimize the likelihood of the most egregious harms resulting from generative AI's proliferation.

The UNGPs are a powerful tool for the task at hand. They articulate the components of a multi-layered governance model needed to advance business respect for human rights in practice: one that moves beyond the false binary choice between voluntary self-regulation and binding law requirements. As John Ruggie, architect of the UNGPs once noted, "(The UNGPs) *are not merely a text. They were intended to help generate a new regulatory dynamic, one in which public and private governance systems, corporate as well as civil, each come to add distinct value, compensate for one another's weaknesses, and play mutually reinforcing roles—out of which a more comprehensive and effective global regime might evolve.*"

With this vision in mind, a UNGPs-informed approach to this task would emphasize that:

I. 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.

II. Regional, national, international and industry-led initiatives focused on advancing responsible generative AI should align to the international standards of business conduct: the UNGPs and OECD Guidelines. This means, in particular, integrating a true risk-based approach to identifying and taking action on impacts.

III. Greater urgency to ensure effective judicial and non-judicial access to remedy for individuals whose human rights are harmed by the development or deployment of generative AI.

The following pages briefly summarize the rationale for these points of emphasis and provide practical recommendations for what States should do in the near-term to advance a "*comprehensive and effective global regime*" for governing generative AI, including through robust engagement with companies and civil society.

I. 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.

RATIONALE: The UNGPs focus on this "smart-mix"[13] because decades of experience have shown that regulation alone is rarely a 'silver bullet' solution that on its own will ensure that respect for human rights is consistently placed at the heart of private sector governance, strategy and conduct. There are varying reasons for this, most notably that some companies will lag behind others in terms of responsible conduct and compliance meaning States invariably need to explore how to make use of diverse legal regimes and policy domains. It is also true that regulators often struggle to keep pace with technological innovation meaning that more nimble methods of governance, alongside regulation, are demanded.

The UNGPs also emphasize the importance of States ensuring coherent action across all State agencies that shape business practice. Where policy coherence is lacking, States will fail to provide private companies developing and deploying generative AI with clear and predictable expectations. This serves to undermine both the effectiveness of State measures and the ability of companies to adjust their practices in a stringent and comprehensive manner.

Special attention to the role of home states within which generative AI investment is most prolific is critical. The action or inaction of these States — notably the United States and China globally, as well as leaders in their region such as Singapore, South Korea, India, Germany, the United Kingdom, Brazil, Chile, Egypt and South Africa[14] — will have an outsized impact on whether generative AI governance and business practice is rights-respecting. The UNGPs reinforce that "States should set out clearly the expectation that all business enterprises domiciled in their territory and/or jurisdiction respect human rights throughout their operations" through applying "domestic measures with extra-territorial implications" or approaches that "amount to direct extraterritorial legislation and enforcement" (GP 2).

[13] The UNGPs state that States should take "*appropriate steps to prevent, investigate, punish and redress human rights abuse through effective policies, legislation, regulations and adjudication*" (GP1), and that States "*should consider a smart mix of measures—national and international, mandatory and voluntary — to foster business respect for human rights.*" (GP3). See also The Geneva Academy of International Humanitarian Law and Human Rights, The relevance of the Smart Mix of Measures for Artificial Intelligence

[14] See the Global AI Index which ranks 62 countries based 111 indicators, collected from 28 different public and private data sources. The indicators are split across seven sub-pillars: Talent, Infrastructure, Operating Environment, Research, Development, Government Strategy and Commercial.

States participating in multilateral fora and multi-stakeholder processes is also an essential component in ensuring the international legitimacy, coherence and effectiveness of State action[15]. Coherent collective action is key to ensure that State can address the fact that the development and deployment of generative AI can occur at a high speed and at great scale across borders. Cooperation between States can take the form of aiding States with less financial resource or technological expertise to implement their own form of the smart-mix needed to govern human rights risks and challenges particular to their national context. Whatever the modalities of collective action, the "Guiding Principles provide a common reference point" and can "serve as a useful basis for building a cumulative positive effect that takes into account the respective roles and responsibilities of all relevant stakeholders." (GP10).

Finally, the UNGPS reinforce that meaningful involvement of civil society and affected groups, as well as investors, academics and business leaders can reinforce accountability for States to prioritize human rights protections and investment into effective ways to address business-related human rights impacts associated with generative AI technologies.

RECOMMENDATIONS: The following are near-term priorities for applying the "smart mix" of measures and ensuring policy coherence within State actions aimed at governing generative AI.

- **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.** In most jurisdictions there are a number of existing laws that can address particular aspects of human rights risks connected to generative AI. Beyond the clear relevance of privacy law, data protection and data security, other relevant domains of law include labor, non-discrimination, copyright, product liability, consumer law and sector-specific regulations (e.g., for finance and healthcare).
- **States should provide effective guidance and associated capacity building to business enterprises on how to respect human rights when developing or deploying generative AI.** Even with State policy and regulatory measures in place, companies — especially start-ups and smaller firms — can benefit from clear direction as to what respect for human rights means in operational terms given their specific place in the generative AI value chain. This could be done through creating new guidance or adapting existing AI due diligence standards[16] to ensure alignment with the UNGPs and OECD Guidelines. This would address current differences in terminology and risk scope which may cause barriers to use for companies who have to comply with guidance in multiple jurisdictions. While guidance is not prescriptive in nature, it may accompany regulation, fill near-term gaps in understanding what good corporate conduct looks like, and even be a testing ground for future regulatory proposals.
- **Authoritative corporate transparency regimes from the corporate responsibility and accountability field should be used to complement technology specific transparency requirements.** For example, EU legislative

[15] The UNGPs note that "*Collective action through multilateral institutions can help States level the playing field with regard to business respect for human rights, but it should do so by raising the performance of laggards. Cooperation between States, multilateral institutions and other stakeholders can also play an important role*" (Commentary to UNGP10).

[16] For example, the U.S National Institute of Standards and Technology Artificial Intelligence Risk Management Framework, the IEEE Standard Model Process for Addressing Ethical Concerns during System Design (IEEE 7000); ISO's Information technology — Artificial intelligence Risk management (ISO/IEC 23894). See also AI Standards Hub.

initiatives — whether in the drafting, transposition or enforcement phase — should make use of the European Sustainability Reporting Standards when seeking to increase transparency by companies developing and deploying generative AI. One obvious benefit is that this can minimize the reporting burden on companies. Equally important, the content of these standards addresses aspects of corporate conduct that are critical to understanding the seriousness and likely quality of generative AI related risk management / due diligence[17]. In addition, innovative multi-stakeholder structures that enable meaningful transparency without increasing risks to human rights or undermining legitimate business interests should be supported by States. The Global Network Initiative's methodology of conducting independent company assessment is a leading model that could be applied to evaluating generative AI related company processes and conduct.

- **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.** The goal of such efforts should be to enable the navigation of the societal and technical complexities of how digital technologies function, the associated risks to people, and the role (including limits) of companies to address these risks. Lawmakers should also include provisions for ensuring appropriate investigative and other capacity exists to meaningfully enforce regulations[18]. Innovative ideas will be needed to achieve this. For example, the Australian Human Rights Commission championed the idea of creating an Australian eSafety Commissioner in order to build capacity in government and industry in relation to promoting and protecting rights when designing, procuring and deploying AI systems. The now-established eCommissioner has published a position statement and guidance on generative AI.
- **States should pursue multi-lateral action focused on the protection and respect of human rights.** Large levels of cooperation and the rapid spread of best practices between States will be crucial for advancing the responsible generative AI. Such multilateral efforts can also minimize the risks of States pursuing their own, however legitimate — economic and geo-political 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.** This will be necessary to better understand how proposals would work in a wide range of real-world scenarios. It is also an important way to reduce the risk of States introducing new legislative or policy measures that are inconsistent with their own international human rights obligations. Brazil's

[17] The European Sustainability Reporting Standards (ESRS) are set to apply to more than 50,000 companies in the EU and at least 10,000 outside. The standards help meet the purpose of the EU's Corporate Sustainability Reporting Directive (CSRD) in ensuring greater rigor and comparability in companies' reporting on their sustainability performance across environmental, social and governance issues. Many aspects of the ESRS can be applied to, reporting by companies developing and deploying Gen AI. For example: The governance of social matters, taking a risk-based approach to the value chain, and clarity about the scope of reporting (including impacts on consumers/end-users). See Shift's Putting the European Sustainability Reporting Standards into Practice.

[18] For example, in its An EU AI Act that works for people and society: Five areas of focus for the trilogues, the Ada Lovelace institute emphasized the importance of well-resourced centralised regulatory capacity to ensure an effective AI governance framework, noting that "The EU central functions should therefore be funded as much as, if not more than, other domains where safety and public trust are paramount and where underlying technologies form important parts of national infrastructure — such as civil nuclear, civil aviation, medicines, road and rail" and that " As in already in place in some of these sectors (for example, 80% of the European Medicines Agency's funding comes from the market entities it regulates) a mandatory fee could be levied on developers over a certain threshold (for example, expenditure on training runs, or compute).

appointment of a Commission of Experts comprised of 18 members with recognized expertise in technology law and regulation, and proposals for stakeholder representation as part of implementation of the EU AI Act are good examples of this.

II. Regional, national, international and industry-led initiatives 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.

RATIONALE: States and stakeholders do not need to reinvent standards of responsible business conduct for companies developing and deploying generative AI technologies. Rather, established expectations of what constitutes responsible business conduct, laid out by the UNGPs and OECD Guidelines, should be the starting point. The UNGPs provide the authoritative definition of responsible corporate conduct in relation to business impacts on human rights: *the Corporate Responsibility to Respect Human Rights*. To meet this responsibility, all companies should have in place "policies and processes appropriate to their size and circumstances" including a "human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights."(GP 15).

Two features of the UNGPs and OECD Guidelines are particularly pertinent to public policy, regulatory and voluntary efforts to guide and govern the conduct of companies developing and deploying generative AI technologies.

First, a risk-based approach grounded in severity of risks to people: The UNGPs expect companies to maintain a wide view of risks, meaning that businesses should identify the actual and potential impacts to all human rights related to the company's business activities and relationships (GP 12). This, by definition, means that companies must anticipate and address risks to people, regardless of whether these lead to reputational, operational or financial risks to business. It is important to note that risks to people and risks to business tend to converge, whether in the short, medium or long-term. But where this is not the case, this does not release companies of their responsibility to address human rights harms.

The UNGPs recognize that companies will often need to prioritize impacts or risks for initial attention, due to the typically large volumes of actual and potential impacts on human rights connected to their operations, products or services. To address this, the UNGPs state that prioritisation should take place on the basis of their likelihood and relative severity from the perspective of those who are or may be affected. Severity involves considering the scale of an impact (how grave it is), its scope (how widespread it is) and its irremediability (how hard it would be to make right)[19]. An impact can be severe overall even if it would only be so in one of these dimensions. The UNGPs are also clear that in the case of human rights impacts, severity should always be the dominant factor over likelihood, particularly where delayed action would make an impact irremediable.

[19] For a more detailed explanation of the severity based risk framework, including considerations for how it applies in the technology sector, please see: Identifying Human Rights Risks Related to End-Use; a B-Tech foundational paper.

To ensure that the most significant harms to people flowing from the development and deployment of generative AI are adequately identified and addressed, this risk-based approach must be better integrated into regulation, technical standards and guidance/methodologies for risk assessment. In practice, this means:

- Establishing that generative AI risk assessments should not conflate evaluations of risks to people and risks to business, but instead treat them as distinct types of risks while acknowledging the relationship between the two. Traditional risk assessment methodologies typically fail to make such distinctions, leaving it unclear on what is driving risk evaluations and decision-making[20]. Efforts to establish greater clarity could usefully build on the reporting concepts of impact materiality, financial materiality, and double materiality[21].
- Requiring that prioritization based on *severity of risks to people* is part of generative AI risk assessment. This is different to high-level classifications of unacceptable, high and low or minimal risk AI systems laid out in the EU AI Act[22]. Rather, this would set expectations that when conducting risk assessment of any generative AI system, developers and deployers apply the scale, scope and irremediability framework when undertaking prioritization of attention needed in the face of an inevitably large set of potential and actual human rights risks[23].

Second, accounting for the nature of a company's involvement with human rights risks and impacts in establishing thresholds of appropriate action. The generative AI value chain is vast and includes a growing number of technology companies involved in the development of generative AI (including foundation model developers and Model hub and MLOps platforms); companies supplying capabilities (such as hardware or computing resources, and investors[24]); and companies, States and individuals using generative AI across diverse industries and contexts.

Under the UNGPs, all companies across the value chain have a clearly defined responsibility to prevent and address negative impacts connected with operations, products or services, wherever they occur in the value

[20] By way of illustration, the NIST AI Risk Management Playbook (Govern 1.3) explains that "AI risk tolerances range from negligible to critical — from, respectively, almost no risk to risks that can result in irredeemable human, reputational, financial, or environmental losses. Risk tolerance rating policies consider different sources of risk. This places risks to people within the types of risks that should be considered, yet in the state of common practice, risks to people are not taken into account sufficiently in risk scoring.

[21] As noted by Shift in Double Materiality: What you need to know "The concept of materiality has been applied for many years in voluntary sustainability reporting standards, albeit with different meanings. For example, for the Global Reporting Initiative, the focus has been on the significance of impacts on people and planet; for the Sustainability Accounting Standards Board, it has been on implications for the financial success of the company. "**Double materiality**," introduced for the first time in the European Sustainability Reporting Standards (ESRS), brings these two approaches together: ESRS 1 makes clear that …a sustainability matter is material "when it meets the criteria for impact materiality or financial materiality or both. Impact materiality and financial materiality are considered "interrelated" by the ESRS: an impact on people or the environment can be financially material from the start, or become financially material over time. The evolving nature of this relationship is sometimes referred to as "**dynamic materiality**."

[22] For an overview of the EU AI Act risk classification see, The EU AI Act: A Primer, by the Center for Security and Emerging Technology.

[23] An existing example of this is the Evaluation of Harms approach within Microsoft's Harms Modelling — "a practice designed to help you anticipate the potential for harm, identify gaps in product that could put people at risk, and ultimately create approaches that proactively address harm.". While it is unclear if the same as the UNGPs factors of severity, this Microsoft framework uses a prioritization approach based on scale and severity, as well as likelihood and frequency.

[24] Under the UNGPs investors at every stage of a company's lifecycle — from start-up to maturity — also have a responsibility to operate with respect for human rights. This is critical because investors, both asset owners and managers, have unique and systematic influence over how companies in the technology industry are governed, make decisions, and act. This B-Tech Investor Briefing provides institutional investors with holdings in digital technology companies with high-level analysis and guidance on how to apply the UNGPs framework to these investments.

chain. This affirms that deployers of AI systems and other actors — including across diverse industry sectors[25] and States — should be within the scope of regulation and other policy interventions. The UNGPs "involvement framework"[26] provides a principled and pragmatic approach to this value chain wide approach, and sets the basis for determining appropriate *action* to address risks and impacts identified as part of risk assessments. The UNGPs make the distinctions that:

- Where a company **causes or may cause** an adverse human rights impact, it should cease or prevent the actual impact. In situations of actual impacts having already occurred, the company should also provide remedy to affected individuals.
- Where a company **contributes to, or may contribute to**, an adverse impact it should cease or prevent its contribution and use leverage to mitigate any remaining impacts to the greatest extent possible. A company can contribute to human rights impacts either in parallel with contributions by third parties, or by enabling or incentivising third parties to cause impacts, whether they are states or other business enterprises. In situations where a company has contributed to actual impacts, the company should also provide remedy to affected individuals.
- Where a **company's operations, products or services are linked through business relationships** to an adverse impact. It does not have a responsibility to provide remedy since it has not contributed to the harm, but may consider contributing to remedy or using leverage to incentvise those causing the harm to do so.

This "involvement framework" is central to the effectiveness of the due diligence approach in the UNGPs for several reasons that are also pertinent to the generative AI value chain:

- It reflects both the extent and the limit of a company's responsibility for human rights harms in their value chains to situations where the harm is connected in some way to its operations, products or services. This also means that companies are not expected to address every negative impact on human rights that is occurring in their value chain or industry. Rather, they should do so when they are causing, contributing or are linked to the adverse impact.
- It clearly differentiates the type of response expected from companies based on the nature of their involvement with an impact, such that the expectations are reasonable and proportionate.
- It emphasizes that effective due diligence must include attention from all companies across the generative AI value chain to how their own practices, business strategies, model and product design decisions, and go-to-market choices might be contributing factor to human rights harms. But also, even where this is not the case, companies should use leverage to address harms that are nonetheless connected to their products via business relationships.
- Causation, and contribution as a form of causation, reflect well-established concepts in civil liability under existing national laws, and doing due diligence can help protect companies against claims made on that basis.

[25] The following Business for Social Responsibility (BSR) resources identify human rights issues associated with AI technologies in healthcare, retail and the financial services sectors, and recommendations for addressing these impacts.

[26] For a detailed explanation of this framework, please see: Taking Action to Address Human Rights Risks Related to End-Use, a B-Tech foundational paper.

RECOMMENDATIONS: The following are near-term priorities for aligning public policy and regulatory initiatives with the core concepts of the international standards of business conduct.

- **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.** Two positive examples of this are the G7 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, and the OECD Declaration on Promoting and Enabling Responsible Business Conduct in the Global Economy, signed by 51 States.
- **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.** Applying the UNGPs severity approach in how developers and deployers of generative AI conduct impact assessments will aid in ensuring robust, well-considered risk evaluations, while also having the important behavioural effect of focusing decision-makers on the most severe risks to people. Using the Involvement Framework provides clarity as to the nature of the responsibilities that different actors in the generative AI value chain for specific impacts, and what role they are therefore expected take in addressing the harm.
- **Establish multi-stakeholder dialogue to deepen appreciation of what a full value chain approach to addressing human rights risks means in practice.** Multi-stakeholder, expert dialogue about the different ways that distinct companies across the generative AI value chain can become involved in human rights harms or risks, could aid in informing what principled and pragmatic/reasonable conduct by each actor should look like in practice. These could be standalone "responsibility sandboxes" or integrated into regulatory sandboxes[27]. To deliver value, these would likely have to focus on singular use case, specific human rights harms or very well-defined tranches[28] of the generative AI value chain. These dialogues could result in recommendations for the best mix of measures aimed at preventing harms, but should also extend to identifying which actors under which circumstances should be prepared to take a role in remedying harms. This is especially important given the particular complexities concerning attribution and explainability of AI-connected impacts on people, the large volume of potentially harmed individuals, and the reality that many harms may occur due to actions of consumers post-deployment of generative AI systems.

III. Ensuring effective judicial and non-judicial access to remedy for individuals whose human rights are harmed by the development or deployment of generative AI.

RATIONALE: Even where law, company commitments, business processes and market incentives are aligned towards avoiding business-related human rights harms, some harms still occur, often in ways that are devastating to victims, families and, at times, whole communities. Consider, for example, workplace discrimination, contamination of rivers or product safety incidents. The development and deployment of generative AI will not be an exception. In fact, totally avoiding negative harms seems highly unlikely. These harms might arise in various ways, including due to malicious use, negligence from a developer, the unpredictability of generative AI model behaviour, and the unpredictability of use cases before generative AI

[27] For a fuller explanation, see the OECD's Regulatory Sandboxes in Artificial Intelligence

[28] An example of such a "tranche" would be synthetic media around which some work has been done, though not yet with reference to the UNGPs involvement framework, to articulate what practices different actors (builders of technology infrastructure, creators, and distributors and publishers) connected to actual or potential harms should take. See: Partnership on AI's Practices for Synthetic Media: A Framework for Collective Action

models and products are released.

Establishing a robust and comprehensive system of remedies for human rights harms connected to generative AI will require the same focus, determination, investment, ingenuity and resources that drive technological innovation today. The right to an effective remedy for violations of human rights is enshrined in international human rights law. The duty of States to provide access to effective remedies for business-related human rights harms, including human rights harms associated with the development and use of digital technologies, is a key aspect of the State Duty to Protect human rights, as laid out in the UNGPs. Responsible technology companies recognise both the commercial and ethical urgency of this task. Not only does the absence of remedies risk undermining companies' social license to operate (which can have significant commercial and legal consequences, both in the short and long term), but it is also the right thing to do.

Placing human rights at the heart of remediation mechanisms and strategies focused on generative AI is vital to ensuring that this technology is aligned with principles and values that allow all human beings to thrive. The positive news is that attention to redress for victims is already a feature of some AI-focused regulations and technical standards. But much more needs to be clarified and implemented.

The UNGPs provide a pragmatic and compelling framework for delivering effective remedies for human rights harm to affected people and communities[29]. Key elements include:

- A focus on the need for a range of remedy mechansims that can respond to all types of human rights risks' as opposed to a narrow set of issues such as freedom of expression or privacy.
- Explanation of the distinct but complementary roles of different kinds of actors (public and private, including companies) in providing remedy, regardless of whether or not harms are intended.
- Reinforcement of the foundational role that judicial systems play within a wider "remedy ecosystem" that also includes State-based non-judicial and non-State-based non-judicial mechanisms.
- Setting out the different forms that effective remedies can take: Restitution, Compensation, Rehabilitation, Satisfaction, and Guarantees of non-repetition. These concepts come from international human rights law and prioritize the importance of understanding and taking proper account of the needs and perspectives of affected people and groups in deciding what kind of remedy is needed in different situations.
- Offering a practical set of effectiveness criteria to guide the design, evaluation and improvement of non-judicial approaches to remedy, including those managed by the private sector.

RECOMMENDATIONS: The following are near-term priorities for making Access to Remedy a central feature in the governance of generative AI.

- **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.** The voice of affected people is a vital, often overlooked, aspect of remedying harms. Processes available to these groups to seek remedy often do not work for them or the remedy

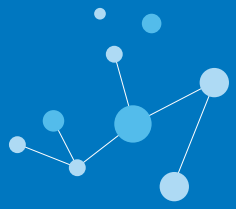
[29] The following B-Tech foundational papers provide more detail about the Access to Remedy pillar of the UNGPs: Access to remedy and the technology sector: basic concepts and principles; Access to remedy and the technology sector: a "remedy ecosystem" approach; Designing and implementing effective company-based grievance mechanisms; and Access to remedy and the technology sector: understanding the perspectives and needs of affected people and groups

delivered is not fully satisfactory when compared to the loss, pain, and suffering experienced. To avoid repeating this pattern with regard to harms that flow from the use of generative AI technologies, States, responsible companies and civil society must find ways to amplify the voice of at-risk individuals in designing an adequate remedy eco-system. This has both process-related elements, such as understanding which types of remedy processes are most accessible to at-risk or harmed people. It also has outcome-related elements, such as clarifying what restitution, compensation, rehabilitation, satisfaction, and guarantees of non-repetition should look like in practice.

- **States should ensure access to judicial remedies where individuals may have been harmed by the development or deployment of generative AI technologies.** This could involve ensuring robust enforcement of legal standards that underpin public law remedies of various kinds when harms related to generative AI occur. Depending on the operation of the regime in question, this could include financial compensation, the criminal sanctions imposed, and binding orders to correct legal breaches and address underlying causes of harm. States should also ensure that people are able to enforce their rights directly when these may have been harmed by the development or deployment of generative AI technologies. This could occur, for instance, under the law of tort, or under a statutory cause of action. States may also then need to make investments and adjustments needed to ensure that people are aware of their rights and that the barriers they face in accessing judicial processes are recognised and addressed.
- **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.** This could include, for instance, raising complaints with regulators about the conduct of technology companies (e.g., with respect to trade practices, anti-competitive behaviour or data-handling); independent complaint and mediation processes led by consumer protection bodies or national human rights institutions; or company-based or collaborative grievance mechanisms. Of particular promise here is the "national contact point" system established under the OECD Guidelines, which is perhaps one of the most widely established but underutilised forms of accountability for AI systems that currently exist. In all cases, special attention should be invested in aligning mechanisms with the UNGPs effective criteria.

HEADLINE THREE

Implementation of thorough human rights due diligence by companies developing foundation models will provide an important basis for risk management across the generative AI value chain.



Attention to the conduct of companies developing foundation models is especially important because the risk evaluations, decisions, business practices and disclosures of these companies can help to mitigate harms at an early stage across the generative AI value chain. This is not to suggest that these companies are the only actors that have responsibilities; as noted above, all business actors across the generative AI value chain must meet their *Corporate Responsibility to Respect Human Rights*.

Identifying good due diligence practices by companies developing foundation models — as well as the challenges and limitations they encounter in the course of this — can deliver considerable benefits. In addition to minimizing the severity or likelihood of harms caused by the use of generative AI systems by malicious individuals, non-State actors or States, good practices implemented by model developers can inform due diligence by application developers and deployers, creating significant efficiencies in risk assessment and mitigation across the generative AI value chain, including establishing sector or use case-specific good practices.

There are also level-playing field and responsible innovation benefits. On the one hand, guidance and tools based on good practices at the foundation model level can aid start-ups seeking to enter the market space, thus enabling responsible competition and innovation. On the other hand, deeper multi-stakeholder consensus about what constitutes good practice can, when accompanied by meaningful regulation and incentives, establish a global, level playing field of conduct under which companies should not be allowed to operate.

The work to clarify good practices has already begun. The field of AI risk assessment and management is rich with academic research, civil society reporting and company policies and processes. As laid out in a supplement to this paper, An Overview of Human Rights and Responsible AI Company Practice some of the most prominent technology companies driving generative AI have long had practices focused on identifying and addressing risks to society from artificial intelligence. Moreover, there is broad alignment between the risk management frameworks and methods being embedded in regulatory proposals or technical standards and the UNGPs. The November 2023 OECD Report *Common guideposts to promote interoperability in AI risk management — comparing AI risk management frameworks* demonstrates that leading risk management frameworks are generally aligned with the top-level steps (define, assess, treat for risks, and govern risk management) of an "Interoperability Framework" based on the OECD Guidelines.

However, there are certain critical elements of Human Rights Due Diligence as laid out by the UNGPs that are absent or under-emphasized in existing guidance and public discussion about company practice. These elements or practices together represent the more transformative propositions of the UNGPs which are designed to guide companies to deliver demonstrably better outcomes for people. They are:

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.

Practice 2 Embedding human rights risk assessment — focused on all human rights with prioritization based on severity — into the working methods and cultures typical of the product-oriented technology organizations developing foundation models.

Practice 3 Evaluating "technical" mitigations with a focus on people in situations of vulnerability or marginalization.

Practice 4 Creatively building and using leverage to address residual risks and enable remedy for harms; and

Practice 5 Engagement with affected stakeholders and human rights experts across the full cycle of human rights due diligence.

Practice 1: Boards and executives identifying the extent to which the company's business model and strategy carries inherent human rights risks, and taking action to address this.

RATIONALE: Under the UNGPs, companies are expected to conduct human rights due diligence across all of their business activities and relationships. As outlined in B-Tech Addressing Business Model Related Human Rights Risks foundational paper[30], this includes addressing situations in which strategic product design/release decisions or business model choices create or increase human rights risks. Business model choices are made and reviewed by the top leadership of an enterprise responsible for strategy. Executives and senior managers then work to ensure that these strategic choices are reflected in the company's operating model and often culture. Where this leads to business processes, incentives, or practices that increase risks to workers, communities or consumers, a tension can arise between a company's business model and its ability to respect human rights.

The intent of identifying features of business models and strategies is not to simplistically label some as rights-respecting and others not. Nor should the existence of risks automatically lead to business model adaptation, as this can sometimes create other, more serious risks. Rather, it spotlights that in certain situations, managing negative impacts connected to specific business activities requires active oversight and involvement from boards and executives making business model and strategy decisions. This is in contrast to tasking teams to manage these risks at an operational level while making higher level decisions that could unwittingly undermine those teams' work.

In the context of developing and deploying generative AI foundation models, typical features of business models and strategies that companies' due diligence programs and practices will need to account for can include:

- *Features of foundation models:* generative AI foundation models are often described as general-purpose technologies meaning that they can be used, misused and abused in endless ways, many of which may

[30] See also B-Tech Institutional Investor Business Model Tool which provides templates for institutional investors to use in engaging technology companies on these business model risks.

not be immediately foreseeable. Harms can also arise when these technologies perform in sub-optimal or unexpected ways[31]. In addition, the inherent complexity and opacity of generative AI models means that undesirable outputs are challenging, though by no means impossible, to explain and so fix.

- *Revenue / monetisation strategy:* The nature of human rights risks will shift based on the actors that a company targets and supports to use its foundation models. These may include developers building applications, enterprise customers in different sectors, public sector organisations or individuals using consumer interfaces. This can be a complex picture but the risks associated with distinct revenue models should be well-understood by board and executives.
- *Nature and speed of deployment:* It is common practice for technology companies to release products and tools incrementally and iteratively in order to gather feedback from users to inform improvements. Companies at the core of generative AI development are taking this same approach, which can focus policy and public attention on how to govern risks of current models and to help prepare to grapple with and address more powerful, future ones. As such, some level of risk to society is inevitable. Moreover, risks may be increased where models may be rushed for release in order to get ahead of or catch up with competitors.
- *The implications of closed, proprietary vs open-source models*: Whether a company pursues an open versus closed source strategy to developing and deploying foundation models can also impact the shape of risks to human rights connected to its products, and how the company can best mitigate risks that flow from end-use[32].

NEXT STEPS: All stakeholders need to be part of creating greater clarity about the appropriate and reasonable oversight role of boards and leadership practices of company founders and executives to address business-model related risks. This should start with deliberations involving executives, civil society, regulators and investors to discuss this issue in more depth, and lead to 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[33].
- 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

[31] See, for example, this study by AlgorithmWatch/AI Forensics. ChatGPT and Co: Are AI-driven search engines a threat to democratic elections?

[32] Open and closed generative AI foundation models each have their own unique set of ethical considerations. For example, Open models can allow for greater transparency and auditability, enabling researchers and developers to scrutinize the model's training data, code, and decision-making processes. This openness can help identify and address potential biases or ethical concerns. However, open models can be misused or repurposed for malicious applications, and can be difficult to control and steer, as changes made by one developer can have unintended consequences for others. Closed Generative AI Foundation Models tend to allow for controlled development and deployment and are generally less susceptible to misuse or repurposing for malicious applications, as they are not publicly accessible. However, closed models can lack transparency and auditability, making it difficult to assess potential biases or ethical concerns. This lack of openness can hinder the identification and mitigation of ethical issues.

[33] This aspect of the governance of sustainability issues is getting increasing attention in corporate reporting requirements. For example, the European Sustainability Reporting Standards ask companies to disclose how they understand and address the relationship between material impacts on people and their business model(s). For more information about this, see Governance, Strategy and Business Models: Four Highlights from the European Sustainability Reporting Standards, by Shift

akin to the "precautionary principle"[34]. This is consistent with the notion that, under the UNGPs, severity of actual and potential impacts should inform company action, even if their likelihood is considered low.

- 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[35].

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.

RATIONALE: The UNGPs place significant emphasis on companies ensuring that human rights due diligence, starting with risk assessment and prioritization, should occur early and on an ongoing basis to enable timely and effective actions to address human rights risks. In particular human rights impact assessments should be undertaken "prior to a new activity or relationship; prior to major decisions or changes in the operation (e.g. market entry, product launch, policy change, or wider changes to the business); in response to or anticipation of changes in the operating environment (e.g. rising social tensions); and periodically throughout the life of an activity or relationship." (GP 17)

However, within the typically product-driven, decentralized technology companies and corporate cultures developing generative AI foundation models there is very little understanding of how to achieve this in practice. This lack of clarity has several negative consequences. First, human rights risks may not be robustly assessed at moments in the life-cycle[36] of generative AI foundation models that would allow these risks to be mitigated most effectively. This could be the result of uncertainty about which moments make sense for human right analysis. Even where clear decision-points for safety and ethical review are established, product teams and engineers may struggle to identify human rights risks and effectively prioritize attention to those risks due to lack of tools, training or access to internal or external human rights expertise.

Second, the ways in which product-oriented tech companies work can themselves be a "black-box" to external stakeholders, especially those without a background in computer science, software development and engineering. This can make it hard for companies to communicate in an impactful way about risk management efforts that are embedded into existing processes. It also reduces the ability for meaningful dialogue about the appropriateness and effectiveness of leveraging product development working methods to identify and assess human rights risks.

[34] The Precautionary Principle was most clearly articulated in Principle 15 of the 1992 Rio Declaration as "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation"

[35] World Economic Forum: Board Duties in Ensuring Company Engagement with Affected Stakeholders provides a brief overview of the role of corporate boards of directors in relation to the concept of "affected stakeholders"

[36] As noted by the OECD in its report, "The lifecycle encompasses the following phases that are not necessarily sequential: planning and design; collecting and processing data; building and using the model; verifying and validating; deployment; and operating and monitoring". This description is also reflected in the AI lifecycle used to structure the U.S. Department of Commerce, NIST Artificial Intelligence Risk Management Framework

Third, methodologies for human rights impact assessments recommended by stakeholders or required in regulations may be too far removed from the pace and iterative nature of developing foundation models. This can push human rights risk assessment and management towards solely being a compliance exercise as opposed to a meaningful tool in influencing corporate conduct towards improved management of risks to people and rights-based, responsible innovation.

Some computer scientists and academics have already begun to frame up important aspects of this topic. The most extensive literature and guidance for companies focuses on the ethical deliberations within agile software development processes[37]. Much rarer is attention to how to integrate human rights considerations into these processes. Where this focus does exist, some researchers are raising important questions about the limits of relying on the latest manifestations of agile product methods to address issues beyond the design of specific features, such as establishing the overall objectives of technology systems, or software products[38].

NEXT STEPS: All stakeholders need to be part of creating greater clarity about what it means in practice to meaningfully embed assessment of human rights risks into the development of generative AI foundation models. This can start with deliberations involving technologists (inside and outside of companies), civil society, academia and business and human rights experts 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. This would ideally cover the spectrum from human rights impact assessments of a full model/system to lighter and quicker assessments as features of foundation models are iterated.
- 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. Instrumental to this will be case studies or hypothetical scenarios detailing the mechanics (who is involved, how long does the assessment take, are external stakeholders consulted, how does it connect/support to other deliberations (e.g., on usability or ethics), what are the pitfalls to avoid etc.).
- Establishing mechanisms to allow external stakeholders to understand, appreciate and inform the quality of human rights risk identification and prioritization practices. This is distinct from important formal reporting by companies developing foundation models about risks and risk mitigation strategies. It is far more about creating safe and honest spaces for innovating robust human rights risk assessments.

[37] See Zuber, N., Gogoll, J., Kacianka, S. et al. Empowered and embedded: ethics and agile processes Humanit Soc Sci Commun 9, 191 (2022).

[38] In his article Can We Move Fast Without Breaking Things? Software Engineering Methods Matter to Human Rights Outcomes, Alex Voss is clear about the merits and potential limits of relying fully on the agile methods. He argues that: "Our choice of methods also affects human rights outcomes for stakeholders. I will argue below that software engineering has regressed as methods have become overly focused on the continuous delivery of new functionality at the expense of overarching and cross-cutting concerns. From this critique, I develop the notion of *rights-respecting software engineering* and outline what it would take to develop methods that make an explicit representation and consideration of rights possible in the development of software products and services.".

Practice 3: Evaluating "technical" mitigations with a focus on people in situations of vulnerability or marginalization.

RATIONALE: The companies developing and deploying the most prominent generative AI foundation models have done a great deal to find ways to make those models accurate and safe. This paper uses the term "technical mitigations" to denote the varied approaches used by these companies to train, tune and guide the behavior of models. Examples of these mitigations include: at the input level, prompt filtering and engineering; at the systems level, supervised fine-tuning, Reinforcement Learning with Human Feedback (RLHF) and Red Teaming; and at the output level, blocklists and classifiers. These same companies also implement post-deployment monitoring to track the model's performance over time, detect potential biases, monitor model usage and ensure model security[39].

It is beyond the scope of this paper to assess the technical effectiveness of these mitigation and monitoring approaches. Company practice however does appear, upon initial review, to be well-aligned with the main process expectations of the UNGPs focused on tracking performance. The UNGPs state that "in order to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response" (GP 20). Companies publish "system cards" and peer-reviewed research to communicate key information about their models, including the technical mitigations that have been applied, with what results and limitations[40]. Some also include reference to how, at a high level, internal and external stakeholders have been involved in evaluating mitigations, which is also consistent with UNGPs' expectations that companies should "draw on feedback from both internal and external sources, including affected stakeholders" (GP 20 b).

However, two important expectations of the UNGPs require greater attention. First, the differential effectiveness of mitigations for people in situations of vulnerability needs to be further addressed. The UNGPs state that "Business enterprises should make particular efforts to track the effectiveness of their responses to impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization" (GP 20). This is echoed in the "Measure" phase of the NIST Artificial Intelligence Risk Management Framework which

[39] The following explanations of some of these mitigations have been sourced from various sites and through the use of Bard and ChatGPT: Mitigations explained: **Prompt filtering:** Some inputs that do not violate law or responsible data policies may be part of producing problematic engagements or outputs. In these cases, it may be appropriate to filter, block, and hard code responses for some inputs until the model can respond in the intended way; **Prompt Engineering:** Direct modifications of the user inputs are used to guide a model behaviour and encouraging responsible outputs. This can be done by including contextual information or constraints in the prompts to establish background knowledge and guidelines while generating the output. Modifications can be implemented in various ways such as with automated identification and categorization, assistance of the LLM itself, or rules engines; **Supervised fine-tuning:** Adapting a pre-trained Language Model to a specific downstream task using labelled data. Reinforcement Learning with Human Feedback: Reinforcement learning from human feedback is a machine learning approach that combines reinforcement learning techniques, such as rewards and comparisons, with human guidance to train an artificial intelligence agent; **Red Teaming:** Red teaming involves many kinds of probing, testing, and attacking of AI systems. It is a best practice in the responsible development of systems and features using LLMs. It helps to uncover and identify harms and, in turn, enable measurement strategies to validate the effectiveness of mitigations; **Blocklists:** A way to prevent the generation of high-risk content is to compile a list of all the phrases that your model should not, under any circumstances, be permitted to include in a response. Many words are easily identifiable as problematic; slurs, for example, are typically offensive no matter their context **Classifiers:** The more effective, but also more difficult, approach is to develop classifiers that detect and filter outputs based on the meaning conveyed by the words chosen. Classifiers, when properly trained on known examples of a particular sentiment or type of semantic content, can become highly effective at identifying novel instances in which that sentiment or meaning is expressed.

[40] See, for example: Llama 2: Open Foundation and Fine-Tuned Chat Models from Meta, GPT-4 Systems Card from OpenAI, and PaLM 2 Technical Report, from Google.

includes provisions that the "Appropriateness of AI metrics and effectiveness of existing controls are regularly assessed and updated, including reports of errors and potential impacts on affected communities".

In the context of generative AI foundation models, companies do point to limitations in existing technical mitigations, but apparently without tracking these implications for groups in situations of high risk. Examples of gaps already document include: safety features built in English that may not have the same effectiveness in other languages, the possibility that humans involved in the human feedback aspects of reinforcement learning may be biased or intentionally promote toxicity, and continuing reports that models are still prone to "hallucinations".

To be consistent with the UNGPs, companies should be tracking quantitative and qualitative data about the extent to which these limits to mitigations are increasing risks or adverse impacts to vulnerable groups. For example: are English trained models likely to result in disproportionate harm in countries where digital literacy is low, or to children whose native languages are not English? To what degree are any human biases being replicated within RLHF disproportionately impacting women or ethnic minorities?

Second, the adverse human rights risks of mitigations themselves need to be evaluated. Under the UNGPs, a company's mitigations are "activities" which need to be assessed for their own human rights impacts. This is already a well-understood principle for many companies in the telecommunications and technology sector. By way of illustration, efforts by online platforms to prevent sexually exploitative content can also result in the removal of content involving nudity for important cultural, educational, and health related reasons. For generative AI, two such examples of mitigations that civil society have identified as carrying human rights risks include watermarking of synthetic media content, and the abuse of human rights of workers within the human feedback phase of RLHF.

Focusing on the extent to which mitigations work for the most at-risk communities should not detract from the critical importance of companies implementing mitigations that have proven to reduce human rights risk for high volumes (at times likely millions) of people. The focus should not be on pitting the human rights of the many against the few. Rather, the idea is that more quantitative and qualitative data about differential effectiveness and the human rights risks of mitigation enables diverse stakeholders to navigate dilemmas and find solutions. Lack of such data will likely simply embed the idea that harms to large swathes of humanity are no more than residual, or worst acceptable, risks and costs of generative AI development and deployment.

NEXT STEPS: All stakeholders need to be part of creating greater clarity about how to evaluate the effectiveness of technical mitigations for the most at-risk groups. This can start with deliberations involving technologists (inside and outside of companies), civil society, affected stakeholders and academia 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. Where this is the case, establishing what constitutes good practice, including about how to communicate externally about insights, should be a priority.
- 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. Where applied, these qualitative research methods should draw on good social science practices.
- 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. This might mean establishing voluntary but legally enforceable 'safe harbour' provisions for all parties.

Practice 4: Creatively building and using leverage to address "residual risks" and enable remedy for harms.

RATIONALE: It is unlikely that generative AI technologies will be deployed with zero residual risks to people and planet. But companies developing foundation models should, consistent with the UNGPs, continue to take action to reduce the severity and likelihood of those risks beyond technical mitigations described above. The UNGPs establish that where companies are not causing adverse impacts, but nevertheless connected to those impacts, they are expected to build and use leverage to effect change to mitigate risk or remediate negative human rights impacts. This is shown in the "involvement framework" described above. As noted previously by B-Tech, leverage can take many forms that commonly fall into the following categories:

- **Bilaterally** with third parties in the context of commercial relationships. For example, via enforcing contractual terms and incentives, or undertaking capacity building.
- **With other companies** — whether industry peers or companies from other industries. For example, through the development of technical standards and associated efforts to incentivise implementation.
- **Via partnerships** with an institution or actor that can play an effective supporting role in influencing the actions of the third party in question. This might include with a home or host State, an international organization or a civil society organization.
- **Through multi-stakeholder collaboration** — whether in the context of formal initiatives or not, and possibly with the aim of establishing relevant public policy and law.

The concept of leverage understood in all its forms can galvanise problem-solving and innovation by companies to tackle the root causes of social externalities in their industry or operating contexts. In addition, it can serve to minimize the outsourcing of responsibility to entities that lack the will, competence or resources to implement it. The notion that companies developing generative AI foundation models should build and use their leverage to address harms that they have contributed to or may be linked to should be reinforced in regulation, technical standards and guidance. This should not in any way reduce efforts to implement technical mitigations. Rather, the creative use of leverage by companies must complement these mitigations.

NEXT STEPS: Companies developing foundation models and their stakeholders should develop greater clarity about the "state of the art" and "art of the possible" in building and using leverage situations of generative AI deployment where, even after technical mitigations, human rights risks exist. The following dimensions should be explored.

- **Responsible use policies, terms of use in contracts, guidance and enforcement.** There is a rich body of practice in this domain which includes companies such as Meta and Google providing guidance and open-source tools for developers to evaluate safety and other features of the applications and products they develop (for example, Meta's Llama 2 Responsible Use Guide, Google's Responsible AI Practices, Microsoft's

Responsible Innovation Best Practices Toolkit and responsible AI training modules). When advancing good practices in the context of generative AI, initial points of emphasis could include: 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:** Know Your Customer (KYC) refers to the set of guidelines, regulations and practices in the financial services sector to verify the identity, suitability and risks involved in maintaining a business relationship with a private sector or government customer. Microsoft, in their report Governing AI: A Blueprint for the Future has already signaled the value and relevance of this idea to address the risks of AI, including generative AI. This is particularly interesting because there has also been considerable attention to the use of leverage by financial institutions to address human rights risks connected to their lending and investments: See, for example, Using Leverage to Drive Better Outcomes for People. When advancing good KYC practices in the context of generative AI, initial points of emphasis could include: 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:** It is broadly accepted that the human rights risks connected to the development and deployment of generative AI will require action from a wide range of actors. And many technology companies already engage in collective action to address upstream labor rights risks and downstream privacy, freedom of expression and cyber-security risks. When advancing collective action good practices in the context of generative AI, initial points of emphasis should be: Clarity about the scope of standards and activities being focused on to avoid signaling that some abuses or root causes of abuse are being addressed when they are not; 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 that will lead to demonstrably better human rights outcomes.
- **Leverage for remedy:** Under the UNGPs, companies have a responsibility to provide for or cooperate in remedy processes where they have caused or contributed to the harm; they may also take a role in enabling remedy where they are linked to the harm, which can be an effective means of reducing risks of their continuation or recurrence[41]. There may also be practical, reputational and social license reasons for companies to contribute to a well-functioning eco-system of avenues for remedies. Building on efforts to advance responsible investment by financial institutions, the use of leverage for remedy by companies developing foundation models could be explored at two levels:
    - **Focusing on customers' "preparedness for remedy" as part of KYC.** In practice, this could involve asking questions to assess the effectiveness of grievance mechanisms that customers put in place and providing proactive support to strengthen customers' or industry-level mechanisms.
    - **"Enabling remedy" in specific cases.** In practice, this could involve executives using their influence to bring greater focus to conversations about remedy when severe impacts occur; supporting

[41] see further UNGP 22.

fact-finding in support of affected stakeholders and customers when impacts are alleged to have occurred, but the facts are disputed; and even contributing financial or other resources to bolster remedy packages.

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.

RATIONALE: The UNGPs strongly emphasize that companies should engage with affected stakeholders or credible proxies and expert stakeholders[42] as part of assessing, mitigating and remediating adverse impacts on human rights that they are, or may become, connected to. This is because these stakeholders typically have a strong understanding - many of them through lived experience - of the interplay between business operations, value chains, products and services and human rights impacts.

The UNGPs state that "To enable business enterprises to assess their human rights impacts accurately, they should seek to understand the concerns of potentially affected stakeholders by consulting them directly in a manner that takes into account language and other potential barriers to effective engagement. In situations where such consultation is not possible, business enterprises should consider reasonable alternatives such as consulting credible, independent expert resources, including human rights defenders and others from civil society" (GP 18)

Meaningful engagement by foundation model developers with the perspectives of affected and expert stakeholders about the human rights risks and impacts connected to the use of these models, or generative AI technologies in general, can improve the quality and credibility of a company's risks assessments. This is especially true given the challenges of fully predicting the ways in which these models will behave or be used and misused post-deployment. Affected stakeholders, as distinct from deployers, may prove to be the most reliable source of information about persistent or emerging harms. Moreover, robust engagement that authentically identifies and addresses human rights-related concerns can help to establish or sustain the social license of generative AI technologies.

The UNGPs affirm that engagement with stakeholders should not stop at risk assessment, but instead take place across all phases of human rights due diligence and as part of remedying harms. In this way, engagement with affected and expert civil society stakeholders can inform model design, risk mitigation and deployment decisions towards mitigating risks to human rights, as well as strategies to ensure victims of harm have access to remedy when harms occur. This ethos, which has been expressed by some as design from the margins in the context of social media is arguably the most promising pathway to the proliferation of generative AI grounded in dignity and equality for all.

Several organisations have already elaborated guidance for stakeholder engagement around AI systems that

[42] Under the UNGPs: **Affected stakeholders** are any individual or group whose human rights has been affected by an enterprise's operations, products or services; **Credible proxies** are individuals or groups who are recognised as legitimate representatives of affected stakeholders. **Expert stakeholders:** individuals or groups with expert knowledge about the impacts of business on people's human rights. In the context of AI systems, it is important to note that affected individuals/communities can be anyone directly or indirectly affected by AI systems or decisions based on the output of AI systems, though they do not necessarily interact with the deployed system or application.

offer a good starting point for advancing good practice by foundation model developers. For example:

- The NIST AI Risk Management Playbook (Govern 5.1) lays out suggested actions to ensure robust engagement with relevant AI actors, which includes affected stakeholders. The playbook emphasizes that participatory stakeholder engagement: assist in identify emergent scenarios and risks in certain AI applications; is best carried out from the very beginning of AI system commissioning through the end of the lifecycle; and is best carried out by personnel with expertise in participatory practices, qualitative methods, and translation of contextual feedback for technical audiences.
- European Centre for Not-For-Profit Law's Framework for Meaningful Engagement as part of assessing human rights impact for AI systems focus on the importance of *Shared Purpose* beyond the interest of the convening organisation; *Trustworthy Processes* that are inclusive, open, fair and respectful and delivered with integrity and competence; and *Visible Impact* i.e., that involvement can make a significant contribution to decision-making, or makes changes to the governance of the organisation, product or service to align it with the public interest.
- Data and Society's Democratizing AI: Principles for Meaningful Public Participation provides recommendations concerning, among other things: Early-stage public participation to ensure that decision-makers do not become wedded to a preconceived decision before receiving public input; the need for equity and social justice commitments to guide every aspect of participation; the design of participation methods for high-quality engagement; and the need to build the technical capacity of communities while also acknowledging that "Affected people do not need to know how to build an algorithm to have an opinion on how automated decision-making systems should (or should not) affect their lives."

In the context of generative AI foundation models, their general-purpose nature presents challenges related to the massive numbers, diversity and geographic location of potentially affected stakeholders that may need to be engaged. On these points, B-Tech Improving Stakeholder Engagement in Tech Company Due Diligence affirms that "technology companies should not interpret the expectation of the UNGPs as meaning that they must engage with every one of the many thousands, even multiple millions, of stakeholders potentially impacted by the use of the company's products and services. Rather, tech companies should seek to hear from a representative mix of stakeholders, with resources prioritized to where risks to human rights are most severe".

B-Tech has also previously called attention to the importance of engagement with expert and affected stakeholders outside of North America and Western Europe. For some companies, membership in robust multistakeholder initiatives, such as GNI has proved a useful way to connect with local civil society and affected groups, as well as the critical work of NGOs with specific focus on digital rights in their region, and international NGOs working to support with capacity building and guidance to these organizations[43].

Broad-based public participation represents another possible avenue for engagement with affected stakeholders. For example, Anthropic and Open AI have partnered with The Collective Intelligence Project to pilot "Alignment Assemblies" aimed at shaping the trajectory of generative AI deployment in society. The project founders also aspire to experiment with other modes of engagement, citing federated citizens' assemblies, retroactive funding

[43] Such as the Paradigm Initiative working across Sub-Saharan; ELSAM in Indonesia, and Association por los Derechos Civiles in Argentina (see, ADC's due diligence guidance focused o marginalized groups), and GNI and Global Digital Partners' Engaging Tech Companies on Human Rights: A How-To Guide for Civil Society

processes for writing better mode evaluations and public red-teaming. Another exemplar innovation from the wider AI domain is the Ada Lovelace Institute's Citizens' Biometrics Council which brings together 50 members of the UK public to deliberate on the use of biometrics technologies like facial recognition.

Finally, all individuals have political and public participation rights that "play a crucial role in the promotion of democratic governance, the rule of law, social inclusion and economic development, as well as in the advancement of all human rights". This is a reminder that formalized, legally mandated mechanisms for ensuring that the voice and interests of specific affected stakeholders are represented in the development of generative AI foundation models should also have a role to play. This could, for example, take the form of regulatory oversight of public participation forums, or establishing modes for specific groups such as data enrichment workers or artists to exercise their collective bargaining rights.

NEXT STEPS: Establishing when, how and under what conditions companies developing generative AI foundation models can most meaningfully ensure engagement with at-risk or impacted stakeholders and civil society organizations requires attention from business, civil society and regulators. Initial points of focus could be 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.

Appendix

Table of 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