Generative AI Human Rights Due Diligence Project
UN Human Rights

Concept Note May 2023

Statement of Need

The newly widespread availability of Generative artificial intelligence ("AI") technology has quickly reshaped the digital landscape. Generative AI, including chatbots relying on large language models such as ChatGPT, can generate novel media content, including text, music, images and videos, in response to text prompts from a user. This technology opens new possibilities for education and knowledge production, personal development, and expression. But optimism about these potential benefits has also, appropriately, been accompanied by concerns about adverse impacts of Generative AI, provoking debate among policymakers and the public at large. Some companies launching products powered by Generative AI, such as Microsoft, Open AI, and Google maintain ethical and/or human rights guidelines for AI more broadly and have confirmed that they are conducting risk assessments and testing processes. Beyond the product level, there also exists a broader debate about the appropriate timing of introducing Generative AI models into the consumer market at scale, and the extent to which the publication of models and codes still under development is appropriate.

Information about the extent to which human rights due diligence (HRDD) has been conducted as part of the development and deployment of these technologies is limited, and there has been little opportunity for learning across the industry about the most effective approaches to prevent and mitigate human rights risks stemming from advances in Generative AI technologies. While ethical frameworks can be useful in guiding company risk management approaches, they are unlikely to cover the full spectrum of internationally recognized human rights standards necessary to ensure human dignity and non-discrimination. As additional companies enter the market, risks will increase as many firms prioritize speed and profits over preventing and mitigating adverse human rights impacts. Some standard setting is under way in different fora, but regulatory frameworks are not keeping pace with technical developments or with effective human rights risk mitigation practices, leaving potentially severe human rights and other societal risks unaddressed.

There is therefore an urgent need to explore what constitutes the appropriate scope and practice of business responsibility in relation to Generative AI. Identifying appropriate responses to this question and building alignment across industry, civil society and standard setters about expectations should draw on international human rights standards. In particular, the expectations set out in the UN Guiding Principles on Business and Human Rights (UNGPs) can provide authoritative and widely accepted guidance. Using these global standards as the initial basis for unpacking the scope and nature of corporate responsibilities can also provide a common foundation for constructive and robust dialogue. Considering this, the UN Human Rights Office is launching this project as a contribution to wider global and societal debates about how to realise the positive potential of Generative AI, while mitigating the many severe, salient risks that may accompany this transformative technology.

Objectives

- Clarify the expectations under the UNGPs for companies developing and launching Generative AI products in order to achieve common and more effective human rights risk management approaches across the industry.

- Raise awareness and facilitate exchange among key stakeholders and interdisciplinary experts to shape a comprehensive understanding about the role the UNGPs can play in governing Generative AI responsibly.
- Inform the debate about policy options for managing human rights risks related to the development and launch of Generative AI, including through mandatory and voluntary measures.

Approach

Building on OHCHR's existing work on tech and human rights, the project will be implemented through an iterative process of research and engagement. This will include conducting a series of exploratory interviews with company practitioners, civil society, technical experts, and other key stakeholders, as well as workshops and other convenings involving multiple stakeholders.

Project implementation will be led by the UN Human Rights B-Tech Project, supported by Shift, a leading center of expertise on the UNGPs, with additional assistance from the Global Network Initiative (GNI), a leading multistakeholder initiative with fifteen years of experience working on tech and human rights. The project team will include business and human rights expertise and be supported by a technical expert group of academics to bring in state-of-the-art computer science research on Generative AI to enable a better understanding of its implications for managing human rights risks. For the company engagements, the project will draw in particular from the team's ongoing engagement with companies in the B-Tech Community of Practice (CoP), while also involving additional companies.

Given the highly competitive nature of efforts in the Generative AI area, the ability to work within an existing framework with some degree of trust among the participants and organizers will be essential for the initial exercise of mapping risks. We aim to create a space that allows for companies to jointly discuss not only best practices but also lessons learned. For civil society engagement, the project will draw on the expertise and perspectives from GNI's academic, civil society, and investor constituencies, and the extensive network of the B-Tech community of digital rights advocates, academics, and other practitioners. The project will be underpinned by research into AI's salient human rights risks and build on work already done on AI risk management, as well as the work of B-Tech and other OHCHR tech-related work, the Global Network Initiative (GNI) and other relevant initiatives on business and human rights in the tech sector.

Activities

The project will be structured into three initial phases:

1) Consultations and Mapping: Map current company approaches to human rights due diligence in order to identify and mitigate salient human rights risks stemming from or being linked to the development, deployment, and use of Generative AI. This will be carried out through desk-based research, explorative expert interviews with company representatives, civil society, and experts, and the establishment of a reference group comprised of experts from AI and data science as well as leaders from relevant policy areas.

2) Peer Learning Workshops and Engagement: Targeted convenings (in-person and virtual) with key stakeholders, in particular with leading tech firms, relating to Generative AI practices, including measures adopted to account for potential adverse impacts on human rights.

3) Consolidation: Based on the input gathered and received, a summary of salient risks and state of play on company practices and learning to date will be prepared, without attribution. Further outputs will be developed based on the needs and approaches identified in the convenings described above. An outline of good practice/standards based on UNGPs expectations is a likely outcome.

Building on the consolidation of the findings from the initial phases of the project, possible additional activities will be determined in consultation with those participating.

Guiding questions to inform project activities and engagements:

- How do existing HRDD practices need to be adapted to be able to incorporate the risks and mitigation strategies stemming from or being linked to the use of Generative AI?
- In which corporate governance structures are risk management processes for Generative AI embedded in companies, e.g., would engaging with product counsel be impactful and if so, how?
- At which points should an evaluation of misuse/abuse be integrated into the AI research/innovation process? How can this be tailored to find the appropriate balance between innovation, actual functional breakthroughs/insights and severity of risks.
- What does meaningful transparency about, and accountability for, use, risks and adverse impacts of Generative AI look like in practice?
- Are existing ethical checks/questions in the R&D process sufficient? What, if anything, does attention to, and prioritization of, a vast number of prospective adverse human rights impacts add to this in practice?
- What does responsible conduct look like for different objectives and modes of AI being made or becoming available for use?
- Which forms of mitigation can be employed?

Outputs

- B-Tech Briefing (max. 15 pages) on key risks, current state of practice, best practices and standards, and UNGPs expectations with regard to Generative AI.
- Additional shorter supporting supplements will be produced that discuss Generative AI-related human rights risks, scenario-based UNGPs interpretations for AI use cases, the state of practice of business respect for human Rights in AI companies and policy coherence in AI Governance aligned with the UNGPs.