UNITED NATIONS
HUMAN RIGHTS
OFFICE OF THE HIGH COMMISSIONER

Responsible AI and Human Rights: An Overview of Company Practices

Supplement to B-Tech's Foundational Paper on the Responsible Development and Deployment of Generative AI.

Introduction

The rapid growth of generative AI and expansion of public access to generative AI tools over the past year have led to questions about the extent to which technology companies are effectively assessing and addressing the risks to people and society associated with their generative AI products and services.

This paper, a supplement to the UN B-Tech Project's foundational paper on generative AI, seeks to contribute to discussions about that question by providing an overview of company practice regarding responsible AI and human rights. The paper demonstrates that:

1. **Companies are largely guided by AI principles or the company's mission. S**ome of these principles reference human rights and this can more readily enable integration of a human rights-based approach to responsible AI.

2. **Management of risks to people and society associated with generative AI tends to lie with product teams and specialized responsible AI teams.** Many companies have invested in human rights expertise to complement responsible AI teams in order to embed human rights into risk assessment and mitigation processes.

3. **Companies take a wide range of approaches to assessing and addressing risks associated with generative AI.** Where it occurs, integration of a human rights-based approach aids in building comprehensive understanding of risk.

4. **Company disclosures focused on responsible AI are largely technical in nature.** However, there have already been increased disclosures about risks to people and society associated with generative AI systems.

5. **Many companies seek to provide access to redress mechanisms in relation to the use of their products and services.** However, remedy for the harms to people and society associated with generative AI require a broader remedy ecosystem.

The growth and maturation of responsible AI programs have laid important foundations for addressing the risks to people and society associated with generative AI. While some companies do incorporate a human rights lens, more consistent integration of a human rights-based approach grounded in the UN Guiding Principles on Business and Human Rights (UNGPs) into their responsible AI policies and processes is needed.

This analysis is informed by a combination of engagement with companies of various sizes on human rights and responsible AI, including generative AI, as well as a review of public company materials. It is important to note that the focus of this paper is on the practices of the technology companies that are the predominant developers and deployers of generative AI today. It does not examine the role of companies in other industries that are increasingly deploying generative AI in a wide variety of domains.

1. Companies are largely guided by AI principles or the company's mission. Some of these principles reference human rights which can more readily enable integration of a human rights-based approach to responsible AI.

The majority of large technology companies developing generative AI have published AI principles in the past few years that are designed to guide how they develop and deploy AI products and services, including generative AI. These principles are often oriented around a set of values that have grounded the responsible AI field, notably: the promotion of human values and human control over technology, fairness and nondiscrimination, transparency, explainability, accountability, safety and security, privacy, and human rights.

Despite early skepticism from external stakeholders about the sincerity of such high-level principles, many companies are successfully utilizing them to ground and guide AI product development and the remit of responsible AI teams whose role is to help the company implement its principles in practice. When AI principles reference human rights explicitly, they can serve as a forcing function to integrate human rights into product reviews–including of generative AI. Company AI principles identified by B-Tech that explicitly reference human rights include:

- Google's AI principles, which include a commitment to not design or deploy "technologies whose purpose contravenes widely accepted principles of international law and human rights."
- Salesforce's AI principles, which pledge to "safeguard human rights and protect the data we are entrusted with."
- NEC's AI principles, which state that their purpose is to "prevent and address human rights issues arising from AI utilization" and to "guide our employees to recognize respect for human rights as the highest priority in each and every stage of our business operations."

Other company AI principles do not reference human rights broadly, but rather call out specific rights. For example, both Microsoft and Meta's AI principles commit to protecting the privacy and security of people's data and ensuring products are fair and work equally well for all people, all of which are effectively human rights commitments.

Most large companies leading on generative AI also have corporate human rights policies that cover activities across their entire value chain, and so in principle also cover the development and deployment of generative AI. Some corporate human rights policies reference AI specifically. For example:

- Meta's human rights policy states, "Human rights also guide our work developing responsible innovation practices, including when building, testing, and deploying products and services enabled by Artificial Intelligence (AI)."
- Microsoft's human rights policy references the company's AI principles, stating, "We seek to mitigate and prevent risks by applying rights-aware decision making throughout our products' lifecycles and business relationships. For example, we are committed to a responsible approach to artificial intelligence (AI) by applying our AI principles to its development and use."

Emerging generative AI companies are quite young and generally do not have AI principles or human rights policies. Instead, they tend to anchor their responsible AI approach on the company's mission or broader ethics commitments. For example:

- OpenAI's mission is to ensure artificial general intelligence (AGI) benefits all of humanity, and its charter includes a commitment to "avoid enabling uses of AI that harm humanity" and to "to doing the research required to make AGI safe."
- Anthropic's purpose is to "build systems that people can rely on and generate research about the opportunities and risks of AI." It has also integrated human rights into the principles that form the "constitution" that guides its AI assistant.
- HuggingFace's ethical charter lists generative AI use cases they want to prevent, which include violations of human rights, as well as ethical principles that guide their work such as transparency and fairness.

2. Management of risks to people and society associated with generative AI tends to lie with product teams and specialized responsible AI teams. Effective collaboration with human rights teams and ensuring responsible AI teams have human rights expertise are key to creating human rights aligned risk assessment and mitigation processes.

Corporate AI principles or broader company commitments form the policy backbone of generative AI companies' commitments. Thus, management of the risks to people and society associated with generative AI tends to lie primarily with the product teams developing generative AI tools and the responsible AI teams that have been created to help companies implement their related commitments (see for example, the work of this Google team and this Microsoft team).

These teams are often named "responsible AI," "responsible innovation", "AI ethics," or "AI safety" teams. They typically do not have human rights as an explicit part of their strategy or governing framework, unless their AI principles or commitments reference human rights. In large companies with established human rights teams, these teams may occasionally provide input into the human rights-related aspects of the responsible AI team's policies and processes, and may conduct or commission HRDD in cooperation. However, this is all largely done via internal collaboration and relationship building rather than formalized governance structures.

Smaller generative AI companies often lack any human rights expertise, or have hired individual "leads" who have been charged with bringing a human rights perspective to broader safety or ethics efforts. Effective collaboration between human rights and responsible AI teams, and ensuring responsible AI teams have human rights expertise on staff, are important for developing human rights-aligned approaches to risk assessment and mitigation.

While both human rights and responsible AI teams typically have a mandate to develop risk assessment and mitigation processes, they generally lack the authority to impose any specific strategy or governing structures. They must therefore collaborate and establish buy-in across research, product, and sales teams in order to embed responsible AI practices throughout a company's generative AI approach. For example, Microsoft has done this by establishing responsible AI leads for each business unit and building out a network of responsible AI "champions" throughout the company to help carry out the day-to-day integration of policies and practices into the development and deployment of AI products and services.

3. Companies take a wide range of approaches to assessing and addressing risks associated with generative AI. Integration of a human rights-based approach can ensure methodological consistency and a comprehensive understanding of risk.

Human rights as a foundation: The interplay of ethics, trust and safety, and human rights-based approaches.

Although some companies assess the human rights impacts of AI products and services in line with the UNGPs, these tend to be on a case-by-case basis. Ethics-based approaches to embedded risk assessment and mitigation processes remain dominant, and because the purpose of generative AI is to generate content, trust and safety approaches are becoming increasingly prevalent as well.

Ethics have dominated the responsible AI field since its inception in academia, and this orientation has been consistently reinforced through the academia-to-technology company pipeline. It has largely been centered around a set of high-level principles–namely fairness, accountability, transparency, and explainability–that are generally compatible with human rights in theory, but in practice have a wide variety of understandings and approaches. There is a general lack of knowledge in the responsible AI field about the relevance of the international human rights framework, the UNGPs, and how they can be utilized.

Trust and safety professionals who previously worked at online platforms are increasingly being hired by generative AI companies to help operationalize ethics and safety efforts.

They are bringing approaches and lessons learned from content governance, which tend to anchor on pre-established taxonomies of harm and the adversarial nature of risks. Ethics, trust and safety, and human rights-based approaches are not mutually exclusive and need not be in conflict. As mentioned previously, many of the ethical AI principles companies have committed to are essentially human rights principles. And most trust and safety issues are also human rights issues as well.

Ethics - based Approaches

- A framework for decision making in situations where right and wrong, good and bad, are not clearly defined.

- Address issues of fairness and social justice where different schools of thought and ethical standards exist; when various choices can be made, and different paths can be chosen.

- Different traditions, cultures, countries and religions may choose different outcomes and priorities suited to specific needs and sensitivities.

Human - rights - based Approaches

- A focus on experiences of the most vulnerable and a holistic recognition of what all members of the society need in order to live with dignity and thrive.

- Based on internationally recognized laws and standards; a common standard of achievement for all people.

- Established rights that should always be protected and respected.

- A minimum threshold and baseline expectation for the responsible use of disruptive technology.

- An internationally endorsed framework for defining company responsibility that considers the critical role of governments.

Source: https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology.pdf

However, because the responsibility to respect human rights is universal for all companies, and the international human rights framework provides the most universally accepted standard, it is important that a human rights-based approach in line with the UNGPs be the foundation for assessing and addressing the risks to people and society associated with AI. Ethics and trust and safety-based approaches that explore a wider variety of issues and decisions outside of the human rights framework can then be integrated from there–for example, when a product may have other challenging social implications, or when there are product abuse behaviors that do not neatly align with human rights framing. Further discussion of the interplay between ethics and human rights-based approaches to responsible AI can be found in this paper.

COMPANY RISK / IMPACT ASSESSMENT APPROACHES

Companies pursue a wide variety of risk / impact assessment models–both technical and non-technical/issue based, and at varying levels of depth. As companies with responsible AI principles steadily work to operationalize them, risk assessment is increasingly being integrated into existing AI product development processes across the product life cycle. However, standalone assessments done at a particular moment in time are also still prevalent, and may be conducted internally or by an external vendor. Some companies have required internal "responsible AI" reviews of all products (e.g. Microsoft's Responsible AI Standard), while for others these reviews only take place voluntarily or upon escalation internally.

Companies tend to develop their own models for risk / impact assessments that are informed by public standards and best practices to various degrees. Because the rapid evolution of generative AI outpaced the responsible AI field's ability to develop and communicate best practices, some of the assessment models companies have pursued for generative AI have been ad hoc and experimental. Although most do not explicitly take a human rights-based approach, these assessments do serve to identify and address some human rights risks. Below are a few examples:

- **Fairness testing** is an increasingly established best practice in the responsible AI field. Fairness testing involves examining the training dataset and probing the AI model to see whether it produces unfair outputs that exacerbate existing societal biases. It is typically both a qualitative and quantitative exercise, and involves technical interaction with both the dataset and the model (for example, see Google's developer guide to fairness testing). For a generative AI product, for example, fairness testing might examine whether a prompt to an image generation tool for images of doctors returns images that are predominantly of men. There are then a wide variety of technical changes to both the model and the dataset that may be made to mitigate these issues. Fairness testing is not typically done with a human rights lens. For example, developers may only consider risks to vulnerable groups in the geographic context they are most familiar with. However, fairness testing effectively seeks to identify and address risks to the rights to equality and non-discrimination.

- **Red teaming** refers to a range of risk assessment methods for AI systems. It usually involves a group of experts from a variety of backgrounds who adversarially test an AI system by identifying flaws and vulnerabilities—for example, ways in which it could produce undesirable outputs, how safety measures can be bypassed, vectors for cybersecurity risks, etc. Red teaming can be both a technical (e.g. technical jailbreaking attempts) and non-technical (e.g. adversarial prompting) exercise, and has gained prominence as a particularly helpful approach to identifying and addressing risks associated with generative AI systems. Both OpenAI and Meta have written publicly about their red teaming approaches and results for their respective generative AI models. One gap demonstrated in red teaming efforts for generative AI thus far is that they have primarily involved technical experts from Western contexts, and have only in some cases included participants with human rights expertise or explicit consideration of human rights impacts (for example, see Microsoft's "harms modeling" approach).
- **Human rights assessments** seek to identify the actual and potential adverse human rights impacts of a given AI product or area technology and make recommendations for addressing those impacts using the methodology and principles outlined in UNGPs. Large companies have been conducting both standalone and integrated human rights assessments of AI products and services for the last several years. These assessments are largely qualitative in nature, although relevant quantitative data and insights from other assessment and testing processes are often considered as an input. A few companies have conducted human rights assessments of generative AI products in the past year, although outputs have not been made public.
- **Algorithmic audits and impact assessments** have emerged as a creation of the responsible AI field over the past several years. There is no standard definition or methodology for either, and despite audits and assessments being fundamentally different processes[1], the terms are often used interchangeably. However, algorithmic audits typically involve quantitative statistical analysis of specific issues. For example, see Twitter's audit of its image cropping algorithm. Both Meta and Open AI conducted and documented algorithmic audit activities of their generative AI models, although they were not referred to as such. At least publicly, algorithmic audits have not taken a human rights-based approach although they may ultimately identify and address some human rights risks.

Algorithmic impact assessments (AIAs) are designed to assess possible social impacts of AI systems, and a variety of largely qualitative tools and methodologies have been proposed by civil society and government entities. Although companies tend not to use the term, many of them conduct AIAs in practice. This is typically done by assessing products or services against their AI principles, or against a predefined taxonomy of risk / harm / impacts. New taxonomies have been proposed to account for the particular risks of generative AI systems. For example, Open AI identified "harms of representation, allocation, and quality of service" in its assessment of GPT-4.

[1] Assessments identify and prioritize risks and make recommendations for addressing them, and are generally forward-looking. Audits determine compliance against a specific standard, involve root cause analyses, and are historical / backward-looking.

Of all the common assessment approaches, AIAs bear the closest resemblance to human rights assessments. However, they lack the methodological robustness and alignment with the UNGPs, such as comprehensive identification of impacts (looking at all internationally recognized human rights vs. a set of predefined risks) and prioritization of impacts based on severity. They can also be highly prescriptive and utilize a checklist type approach. Because of these issues AIAs can miss a wide range of risks to people and society, and are more prone to the inherent biases and viewpoints of the assessors–e.g., geographic and cultural context. They also do not always involve input or consultation from affected stakeholders, which is vital for understanding risks to vulnerable groups and how the context in which generative AI systems are deployed can affect impacts.

In addition to assessments, there are several responsible AI development best practices that implicitly identify and address certain human rights risks. Companies have utilized many of these approaches for generative AI, for example:

- **Data quality reviews** involve the examination of the raw data used to train AI models to look for issues such as incorrect labels, representativeness, accuracy, and bias, that may lead to inaccurate or problematic outputs. In generative AI systems, poor quality data hinders the model's ability to generate accurate and meaningful outputs. These kinds of assessments are therefore key for identifying and addressing risks to human rights that can come from inaccurate or problematic generative AI outputs.

- **Privacy best practices and** privacy preserving approaches to collecting data and training and operating AI models are designed to address the myriad privacy risks associated with training AI models on personal and often sensitive data and using AI models in sensitive contexts. This is also an evolving area of research and practice across the responsible AI field. Established approaches include data cleaning (e.g., removing personal information and metadata that could be used to re-identify people), on-device processing (aka "edge AI"), and federated learning, which can also be applied to generative AI products and services.

- **Human-centered design** is a longstanding approach to developing technology, including AI systems, that is focused on meeting people's needs and is aligned with societal values. It often involves user research with particular demographic groups that can end up surfacing human rights-related issues. Part of human centered design involves not causing harm to users, and so these processes also typically involve designing technology in a way that respects key human rights such as via "privacy-by-design" or "safety-by design" approaches.

- **Model training / fine tuning approaches** for generative AI models are designed to address many of the common issues related to bias and inaccuracy in datasets that can be reproduced by the models and ultimately lead to adverse human rights impacts. Two dominant examples in the generative AI space are Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF), although they also have notable limitations.

Importantly, assessing and addressing the human rights impacts associated with the procurement of data enrichment services, which is typically required to train AI models–including generative AI–is notably absent from existing company approaches to responsible AI. This may be because data enrichment tends to fall in a significant human rights risk management gap inside of companies. It is technically a supply chain issue and primarily involves labor rights issues, yet it is not covered by supply chain sustainability / human rights teams because it does not involve the kind of physical products (e.g. hardware) that supply chain teams typically cover.

4. Company disclosures are largely technical in nature, however, there has been increased disclosures about risks to people and society associated with generative AI systems.

Companies have increasingly disclosed information about their broader approaches to responsible AI, describing the principles, management, and processes that apply across their AI products and services. However, to-date there has been relatively limited company disclosure about their understanding of risks associated with their AI products and services and what they are doing to address them. Disclosures have largely been limited to research advancements inside of companies rather than products, such as the development and release of a new model. They have also been largely geared toward a technical audience in the form of research papers and "model/system cards" or "datasheets for datasets" that can only be meaningfully understood by AI researchers and developers.

Many of these types of disclosures contain little if any discussion of risks and mitigations. However, to date there has been some improvement with generative AI–for example, both Meta and Open AI released technical papers and guidance that clearly articulated risks to people in society. Non-technical disclosures about risks–for example the results of human rights or algorithmic impact assessments of any AI products–are exceedingly rare. (See for example Microsoft and Google). Disclosure about risks and mitigations is not only an important aspect of upholding the responsible AI principles of transparency, explainability, and accountability to which many companies have committed, but also in adhering to the UNGPs.

5. Many companies seek to provide access to redress mechanisms in relation to the use of their products and services. However, remedy for the harms to people and society associated with generative AI require a broader remedy ecosystem.

There are significant challenges related to remedy for adverse human rights impacts associated with generative AI. The diffuse nature of generative AI systems can make it challenging to determine whether a company is causing or contributing to adverse human rights impacts and should therefore provide or cooperate in remedy. It can also be challenging to determine whether responsibility for remedy best lies with the developer vs. the deployer vs. the user of a generative AI tool. Currently, the companies that are both the developers and deployers of consumer facing generative AI tools–e.g., ChatGPT, Google BARD–provide remedy within their own products and services by offering a reporting channel for users to report problematic outputs or other issues they encounter. These channels serve as a sort of guarantee of non-repetition[2], as they enable the company to take effort to ensure the issue doesn't happen again. Moving forward, it will be important for companies to ensure these reporting channels fully meet the UNGPs effectiveness criteria for non-judicial grievance mechanisms[3].

An additional challenge to remedy is that many of the impacts of generative AI will be cumulative and cross-society, and therefore will necessitate cooperation across a variety of both state and non-state actors in a broader "remedy ecosystem." This may take the shape of multi-stakeholder collaboration to develop specific standards for generative AI uses in different sectors, or cooperation in directing individuals toward remedy mechanisms. For example, several companies have signed onto the Partnership on AI's Responsible Practices for Synthetic Media.

Conclusion

The growth and maturation of responsible AI programs have laid important foundations for addressing the risks to people and society associated with generative AI. Generative AI is just the latest development in AI, and over the past several years many technology companies have been building out their responsible AI policies and processes more generally. Although the specifics of risk assessment, mitigation, and management may differ for generative AI due to its unique characteristics, these broader responsible AI policies and processes apply to generative AI as well.

These efforts have contributed to significant advances in the responsible AI field more broadly, and while they generally do not use the terminology of human rights, in practice they help address many of the human rights risks associated with generative AI. While some companies do incorporate a human rights lens, more consistent integration of a human rights-based approach grounded in the UNGPs into their responsible AI policies and processes is needed.

[2] There are five pathways to remedy under the UNGPs—apology, restitution, compensation, rehabilitation, and non-repetition. See https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf and further description here https://www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-and-guidelines-right-remedy-and-reparation.

[3] Under the UNGPs, in order to be effective non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

The responsible AI field has wrestled with many dilemmas that the UNGPs can help address, including how to identify harm/impacts, how to understand the severity of harm, how to prioritize risks, and how to deal with trade-offs/competing values. There are therefore significant opportunities for responsible AI teams to leverage the UNGPs as a foundation for risk assessment and mitigation, rather than reinventing the wheel with new methods.

Acknowledgements

The UN B-Tech team expresses thanks to all the experts and stakeholders that provided input into this Generative AI Project supplement. The team is especially appreciative to BSR, the lead author of this paper.