A Human Rights Assessment of the Generative AI Value Chain

February 2025

BSR®

Contents

ACCOMPANYING RESOURCES

A HRA of the GenerativeAI Value Chain

Overview of the Practitioner Guides

Guide 1: Human Rights Fundamentals

Guide 2: Governance and Management

Guide 3: Impact Assessment

Guide 4: Risk Mitigation

Guide 5: Stakeholder Engagement

Guide 6: Policies and Enforcement

Guide 7: Transparency and Disclosure

Guide 8: Remedy for GenAI Related Harms

1. Introduction

1.1 Background

The list of generative AI's (genAI) capabilities is lengthening by the day. GenAI models are used for direct interface with computers, rapidly creating life-like videos, assisting with coding, summarizing research papers into audio discussions, live translation, and much more. However, these capabilities also pose risks to people and society, such as the creation of synthetic nonconsensual intimate imagery or significant job loss in the creative industry. To address these risks, some companies that design or deploy AI technology have created AI governance systems comprising AI principles, model evaluations, risk/impact assessments, and technical risk mitigations. However, many of these governance systems do not adequately integrate human rights principles and methodologies, or do not include them at all.

There is thus a gap between the practices of companies and of stakeholders such as regulators, NGOs, and academics, who are increasingly using a "rights-based approach" to address the risks associated with genAI. Civil society has long been calling for companies to take a human rights-based approach to identifying, assessing, and mitigating risks associated with AI.

More recently, the EU Artificial Intelligence Act (the AI Act)—the world's first comprehensive AI regulation—aims to ensure "a high level of protection […] of fundamental rights." The AI Act places obligations on developers and deployers of AI systems to assess AI products for their impacts on fundamental rights, including a requirement that deployers of certain high-risk AI systems undertake "fundamental rights impact assessments."

A granular human rights analysis of the "value chain" of genAI is also important for a comprehensive understanding of human rights risks. The value chain of genAI refers to the different actors involved in creating and deploying genAI products, such as suppliers, foundation model developers, and individual users. Media attention has focused on the AI governance of a few frontier AI labs—OpenAI, Google DeepMind, and Anthropic—that build both foundation models (e.g., GPT-4o) and applications that are powered by them (e.g., ChatGPT). Less attention is paid to the other actors that comprise the genAI value chain, such as data vendors, although their decisions or omissions (i.e., failure to act) also shape human rights impacts.

The interdependence of the value chain means that decisions made by one actor may have cascading effects across the value chain. Dataset suppliers provide or label the data from which

models learn natural language and multimodal capabilities. Foundation model developers use that data to train their models. Downstream developers build products on top of those models. Deployers make decisions about where and how genAI products or features are deployed, which may create contextual risks. Individual users make choices, such as the prompts they enter into genAI chatbots, that can influence outputs. This interconnectedness means that considering the entire value chain is crucial to assessing the human rights impacts of genAI. This human rights assessment (HRA) of the genAI value chain does this by identifying and assessing human rights impacts across the genAI value chain.

This HRA was undertaken using methodologies based on the UN Guiding Principles on Business and Human Rights (UNGPs), including a consideration of the various human rights principles, standards, and methodologies upon which the UNGPs were built. It was informed by desk research, a variety of non-public human rights assessments (HRAs) of genAI products BSR has conducted over the past two years, and interviews with industry stakeholders representing all points of the genAI value chain, as well as with value chain adjacent stakeholders such as researchers and civil society organizations.

1.2 What This Human Rights Assessment Provides

This HRA identifies and assesses the actual and potential human rights impacts (i.e., risks and opportunities) associated with genAI through a value chain lens. It also identifies how different value chain actors are connected to potential impacts,[1] and proposes actions those actors may take to address risks and provide remedy for harms. This HRA consists of the following sections:

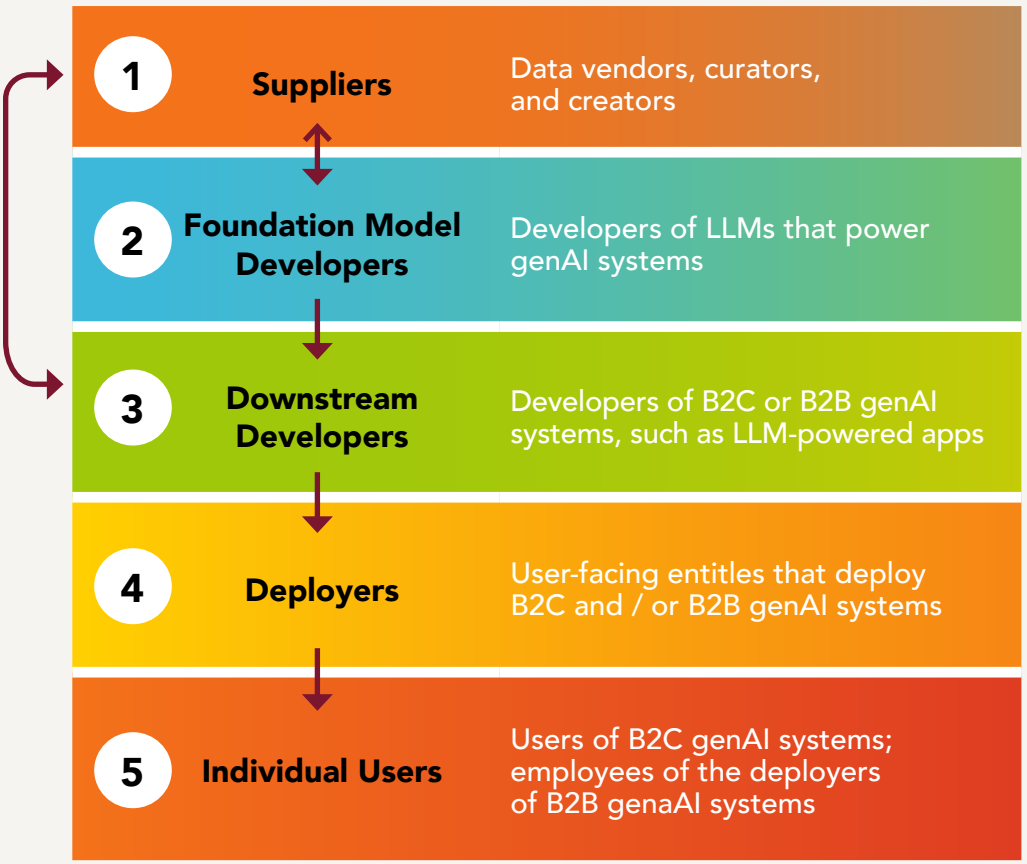
1. **Introduction:** The background and purpose of this assessment; an overview of a human rights-based approach to AI and its value, the genAI value chain, and key observations.
2. **Human Rights Assessment Methodology:** A description of the methodology this HRA uses to identify, prioritize, and make recommendations to address the human rights risks associated with genAI.
3. **Overview of the Generative AI Landscape and State of Play:** An outline of high-level factors that create or amplify the human rights risks associated with genAI.
4. **The Five Parts of the Generative AI Value Chain:** An explanation of the entities that make up the genAI value chain: suppliers, foundation model developers, downstream developers, deployers, and individual users.
5. **Human Rights Risks and Opportunities:** A high-level assessment of six categories of salient human rights risks for the genAI value chain, and how the actions or omissions of value chain entities may be connected to those risks.
6. **Recommendations:** Recommendations for value chain entities to address human rights risks.
7. **Appendix:** A glossary of terms and additional resources.

1 The technical term for this under the UNGPs is "attribution."

BSR has been conducting HRAs of AI products, services, and use cases for several years; however, prior to the publication of this assessment, few HRAs of AI have been published in full.[2] This has created challenges and reduced opportunities for learning and advancement on a shared human rights-based approach to identifying and assessing AI-related risks. We hope this HRA can therefore serve as a useful example for Responsible AI practitioners and encourage greater transparency across the field.

Accompanying this HRA are a series of practitioner guides that advise those who work on Responsible AI on how to incorporate a human rights-based approach to their work. These guides can be found here. In BSR's experience, Responsible AI practitioners have found human rights principles and methodologies to significantly improve work lows across risk assessment, risk mitigation, policy development and enforcement, and transparency and disclosure. These are necessary elements of AI governance that are not unique to genAI. Therefore, the practitioner guides may be used across all AI product development and deployment.

Furthermore, this assessment takes a "cross-value chain" approach to considering AI risks and mitigations. The value chain refers to the different actors involved in the creation, development, deployment, and utilization of genAI:



Suppliers of genAI system components, such as hardware or infrastructure, are not in scope for this HRA.

Some entities straddle multiple categories. For instance, OpenAI is a supplier, a foundation model developer, and a deployer

The double arrows between foundation model / downstream developers and suppliers indicates that these actors work closely, particularly during the training and evaluation process. Both types of developers may go back to suppliers multiple times for additional data to satisfy data needs for model fine-tuning.

Other actors, such as regulators or investors, may be "value-chain adjacent." This means that their activities have impacts across the value chain. They are excluded from this image for simplicity.

2 For some examples of published HRAs, see Microsoft's human rights assessment of its Enterprise Cloud and AI.

BSR created this representation of the value chain to clearly delineate how genAI experts across both technical and nontechnical domains understand the relationships of enablement between value chain actors. It was refined based on input from leading industry and civil society stakeholders. BSR notes that there is no definitive description of the genAI value chain, and other entities may depict the value chain differently.[3]

This assessment draws attention to the value chain to illustrate how decisions or omissions made by the full spectrum of value chain actors, not just foundation model developers or downstream developers, may impact human rights. For instance, the choices that suppliers make when curating the data that is used to train genAI models may have significant impacts downstream (e.g., a training dataset that contains personal information may be associated with privacy-violating data leakage in model outputs). More detailed analysis on this point is provided in Section 4: The Five Parts of the Generative AI Value Chain.

This HRA focuses on recommendations that speak to how value chain actors may influence others to mitigate risk. A core message of this HRA is that risks are the byproduct of interdependent actions or omissions; individual value chain actors can enhance their risk identification and mitigation efforts through collaboration. BSR accordingly explores broad interventions that apply to some or all value chain actors to mitigate risks with collective roots.

1.3 How to Read This HRA

This HRA is thorough and contains a lot of content. Depending on your background and interests, BSR recommends prioritizing the following sections:

5 minute read: Section 1.5 (Key Observations).

Responsible AI practitioners (e.g., Trust and Safety, AI governance teams) interested in learning more about the human rights-based approach: Section 2, Section 5, and Section 6 (reading both the "ecosystem recommendations" subsection and the subsection(s) that are relevant to their value chain entity). Also refer to the Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI.

Any staff at entities in the genAI value chain ("value chain entities"):

- **Suppliers (data vendors, curators, and creators):** Section 4.1, Section 4.2, Section 6.1, Section 6.2, and practitioner guides that are of interest.
- **Foundation Model Developers (entities that develop and pretrain foundation models):** Section 4.1, Section 4.3, Section 6.1, Section 6.3, and practitioner guides that are of interest.
- **Downstream Developers (designers of products or features powered by foundation models, or integrators of genAI into existing products, features, or services):** Section 4.1, Section 4.4, Section 6.1, Section 6.4, and practitioner guides that are of interest.

3 For example, see: https://one.oecd.org/document/DAF/COMP(2024)2/en/pdf#page=14.

- **Deployers (businesses that deploy genAI-powered products, features, or services directly to users):** Section 4.1, Section 4.5, Section 6.1, Section 6.5, and practitioner guides that are of interest.

Researchers and policymakers interested in understanding the value chain of genAI: Section 3 and Section 4.

Human rights professionals (e.g., from NGOs and IGOs) who wish to learn more about the genAI value chain and how it connects to human rights: Section 3, Section 4, and Section 5.

1.4 About BSR and Acknowledgments

BSR is a nonprofit consultancy and sustainable business network that works with the world's largest companies to achieve a just and sustainable world. This HRA and accompanying practitioner guides are authored by J.Y. Hoh, Samone Nigam, Lindsey Andersen, and Hannah Darnton, members of BSR's Technology and Human Rights team. This team specializes in advising companies on how to incorporate effective human rights practices into the development and deployment of technology. More than 100 of the world's largest tech companies are among BSR's members, including Google, Amazon, Meta, and Microsoft. BSR has conducted more than 120 HRAs of technology products, governance processes, and business operations; this includes several HRAs of genAI products.

BSR also collaborated with the B-Tech Project at the UN Office of the High Commissioner for Human Rights to produce three foundational papers on human rights and genAI:

- Advancing Responsible Development and Deployment of Generative AI: The Value Proposition of the UN Guiding Principles on Business and Human Rights
- Taxonomy of Human Rights Risks Connected to Generative AI
- Responsible AI and Human Rights: An Overview of Company Practice

These papers form the theoretical foundation for this HRA and the accompanying practitioner's guides, which seek to turn theory into practical advice and examples that companies may use to inform their operational decision-making. BSR thanks the B-Tech Project for providing input on this HRA and accompanying practitioner guides.

For their thought partnership and input, BSR also thanks Amy Winecoff and Ruchika Joshi (Center for Democracy & Technology), Dunstan Allison-Hope (independent), Kim Malfacini (OpenAI), Marlena Wizniak and Vanja Skoric (European Center for Not-for-Profit Law), Owen Doyle (Harris Research Group), Rashad Abelson (OECD), Richard Wingfield and Lale Tekisalp (BSR), and Stephen Pfohl and Unni Nair (Google).

1.5 Key Observations

› The human rights-based approach provides an internationally-tested methodology for identifying and mitigating risks to people and society.

Human rights complements rather than conflicts with existing ethics-based or trust and safety approaches to Responsible AI: there is significant overlap between the overarching purposes and core concepts of these methodologies (for more on this, see the "Understanding Ethics vs. Human Rights vs. Trust & Safety-based Approaches" section in the Practitioner Guide 3 on a "Human Rights-Based Approach to Risk Assessment"). For more on how Responsible AI practitioners may apply the human rights-based approach, please see the accompanying Practitioner's Guides, which contain detailed instructions on integrating human rights into various AI governance workflows such as impact assessment, stakeholder engagement, and policies and enforcement.

› Collaboration and communication across the value chain is needed for effective human rights risk identification, mitigation, and remedy.

Since value chain entities collectively make essential contributions to the final genAI product or feature, their cooperation is also necessary to identify and mitigate risk. For instance, effectively preventing genAI chatbots from generating hate speech may require updates to the application's safety filters, new fine-tuning of the foundation model, and the cleaning of datasets of harmful content and associations. Each of these are core activities of different value chain entities; they cannot be achieved alone. Value chain actors should establish effective means of communication with one another in order to collaborate on remedy and risk mitigation.

› Where direct remediation by businesses is required, and that remedy requires collaboration across the genAI value chain, the duty to coordinate that remedy should lie with a single point of contact.

A single point of contact ensures that remedies are accessible and adequate for affected stakeholders and prevents "finger-pointing" by value chain entities that may frustrate the effective resolution of grievances. This single point of contact should coordinate different elements of remedy across the value chain and communicate those actions to affected stakeholders.

› While value chain actors closest to a use case or deployment context may be better positioned to effectively identify and address human rights risks, they may lack the resources and skills to do so.

Large-scale developers and deployers may have the resources to create and maintain robust Responsible AI programs to identify and address risks. By contrast, many downstream developers and deployers are small entities and may lack the resources for such programs. Additionally, deployers may lack the technical expertise required to identify risks and take action to address them, particularly in the case of technical mitigation measures. They will need to coordinate with upstream actors that have the skills to intervene at the model and application levels.

› Suppliers' decisions can have cascading human rights impacts downstream, which makes it especially important for developers to influence their practices to be human rights-respecting.

Model training is a resource-intensive process, which makes it costly to retrain a model that encoded patterns in nonrepresentative or privacy-invasive datasets. The suppliers that control these datasets have direct commercial relationships with developers, who consequently have significant leverage to influence human rights-respecting data curation, documentation, and annotation practices. A proactive approach to supplier engagement mitigates downstream human rights risks from the beginning of the value chain.

› Comprehensive risk mitigation at the foundation model level may not be possible or appropriate.

Foundation models are general purpose, in that they are utilized in a vast number of downstream products, use cases, and contexts. At the foundation model development stage, use cases are likely undetermined. This makes it challenging to introduce targeted risk mitigations. Furthermore, introducing too many generalized mitigations at the foundation model level may also inhibit performance in downstream applications, because what is considered to be harmful in one context may be an acceptable use case in another. Foundation model developers are therefore limited in the range of safety mitigations they can introduce without obstructing the range of downstream use cases. This is especially the case for open source foundation models, for which developers have even less ability to be aware of and control downstream use cases and to introduce post-release safety mitigations.

› The extent of direct relationships between downstream developers and foundation model developers will impact these entities' ability to mitigate human rights risks.

Where there are direct relationships, this may enhance risk mitigation efforts because foundation model developers and downstream developers are able to work closely on implementing mitigations. However, downstream developers may not always have direct relationships with foundation model developers, which could in turn hinder risk mitigation efforts because it may be harder for downstream developers to get necessary information about the foundation model in order to implement appropriate and effective mitigations.

› Deployers are closest to the use of genAI systems, making them better positioned to identify risks related to their specific use case and/or issues that arise during deployment.

However, deployers may not have the requisite information or resources to implement technical mitigations for observed issues in system performance. Technical mitigations at the model level are more appropriately addressed by foundation model developers and by downstream developers at the application level. For this reason, the level of proximity a deployer has to the downstream developer may impact the effectiveness of ongoing risk mitigation. When there is less proximity between downstream developers and deployers, it may be difficult for deployers to alert developers when issues arise in a specific deployment context, potentially including instances of misuse or abuse.

2. Human Rights Assessment Methodology

This HRA was undertaken using methodologies based on the UN Guiding Principles on Business and Human Rights (UNGPs), including a consideration of the various human rights principles, standards, and methodologies upon which the UNGPs were built. It is informed by desk research and interviews with industry stakeholders representing all points of the genAI value chain, as well as with value chain adjacent stakeholders such as researchers and civil society organizations.

The core elements of an HRA are described below. BSR has included additional analysis in this assessment—such as an exploration of the genAI value chain and "risk pathways"—that is important for understanding the human rights impacts associated with genAI and how they can be addressed. This level of additional analysis is not typically necessary for HRAs targeted to specific products and services, or where the audience is already well-versed in the context.

2.1 Identifying Human Rights Impacts

Principle 11 of the UNGPs state that companies should respect all internationally recognized human rights. Human rights are indivisible, interdependent, and interrelated: the protection of one right can facilitate advancement of the others; the deprivation of one right can adversely affect others.

The development and deployment of genAI presents both risks and opportunities. Principle 11 of the UNGPs states that businesses "should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved," but also that companies "may undertake other commitments or activities to support and promote human rights, which may contribute to the enjoyment of rights."

However, Principle 11 also makes clear that positive impacts do not "offset a failure to respect human rights." For these reasons, it is important to note that when we list positive impacts in this assessment, they should not be balanced or offset against adverse impacts. This means that a company should not use the potential benefits of a genAI system to justify a refusal to

address its risks. Instead, companies, to the greatest extent possible, should seek to maximize the human rights opportunities while addressing risks.

In this HRA, BSR identifies potential human rights impacts using the universe of rights codified in the following international human rights instruments:

- The Universal Declaration of Human Rights (UDHR)
- The International Covenant on Civil and Political Rights (ICCPR)
- The International Covenant on Economic, Social and Cultural Rights (ICESCR)
- The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)
- The Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW)
- Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)
- International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICMW)
- Convention on the Rights of Persons with Disabilities (CRPD)
- The eleven International Labour Organization (ILO) Core Conventions
- The Convention on the Rights of the Child (CRC)

2.2 Rightsholder and Stakeholder Consultation

An HRA should involve meaningful engagement with affected stakeholders (also called "rightsholders")—people whose human rights may be impacted by the company—with particular attention to human rights impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization. Where direct engagement with affected stakeholders is not possible, the UNGPs suggest that companies should use reasonable alternatives, such as engaging with independent expert resources, human rights defenders, and other representatives from civil society.

The development and deployment of genAI products will impact over millions of users. It will also affect people who do not use genAI products themselves, but whose rights may be impacted by those who do or by the widespread adoption of genAI into systems across different segments of society. Identifying the connection of these impacts to different value chain actors, and their corresponding duty to remediate harms, was a key object of this HRA. (For more information on attribution, see Section 5.1.) For this reason, BSR consulted with a range of independent academics and civil society organizations to interpret the significance of the genAI value chain, including technical experts, Responsible AI practitioners, and organizations specializing in digital rights.

BSR also engages with a diverse range of stakeholders when undertaking human rights due diligence for companies across all industries. BSR supplemented stakeholder inputs from engagement conducted for this HRA with insights from previous HRAs that we conducted of various genAI products.

2.3 Prioritizing Human Rights Impacts

Principle 24 of the UNGPs acknowledges that while companies should address all their adverse human rights impacts, it is not always possible for companies to address them simultaneously, and therefore companies "should first seek to prevent and mitigate those that are most severe or where delayed response would make them irremediable." To inform prioritization, this HRA draws upon the human rights concepts of severity and vulnerable groups.

Severity: Scope, Scale, and Remediability

There are three main criteria for assessing severity:

- **Scope**—The number of people affected by the harm.
- **Scale**—The seriousness of the harm for the victim.
- **Remediability**—The extent to which a remedy will restore the victim to the same or equivalent position before the harm.

The UNGP-guided HRA methodology is a flexible one that may be applied to a specific product, service, policy, process, or operational decision. Alternatively, HRAs may also be conducted at a high level—to a family of products, a business strategy, or in the case of this HRA, the genAI value chain.

Because the genAI value chain is so broad and far reaching, billions of people may be impacted rightsholders. It is thus challenging to conclusively determine the scope, scale, or remediability of potential impacts, which will ultimately vary depending on the situation.

Scope may be inferred by reviewing current usage trends to estimate the baseline number of people who could be affected. However, given the sheer number of users of genAI products, the scope of the harm is almost always likely to be very large.

However, scale and remediability are much more difficult to extrapolate in the context of genAI products because the seriousness of the adverse impacts suffered by a rightsholder varies significantly according to the context in which it occurs. In this HRA, BSR has considered these criteria at a high level and with an understanding that they may vary from case to case.

BSR also typically considers the **likelihood** of the potential impact occurring in the near future. However, the broad scope of this report, focusing on the entire genAI value chain, makes a likelihood analysis less informative than one for a specific product. BSR notes that (1) there is certainty that bad actors will attempt to exploit genAI products and services, (2) many harmful outputs have already been produced by genAI products and services, and (3) genAI is still

in its infancy, and therefore it is necessary to be forward-looking because many impacts will emerge as genAI applications expand and evolve.

Vulnerable Groups

As established in the UNGPs, companies should pay "particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized." Vulnerable groups generally face heightened risks, or different risks, compared to others and are less likely to have their needs represented in decision-making processes. These groups may be disproportionately impacted by the adverse human rights impacts of genAI. For instance, minorities are more likely to be affected by discriminatory model outputs, while women and children are more likely to be impacted by synthetic nonconsensual intimate imagery.

Vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. These groups include, but are not limited to, human rights defenders, journalists, political dissidents, environmental and community activists, women, children, members of ethnic and religious minorities, indigenous groups, older people, members of the LGBTQIA+ community, disabled people, and those who are illiterate or have low levels of digital literacy. BSR's human rights methodologies for identifying vulnerable groups are based on four dimensions:

- **Formal Discrimination**—Laws or policies that favor one group over another.
- **Societal Discrimination**—Cultural or social practices that marginalize some and favor others.
- **Practical Discrimination**—Marginalization due to life circumstances, such as poverty.
- **Hidden Groups**—People who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants and sexual assault victims.

Vulnerability is heavily impacted by a variety of factors, including geography, culture, or the use case. In countries with a history of widespread human rights violations and/or conflict, vulnerable groups are especially at risk. There may be vulnerable groups outside of the typical categories who are vulnerable specifically in that geographic context.

2.4 Determining Appropriate Action

BSR's HRA methodology considers the appropriate action for an entity to address adverse human rights impacts using factors contained in Principle 19 of the UNGPs. Appropriate action is informed by two concepts: **attribution** and **leverage**.

Attribution assesses how closely connected the company would be to the human rights impact, where connection is determined using the following factors:

- **Caused the impact**—The company should cease or prevent the actual impact. In situations

of actual impacts having already occurred, the company should also provide remedy to affected individuals.

- **Contributed to the impact**—The company 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 incentivizing 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.
- **Directly linked**—The company is linked to the impact through its products, services, or operations arising from its business relationships, including with users. 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 incentivize those causing the harm to do so.

Leverage is the extent of the company's ability to effect change in the wrongful practices of an entity that causes harm. Companies may increase their leverage, such as by collaborating with other actors.

As with assessing the severity of a given impact, attribution and leverage will vary depending on the situation. Given the high-level nature of this HRA, we also discuss attribution and leverage at a high level.

Counterbalancing Rights in Tension

As noted above, all human rights are indivisible, interdependent, and interrelated. The protection of one right can facilitate advancement of others; the deprivation of one right can adversely affect others. For example, privacy is a necessary condition for the full realization, promotion, and protection of many other human rights, such as the rights to freedom of expression, freedom of assembly and association, freedom of movement, and freedom of belief and religion.

However, human rights can come into conflict with one another for legitimate reasons, and rights-based methods can be used to define a path forward when two competing rights cannot both be achieved in their entirety. Since human rights are interdependent and of equal importance, businesses cannot decide to "offset" human rights risks and benefits—i.e., using the benefits of a product or business decision to cancel out the need to address its risks.

Instead, it is important to pursue the fullest possible expression of both rights and identify how potential harms can be addressed. BSR has created a methodology named "counterbalancing," based on international human rights standards, to identify ways to secure the fullest possible expression of rights without unduly limiting others by applying established international human rights principles such as legitimacy, necessity, proportionality, and nondiscrimination. This methodology is consistent with the notion that most human rights are not absolute and can be limited in certain legitimate circumstances. For an example of counterbalancing, please see the section on "Addressing Tensions and Trade-offs" in Practitioner's Guide 4: A Human Rights-Based Approach to Risk Mitigation.

GenAI's impacts contain many instances of competing rights. One clear example of this is the tension between the helpfulness and harmfulness of genAI systems. Agents that helpfully answer various queries can enhance access to information, education, and science; but they can also output dangerous content and instructions that create risks to freedom from discrimination and bodily security rights. Defining how to balance these competing rights is challenging, particularly because there is no definitive hierarchy of human rights—none can be considered more important than others.

BSR's approach to counterbalancing rights in this HRA is shaped by the following established international human rights principles:

- **Legitimacy**—Restrictions to a right must pursue an objectively legitimate purpose and address a precise threat.
- **Necessity and proportionality**—Only restricting a right when the same goal cannot be achieved by other means, and using restrictions that are the least intrusive to achieve the legitimate purpose.
- **Nondiscrimination**—Restrictions to a right must be implemented in a nondiscriminatory manner.
- **Reverting to principle**—Focusing on the underlying principle of the right being restricted and identifying ways to uphold the core principle, even if not the exact right.

2.5 The Relationship Between Human Rights and Responsible AI

The human rights-based approach is based on internationally recognized human rights instruments, the first of which was the 1948 Universal Declaration of Human Rights (UDHR). At the heart of human rights is the recognition that all people have certain inalienable rights, such as the right to privacy or freedom from discrimination, that should be protected. These rights are listed in the UDHR and other human rights instruments, such as the International Covenant on Civil and Political Rights and the International Labour Organization's core conventions.

While human rights initially focused on protecting individual rights against government interference, the end of the 20th century saw increased stakeholder attention toward businesses' impacts on human rights as well. The UN Guiding Principles on Business and Human Rights (UNGPs), unanimously endorsed by the UN Human Rights Council in 2011, created a global standard for responsible business conduct that has been used by companies across all sectors to identify, assess, and mitigate business and human rights risks for more than a decade. This made the responsibility to protect and respect human rights universal for all companies and governments.

Over time, the international human rights system has developed a wealth of useful principles, concepts, and frameworks for prioritizing risks for action and determining appropriate mitigations that can be applied to help identify and address the spectrum of risks to

people and society associated with genAI. These are explored further in the accompanying practitioner guides.

Many companies currently use "ethics-based" normative frameworks that underpin AI governance systems.[4] These ethics-based frameworks, often anchored by a set of high-level "AI Principles" such as fairness, accountability, and transparency, are used to guide AI governance. Global attention to the genAI boom has triggered a proliferation of these ethical frameworks, some advanced by companies, while others have been authored by states or international organizations.[5] While there are some commonalities to these frameworks, there is still significant fragmentation in terms of the underlying ethical principles and how they are interpreted and operationalized.

With the growth of genAI, trust and safety-based approaches to Responsible AI are also becoming increasingly prevalent. Trust and safety professionals who previously worked at online platforms are increasingly being hired by genAI companies to help operationalize ethics and safety efforts. Trust and safety-based approaches focus primarily on the practical application of ensuring genAI tools are "safe," and are bringing approaches and lessons learned from online platform content governance. In contrast to high-level ethical principles, trust and safety-based approaches tend to anchor on pre-established taxonomies of harm based on how a product may be misused or abused.

Both ethics-oriented and trust-and-safety-oriented approaches to Responsible AI are often closely aligned with human rights. Companies need not choose between a human rights-based approach and an ethics-based approach. They can be complementary, with the human rights framework providing useful tools and methodologies to address common problems and dilemmas in AI governance. Given that AI products and services have global impacts that cross geographic borders, human rights can be a unifying force in AI governance because it has international legitimacy and is the normative framework with the strongest claim to universality for both governments and companies.

Human rights also form a "common language" that can facilitate further cooperation between AI companies and external stakeholders such as civil society and regulators. Furthermore, international human rights standards provide a wealth of ready-made concepts and frameworks that can be applied to address ethical problems or trade-offs in AI governance. The practitioner guides explore in detail how a human rights-based approach in line with the UNGPs can serve as the foundation for assessing and addressing the risks to people and society associated with AI—a foundation upon which other approaches can be integrated.

4 This is due to a range of reasons, including the academia-to-technology company pipeline and the absence of human rights-trained staff within these companies. See, for example, https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/overview-human-rights-and-responsible-AI-company-practice.pdf and https://www.chathamhouse.org/sites/default/files/2023-01/2023-01-10-AI-governance-human-rights-jones.pdf#page=12.

5 See, for example, the OECD's AI Principles, https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=The%20OECD%20AI%20Principles%20promote,stand%20the%20test%20of%20time or China's 'Governance Principles' for 'Responsible AI', https://perma.cc/V9FL-H6J7.

3. Overview of the Generative AI Landscape

The genAI industry has various factors that influence human rights risks and opportunities, as well as their likelihood of occurrence. This section highlights numerous "landscape" factors that provide essential context for the assessment of human rights impacts. One example of a landscape factor is the intense competition between AI companies, which may incentivize rapid model and product development at the expense of safety. Some of these factors extend beyond genAI, applying to AI more broadly. The issues raised are not exhaustive; rather they are those which are most likely to contribute to the overall human rights risk profile of genAI.

3.1 Characteristics of the Field

The following characteristics of the field of genAI development and deployment provide a foundational knowledge for assessing human rights risk associated with the technology and determining appropriate action to be taken by specific actors in the value chain to address risk.

› Competition:

The genAI ecosystem is characterized by significant competition between leading AI developers that has been described as an "arms race." While the competitive ethos may relax with time, initial staunch competition between genAI developers stymied immediate collaboration on common safety issues. While some collaborative efforts have since emerged (e.g., the Frontier Model Forum), concerns over commercial confidentiality or competitiveness remain an impediment to industrywide collaboration on safety.

› Acceleration:

The tense competition between genAI developers engenders a climate marked by accelerated innovation, as various actors move quickly to release new and competitive products. This includes accelerated innovation of foundation models, as well as new applications built on top of these models. During a February 2023 launch event for the new genAI-powered Bing search engine, Microsoft CEO Satya Nadella stated, "Rapid innovation is going to come. The race starts today." However, accelerated innovation has trade-offs, including

the inability to conduct thorough risk assessments and safety evaluations prior to product launch and to implement appropriate safeguards to mitigate risk. Such trade-offs may result in decreased model safety, which in turn could lead to downstream human rights impacts.

› Lack of transparency:

The competitive climate of genAI development has led to secrecy about the characteristics and processes related to proprietary models. While some foundation model developers—such as Meta—are moving in a different direction by publicly releasing information about how foundation models are built, trained, and evaluated, as well as their energy consumption, developers of other prominent foundation models are reluctant to release information about things like model architecture (including model size), hardware, training compute, dataset construction, training methods, and other details. Insufficient transparency makes it difficult for relevant stakeholders in, or adjacent to, the genAI value chain to consider safety issues, responsible use, or appropriate regulation of the technology. For example, insufficient transparency limits the ability of human rights experts to assess particular genAI models for their potential to lead to downstream human rights impacts related to training methods, training data, or other more technical model characteristics. An inability to perform robust human rights due diligence on a model increases its human rights risk profile. At the same time, increased transparency may pose its own risks, such as empowering bad actors to understand and circumvent safeguards. This may increase the likelihood of model misuse and abuse.

› Information asymmetries:

Actors at different points along the genAI value chain have varying levels of information about the factors and conditions that may shape potential downstream human rights impacts and their possible mitigations. For example, foundation model developers hold the greatest information pertaining to how a foundation model was built (training data, parameters, model architecture, etc.), while deployers have the most granular information about how the model is being used and the associated use case-specific risks. A downstream developer may be best positioned to deploy a given risk mitigation but may have insufficient knowledge about the foundation model or the deployment context to do so effectively.

› Resource discrepancies:

Not all actors in the genAI ecosystem are equally resourced. These resource discrepancies may create pathways to human rights risk. For example, while large companies that develop foundation models may have the resources to perform comprehensive risk assessment and model evaluation to understand and mitigate risks associated with their models, small-scale app developers may not have the resources necessary to perform robust safety evaluations prior to deployment. This may lead to products launching without comprehensive risk mitigation in place. Under-resourced data suppliers may also be constrained in their ability to procure, vet, and supply high-quality training data to foundation model and downstream developers.

› Lack of standardization:

While efforts are underway by domestic and intergovernmental standard-setting bodies to govern the ethical, responsible, and safe development and deployment of AI, regulatory efforts are still nascent. Furthermore, within the field of AI development, there are no widely agreed-upon normative thresholds for benchmarking model safety. In the absence of government regulation and standardized industry approaches, it is up to individual AI developers and deployers to determine minimum thresholds for model performance and safety.

The lack of standardization in model performance requirements and safety evaluations may be associated with adverse human rights impacts. For example, in the public sector, government agencies may purchase off-the-shelf genAI tools, or tools with some or full customization, that play a role in the delivery of public services. It is then largely up to the individual public sector agency to identify potential risks and determine appropriate levels of model safety, which may require technical skills and resources it does not have. Because many of the core functions of public sector agencies deal with the realization of rights, shortcomings in model performance or safety could directly impact human rights.

› Western-centric:

While non-Western companies are steadily emerging as developers in the genAI ecosystem, at the time of this assessment, the field of genAI is Western-centric. The majority of leading developers are American companies that operate primarily in Western countries, although leading Chinese firms have also developed powerful foundation models. There are also concerns that unequal access to the technology will increase the digital divide in the long run. If developers and deployers are not aware of gaps in representation and inequities in access, these asymmetries may go unaddressed in the development and adoption of genAI systems, potentially leading to adverse human rights impacts.

› Difficulties in measuring impact:

Even with safety evaluation processes in place and a comprehensive understanding of the risks posed by genAI, connecting risk to impact can sometimes be challenging. Some impacts may be easier to measure (e.g., the number of victims of scams that leverage genAI), while others may be less quantifiable. For example, while it may be possible eventually to measure misinformation-related misuse via watermarking and other provenance methods (e.g., the amount of AI-generated misinformation circulating online), it is more difficult to measure the downstream impact of this misuse (e.g., the extent to which AI-generated misinformation influences people's beliefs). The challenges in measuring the subtler, yet important, adverse impacts of genAI complicates effective mitigation, which increases human rights risk.

› Accountability:

Accountability for the impacts of genAI systems, and how that accountability may be enforced, raises complex questions. For example, who should be accountable when a genAI system is used as intended but is still associated with harm, when a system is misused to facilitate harm, or when harms result from system failures? In addition to pinning down "who" is accountable, there are questions pertaining to the "how" of accountability, with arguments for new regulation, risk/impact assessments, system evaluation and

auditing requirements, legal liability, and remedy for harms associated with genAI systems. Although many genAI developers take measures to reduce the likelihood of their products being connected to harm, a lack of transparency about such measures further obscures the question of accountability in risk mitigation processes. The current uncertainty and opacity around accountability impacts the human rights risk profile of genAI because effective human rights risk mitigation relies on clearly assigned accountability and transparent methods for enforcing it.

› B2C vs. B2B applications:

GenAI developers are releasing technologies both for use by individuals (B2C) and enterprises/government entities (B2B). Using the value chain outlined in Section 4, another way of thinking about this directional flow of products and services is from developer to individual user (B2C) or developer to deployer (B2B). Human rights risk profiles may differ significantly across and within B2C and B2B contexts.

› Sensitive domains:

The application of genAI in particular domains may pose increased human rights risk. Sensitive domains are those in which there are greater risks to human rights due to the nature of the activity, often involving high stakes decision-making that can have significant impacts on the lives of those affected. They include, for example, the criminal legal system, military, immigration, medical services, legal services, financial services, education, and public benefits (e.g., welfare). The integration of genAI into systems that facilitate the core operations or processes of these domains may therefore be inherently high risk and could perpetuate societal inequities. At the same time, because decision-making in sensitive domains is so consequential, improvements to operations in these domains enabled by genAI may have positive human rights impacts. For example, if genAI tools were used to process asylum applications more efficiently, rightsholders may be able to access state benefits or receive asylum decisions more quickly.

› Use case-specific risks:

At the time of this assessment, the adoption of genAI across sectors and domains is still in an exploratory phase. As adoption increases, high-risk use cases may emerge across domains. This could include use cases which require the processing of personal data or other sensitive data, such as biometric data, or the use of genAI models in ways that result in allocation harms (i.e., use cases designed to assist in the allocation of resources or opportunities that may make decisions that are imbalanced with regard to different social groups). Because genAI applications are in an experimental phase, ongoing human rights due diligence is necessary to identify high-risk use cases that may emerge in any sector.

3.2 Characteristics of Generative AI

Gen AI models and large language models (LLMs) (See the glossary in the Appendix for definitions) feature unique characteristics that shape their human rights risk profiles. To understand the actual and potential human rights impacts associated with the technology—

impacts that are occurring now and those which could emerge in the future—it is important to understand key characteristics about genAI models and LLMs, how they function, and their training and evaluation.

› Environmental impact:

Training and operating such large models requires huge amounts of computing power, which has a significant environmental impact. Pretraining of Meta's LLM, Llama 3, caused 2,290 metric tons of carbon dioxide emissions, equivalent to the average annual energy use of about 300 US homes. Additionally, increased reliance on AI requires greater data center capacity. Microsoft reported a 30% increase in emissions in 2023 corresponding to AI-related investments, notably the construction of new data centers. Some experts implore us to consider whether the value LLMs and associated use cases may add to society is worth the environmental cost. Others argue that the field of machine learning research and development must make carbon emissions and other resource considerations central to decisions about how the field innovates and evolves.

› Model size:

As the name implies, LLMs are very large machine learning models. They are trained on huge amounts of data (with recent models having over a trillion tokens and training parameters), and at such enormity it is not possible to thoroughly curate or document datasets ingested by the model. Dataset curation is important for ensuring data is balanced, fair, and representative, and that harmful data has been removed. Dataset documentation is important for understanding sources of bias in model outputs and identifying mitigation strategies. The immensity of LLMs and the prioritization of size over intentional dataset curation and documentation may lead to a number of challenges that have human rights implications. These challenges include dangerous capabilities (e.g., image generation models that are capable of producing synthetic Child Sexual Abuse Material (CSAM) due to the presence of CSAM in the training data), discrimination (e.g., LLMs that produce representational harms such as stereotyping, failing to perform equitably for all social groups, or misrepresenting or denigrating social groups), and harm without remedy (e.g., harms that, due to the size of the model, cannot be mitigated via identifying and understanding the training data characteristics that lead to the harm).

› Static data:

Machine learning identifies patterns in data and uses them to make predictions. Because machine learning models project patterns from the past (i.e., patterns found in past data) into the future (i.e., future-facing predictions based on past data), the outputs from machine learning models may be inherently conservative, and the widespread adoption of these models across particular societal domains may stymy social change. In the case of LLMs, the use of older text as training data may reify former, less inclusive world views that don't account for social progress. For example, even though some societies may be generally heading in a direction that increasingly acknowledges gender expansiveness, LLMs trained on dated text data may not reflect this (e.g., LLMs may assume that someone who uses the name "Max" must be male). Entrenching dated social norms and practices may in turn limit progress toward the socialization of human rights.

› Poor interpretability:

Sometimes used interchangeably with "explainability," "interpretability" refers to a human's ability to understand how or why an algorithmic system has produced a particular output. Generally, interpretability decreases as the complexity of a system increases. Neural networks, which underlie LLMs, are a highly complex machine learning process with low levels of interpretability. As developers build increasingly capable LLMs, they continue to have poor interpretability, making it challenging, if not impossible, to determine with full transparency the logic behind model outputs. While techniques are being developed to increase transparency of AI decision-making, poor interpretability may impact the human rights risk profile of genAI and LLMs. If developers are unable to understand why a model behaves a certain way or generates a particular output, it creates challenges for diagnosing and correcting issues in the model that may lead to bias, inaccuracy, toxicity, discrimination, or other harmful outputs.

› Emergent behavior:

As models scale they often develop novel capabilities not present in smaller models, a phenomenon referred to as "emergent behavior" or "emergent abilities." Emergent behaviors can pose different levels of risk. For example, relatively low-risk emergent behaviors include improved arithmetic and language understanding, while potentially high-risk emergent behaviors could include a model's ability to create and act on long-term plans, accrue power and resources, or act with greater agency, implying that the model is capable of performing tasks without command or human control. However, it is not only high-risk emergent behaviors that may impact human rights. The fact that models develop capabilities sometimes unknown to developers is an issue in and of itself. This is because it could have significant impacts on the ability of developers to identify dangerous capabilities in a model, design effective safeguards to mitigate them, and prevent model misuse.

› Training methods:

The pretraining stage involves LLMs. After pretraining, LLMs are fine-tuned through two further steps: supervised fine-tuning (SFT) and reinforcement learning (RL).

- **SFT:** A pretrained LLM will have powerful natural language capabilities, but its outputs often require further adjustment to be consistently helpful, harmless, and aligned with specific tasks. The pretrained LLM may produce generic, irrelevant, or unsafe responses because it has not been trained for specific tasks, such as following user instructions. For instance, when asked to "summarize this paragraph," a pretrained LLM might produce unrelated text or a continuation of the input rather than a concise summary.[6] Supervised fine-tuning (SFT) trains the LLM on a dataset of human-labelled conversations, enabling it to better understand and execute instructions, produce task-appropriate responses, and reduce the likelihood of harmful or unhelpful outputs.
- **RL:** Reinforcement learning builds on SFT by creating an ongoing process that allows models to be continuously updated. SFT is a one-and-done process: a dataset is collected and used to train the model, after which subsequent updates would require a new dataset being collected. By contrast, RL is dynamic: it uses an innovation named

6 https://openai.com/index/instruction-following/.

a reward model to update the LLM. The reward model is trained on human preference data to predict what types of outputs would be aligned with those preferences. Then, the reward model is used to update the LLM. Importantly, the human preference data can be dynamically collected and utilized to update the LLM. This includes user data while the model is deployed live; for instance, users of ChatGPT are sometimes asked which "response they prefer," an example of RL. This permits the LLM to be dynamically fine-tuned.

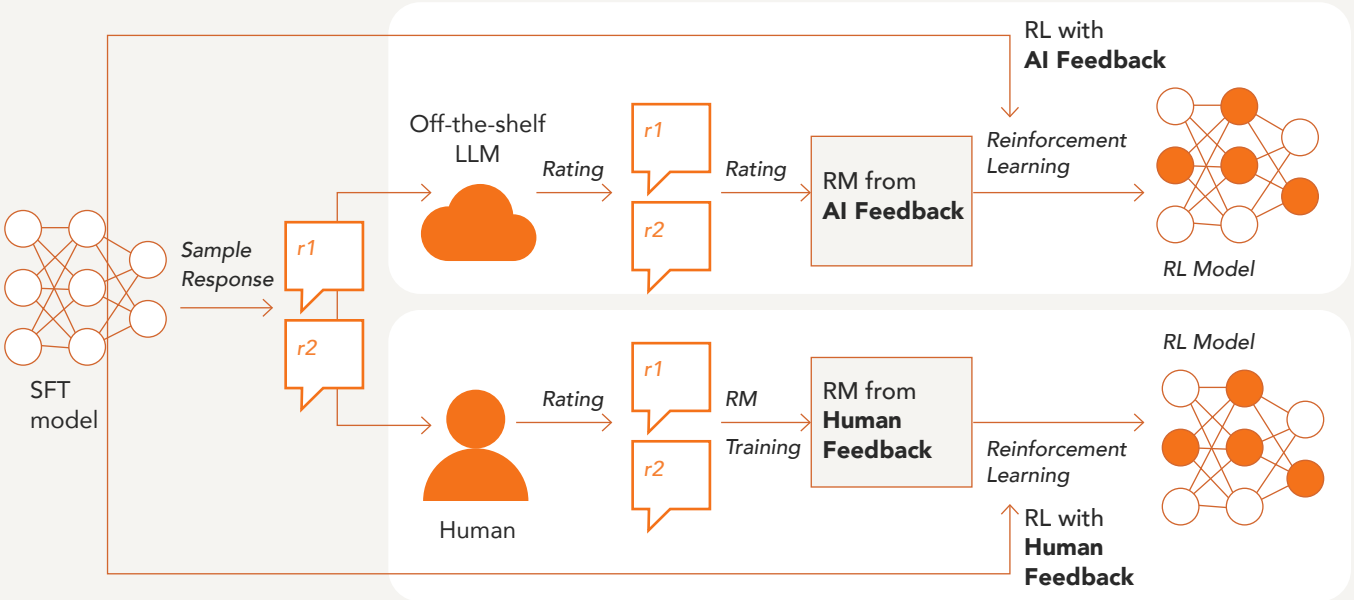
Two methods for RL are Reinforcement Learning through Human Feedback (RLHF) and Reinforcement Learning through AI Feedback (RLAIF):

RLHF: At the time of this assessment, RLHF is the most common strategy for fine-tuning LLMs after SFT and pretraining. In a three-step process, humans evaluate and label model outputs, and this in turn is used to improve model performance. In other words, through RLHF, models learn to predict what types of outputs satisfy human preferences. There are significant challenges associated with RLHF that may ultimately impact the human rights risk profile of LLMs trained using this method. RLHF can be costly, time-consuming, and difficult to scale, which may create barriers to adequate training as models trend toward increased size. The quality of human feedback may be inconsistent or suboptimal, or may be influenced by human annotators' harmful biases. Human annotators may not be sufficiently representative of diverse ages, backgrounds, cultures, experiences, or beliefs. Because of the reliance on a single reward function,[7] there are technical challenges in training models to represent the diversity of human preference and opinion. These challenges could potentially result in harmful model behavior—such as biased, nonrepresentative, inaccurate, sycophantic, or privacy-violating outputs—which may have associated human rights impacts.

RLAIF: RLAIF is proposed as an operable alternative to RLHF. With RLAIF, an AI assistant replaces human annotators to evaluate and label model outputs based on a defined set of principles, which is then used to improve model performance. Because this process does not substantially rely on human annotation, it may be quicker and more scalable. However, there are challenges associated with this method. Complex AI systems already lack interpretability, and using AI to fine-tune large models could further obscure AI decision-making. This could exacerbate above-mentioned issues associated with poor interpretability.

7 Also known as a "reward model" or a "preference model." Labeled outputs from a pretrained model are used to train a reward function, which is then used to optimize the LLM. For more information on the three-step process of RLHF see: 2203.02155 (arxiv.org).

A diagram depicting RLAIF (top) vs. RLHF (bottom)
Image Source: Google Research



› Sycophancy:

Sycophancy is the tendency of an LLM to respond to subjective user prompts in a way that aligns with the user's stated beliefs. This likely happens in part due to RLHF, a training method that rewards system performance that adheres to human preference. There is some evidence to suggest that during RLHF, human annotators may be likely to rate sycophantic responses more preferably, which in turn could cause models to produce outputs that appeal to the user even when the outputs are flawed or inaccurate. This tendency in model performance may impact users' ability to access quality and accurate information, and in an extreme, could serve to uphold biases by flattering, rather than challenging, users' potentially harmful beliefs.

› Expansive outputs:

While some prompts provided to a genAI system should elicit a single response, such as queries about commonly accepted facts (e.g., What year did India gain its independence from the United Kingdom?), other prompts can result in a large range of outputs that are all relevant to the prompt. This differs from other types of machine learning where correct outputs are finite. The expansiveness in potential responses to a single prompt creates challenges in defining model safety and fairness. Given that "harm" is not explicitly definable and is context-specific, the range of what constitutes harm may be as expansive as the possible responses to a single prompt. The amount of possible relevant outputs for a given input complicates risk mitigation efforts, which affects potential human rights impacts.

› Modality:

Foundation models are moving toward multimodal capabilities, meaning they can interpret inputs and produce outputs in multiple modalities, such as text and image. Different modalities present different levels of risk. For example, harmful images or video could be more immediately psychologically distressing to viewers than harmful text. Having a single tool that can process and generate content in a variety of modalities may pose greater risk than text-to-text models. Unique consideration should be paid to the potential human rights impacts associated with the increased pervasiveness of multimodal genAI systems.

› Impact beyond users:

In some application domains, system operators may be using genAI tools to make decisions that impact others. For example, in the healthcare setting, clinical care providers may use genAI tools that ultimately impact their patients. In the public sector context, public welfare agency employees may use genAI tools in processes that impact welfare applicants. In this way, individual users may not always be the final affected stakeholder. Value chain actors should therefore seek to understand the potential impacts of genAI systems not only on users, but on all potential affected stakeholders and society broadly.

› Closed vs. open source:

GenAI models can be released either as closed or open source or with varying levels of openness. Open source refers to when developers enable full or partial access to the model, its code, training data, and/or other training information. Closed source refers to models that are not widely nor freely available to the public, and where access, distribution, and use is typically controlled by the organization or company that developed the model. There is ongoing discussion about which is the safest strategy. Each option, and their gradients, affect the human rights risk profile of genAI.

Open source models offer greater transparency. They allow relevant stakeholders—such as researchers, academics, experts, policymakers, or others—to examine systems, diagnose issues, and identify potential sources of harm. Open-sourced models are also a significant boon to researchers, who may utilize them in their research projects at no cost. Open access to models also empowers actors other than well-resourced labs to innovate on AI research, increasing market competition. Broader access also facilitates the incorporation of diverse perspectives into AI systems and the industry more broadly, which could help mitigate bias and improve representation in system performance.

However, open sourcing a model makes it almost impossible to ensure safeguards remain effective. An open source model may be downloaded by anybody, severing the connection between the foundation model developer, which installs safeguards, and downstream actors. Downloadable open source models can be manipulated to change or override programmed safeguards and re-released. There is almost nothing a foundation model developer can do to mitigate risk once a manipulated model has been re-released or downloaded by other users.

Closed source models minimize the potential for model misuse or manipulation by enabling greater control by the original developer. Research has shown that value alignment in open source models is vulnerable to manipulation, and therefore closed source

models may be better at preventing the use of LLMs to produce harmful content. However, closed source models may also concentrate power among developers, limiting input from other relevant stakeholders whose perspectives could benefit the quality and safety of system performance.

3.3 Known Risks and Challenges Associated with Generative AI and LLMs

Collective understandings of LLMs and genAI models are evolving as rapidly as the models themselves. Because the field of genAI is characterized by accelerated development, new challenges present themselves on an ongoing basis, as do efforts to mitigate these challenges. Despite the rapidly evolving nature of the technology, there are known challenges that have been thoroughly documented. These challenges can lead to human rights risks, and it's therefore important to understand them to inform a comprehensive human rights assessment of genAI technologies.

Known challenges associated with genAI and LLMs are presented below. Solutions to some of these challenges are under development and, with time, this set of challenges may recede while new ones surface. For this reason, the list below is not exhaustive and should be regarded as a snapshot in time.

› Dangerous capabilities:

As LLMs increase in size, they sometimes develop capabilities that are not specifically programmed (see "emergent behaviors" above). Some emergent behaviors may be dangerous and pose a number of human rights risks. Dangerous capabilities could include cyber offenses, fraud, deception, human persuasion and manipulation, political strategizing, weapons acquisition, multistep planning, ability to create other AI models, situational awareness, self-proliferation, or privacy violations. Regardless of use case, models that display dangerous capabilities may be associated with adverse human rights impacts depending on the type of behavior and how long it takes to identify and address it.

› Inaccuracy:

GenAI models may present inaccurate information as though it is fact, referred to as "hallucinations." While ongoing improvements to genAI are reducing the occurrence of inaccuracies, eliminating this risk is not possible. Inaccuracies may be of greater consequence in high-risk use cases or sensitive domain applications, potentially leading to adverse human rights impacts.

› Inferring sensitive characteristics:

GenAI models may infer sensitive characteristics about an individual or a user by combining pieces of data or through user interactions. This ability has significant privacy implications and could lead to unforeseen human rights impacts, which may vary in severity depending on the application domain.

› Harmful content:

Harmful content includes any content that contains disinformation, bias, hate speech, suicide or self harm, terrorism or violent extremism, child sexual abuse material, or any other content that may be associated with psychological or other real world harm. While developers often implement safety guardrails to reduce the likelihood that genAI models produce harmful content, when this type of content is present in the training datasets, it is liable to surface in model outputs. The proliferation of harmful content could lead to downstream human rights impacts.

› Representational harms:

Representational harms refer to issues in how genAI models represent individuals, groups, communities, cultures, beliefs, or opinions. Due to quality issues in training data, models may learn demeaning stereotypes and/or perpetuate patterns of erasure. For example, assessments of text-to-image models illuminate a tendency of models to reinforce racial and gender stereotypes, promote American norms, and idealize whiteness. Additionally, some LLMs have been found to treat gender as binary, raising concerns about nonbinary and trans erasure. Representational harms may give rise to a number of related human rights risks.

› Jailbreaking / vulnerability to adversarial prompting:

Jailbreaking refers to successful user attempts to override a generative model's programmed safety guardrails in order to produce disallowed content. When this happens model outputs may divert away from human rights values and/or may be used to facilitate adverse human rights impacts (e.g., scams, phishing, fraud, etc.).

› Unevenly resourced GenAI value chain entities:

Large-scale GenAI developers and deployers may have the resources to create and maintain robust Responsible AI programs to identify and address risks. However, independent app developers or dataset vendors may lack the resources for such programs. Additionally, deployers may not lack the technical expertise required to identify risks and take action to address them, particularly in the case of technical mitigation measures.

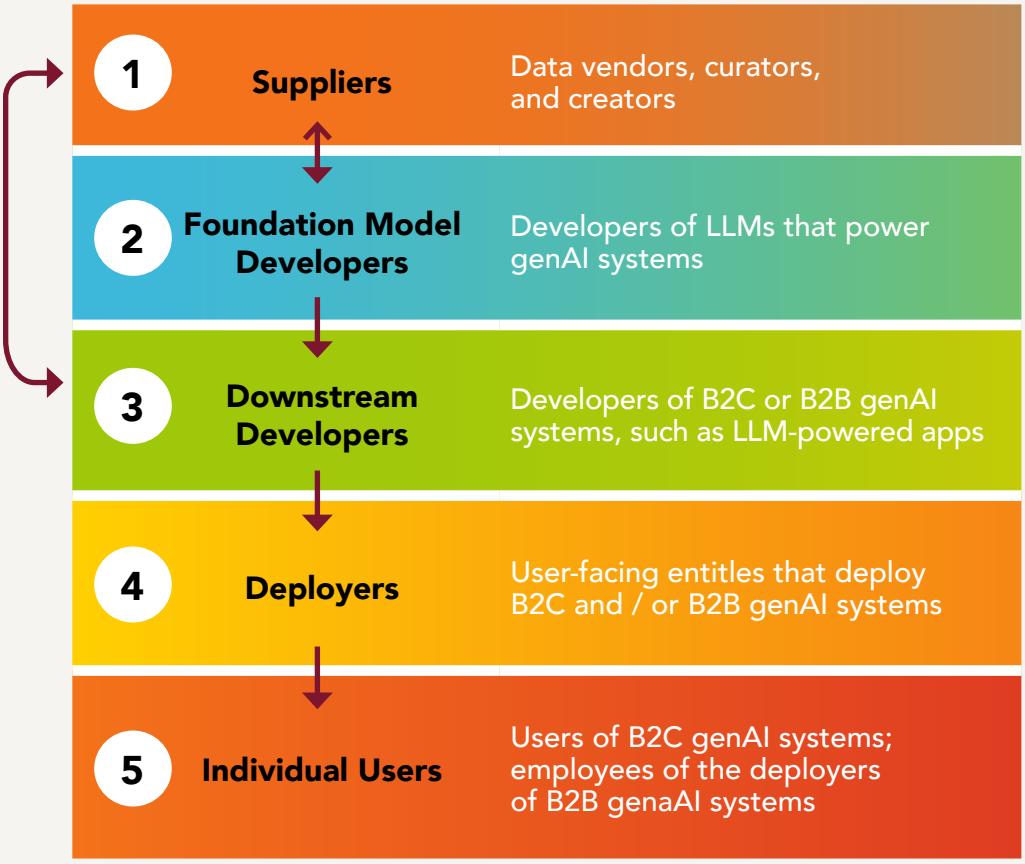
› Uneven communication between GenAI value chain entities:

A relationship between a data supplier and foundation model developer is a precondition for model creation. However, it is not commercially necessary for foundation model developers to engage with downstream developers—or for downstream developers to engage with deployers—beyond the sale of a product. Low or no engagement between these entities can create numerous pathways to human rights risk, including poor understanding of the product, a mismatch between model capabilities and the deployment context, or deployer choices that override safety guardrails.

4. The Five Parts of the Generative AI Value Chain

4.1 The Generative AI Value Chain

Understanding the genAI value chain is crucial for understanding how decisions made across the entirety of a genAI system's life cycle may contribute to downstream impacts on individuals and society. The genAI value chain is divided into five categories: 1) suppliers, 2) foundation model developers, 3) downstream developers, 4) deployers, and 5) individual users (see infographic below).



Suppliers of genAI system components, such as hardware or infrastructure, are not in scope for this HRA.

Some entities straddle multiple categories. For instance, OpenAI is a supplier, a foundation model developer, and a deployer

The double arrows between foundation model / downstream developers and suppliers indicates that these actors work closely, particularly during the training and evaluation process. Both types of developers may go back to suppliers multiple times for additional data to satisfy data needs for model fine-tuning.

Other actors, such as regulators or investors, may be "value-chain adjacent." This means that their activities have impacts across the value chain. They are excluded from this image for simplicity.

EXAMPLE A

OpenAI, GPT-4o, and ChatGPT

In this ChatGPT example, OpenAI is simultaneously a supplier (although it also sources data from other suppliers), foundation model developer, downstream developer, and deployer.

1. **Supplier:** OpenAI has partnered with data suppliers such as Shutterstock, Vox Media, and the Associated Press to obtain paywalled content, archives, and metadata. It also obtains datasets from other external sources, such as public datasets. Additionally, OpenAI is itself a data supplier, curating or obtaining data from internal sources, such as user data (with consent), synthetic data generated by its models, or datasets created by internal staff such as red teamers. These datasets are used to train and fine-tune its foundation models.
2. **Foundation Model Developer:** OpenAI develops foundation models such as GPT-4o and o1, which power the ChatGPT application. OpenAI is responsible for core model development and training workflows, such as model pretraining, fine-tuning, evaluations, and model-level mitigations such as model policies and output filters. It uses the datasets it obtains from internal and external sources for these workflows (e.g., datasets are used for model pretraining and fine-tuning).
3. **Downstream Developer:** OpenAI develops LLM-powered applications such as ChatGPT on top of its foundation models. OpenAI conducts system-level model fine-tuning, using SFT and RLHF to align GPT-4o to the purposes of ChatGPT, such as responding helpfully and harmlessly to user instructions. OpenAI also designs the ChatGPT application, including the interface and features that ChatGPT users interact with.
4. **Deployer:** In the B2C context of ChatGPT, OpenAI is the deployer, as it is the entity that directly interfaces with the user.
5. **Individual User:** The individual user is anyone who uses ChatGPT.

EXAMPLE B

B2B GenAI Call Summarization Tool (all entities fictional)

This example follows the value chain of a genAI tool used in enterprise settings to create automated summaries of video calls.

1. **Supplier:** A dataset vendor provides foundation model developers with a variety of datasets, such as multimodal data, to supplement public data for pretraining. The dataset vendor also provides downstream developers with smaller datasets curated for the eventual use case, such as videos of diverse groups of people speaking on professional remote call settings.
2. **Foundation Model Developer:** AI Lab X develops and trains its flagship foundation model, Matrix-1. Matrix-1 is a general purpose model, meaning it can be adapted to a wide range of downstream use cases. AI Lab X makes Matrix-1 available to third-party developers through an application programming interface (API), who may build on top of it to create genAI-powered products and features.
3. **Downstream Developer:** SoftCo, a medium-sized software company, accesses Matrix-1 through AI Lab X's API for a fee. SoftCo designs an app that uses Matrix-1's multimodal capabilities to process audio from videos and converts it into meeting notes. SoftCo contacts data suppliers to obtain curated datasets of videos of people speaking about work to fine-tune the model that powers the app. SoftCo puts its product on an app marketplace for enterprise customers.
4. **Deployer:** Z Corp, a large pharmaceutical company, purchases SoftCo's app. Z Corp then deploys it to its employees.
5. **Individual User:** Z Corp's employees use SoftCo's app to summarize notes of recorded meetings.

The human rights impacts of genAI detailed in the following section are organized by this five-part categorization of the genAI value chain. However, there are nuances to this categorization that are important to bear in mind:

› Relationships of enablement:

In some cases, the distinctions between value chain categories may be subtle, so for the purpose of this HRA they may be thought of as "relationships of enablement." Upstream categories of the value chain enable the core operations of the categories downstream of them. For example, because developers who build foundation models enable downstream developers to build apps and programs that leverage those models, foundation model developers and downstream developers belong to different value chain categories. Using this framing, suppliers enable the core operations of foundation model and downstream developers, which enable the core operations of deployers and individual users. When relevant, deployers may also enable the core operations of individual users. For example, an organization that deploys genAI tools internally enables its employees to use these tools as a function of their jobs.

› Overlapping categories:

GenAI value chain actors may fall into more than one category. The five categories of the value chain are defined by the activities and operations associated with each. A single company, or different teams within a single company, may engage in activities and operations associated with all parts of a genAI system's life cycle, placing the company in multiple value chain categories. For example, OpenAI is simultaneously a developer of foundation models (e.g., GPT-4), a developer of AI systems (e.g., ChatGPT), and a deployer of those tools (i.e., it has a direct interface with ChatGPT's users). The type of human rights due diligence performed at every step of a genAI system's life cycle should be informed first and foremost by the activity underway. For example, an actor that is primarily involved in deployment of genAI systems may also engage in supplier activities—such as providing proprietary datasets for model fine-tuning— and should conduct appropriate human rights due diligence given their involvement as both suppliers and deployers.

› Nonlinear:

The value chain is nonlinear in nature, and operations may move upstream before eventually moving back downstream. For example, a foundation model developer may procure data from a supplier for foundation model pretraining. After initial deployment, foundation models may be fine-tuned to execute effectively on specified tasks or use cases, which may occur at the foundation model and application levels. This requires the foundation model and downstream developers to go back to data suppliers for additional training data. This phase often happens between the initial development and deployment of the model.

› Value chain-adjacent actors:

There are a number of relevant stakeholders not captured in the below sections who are adjacent to the value chain and impact the genAI ecosystem and associated human rights

risks. Public, private, academic, and civil society stakeholders—such as policymakers, investors, researchers, and human rights practitioners—impact the environment of genAI development and deployment through regulation, financing operations, research, standards setting, and advocacy. Value chain-adjacent actors are also driving conversations related to harm and safety in the context of genAI. These stakeholders hold unique leverage to illuminate and push for mitigation of the human rights risks associated with genAI.

› Affected stakeholders:

Lastly, people may be affected by operations along all points of the value chain. In some cases the user of a genAI system may not be the most affected stakeholder. For example, a paralegal who uses genAI tools when putting together documents for a client's court proceedings is the system's user, while the client is the affected stakeholder. The affected stakeholders of a genAI system depend on the use case, operating context, and application domain. However, some individuals and communities may be more vulnerable to risk than others. An assessment of the most vulnerable rightsholders can be found in the human rights tables in section 5, "Human Rights Risks and Opportunities."

All actors along the genAI value chain can increase the effectiveness of human rights impact identification and mitigation through collaboration. Even if some mitigations are most effective if introduced at specific points in the development or deployment process, human rights due diligence should be conducted by all actors along the value chain, at all points in a genAI system's life cycle, and on an ongoing basis.

The following five subsections are deep dives into each value chain actor, fleshing out their roles, core activities, and relationships of enablement with other value chain actors. The sections also include key observations on how each value chain actor's position and relationships create human rights risks or opportunities to address those risks.

Overview of the GenAI Value Chain

Value Chain Entity	Core Activities	Risk Pathways
Suppliers Data vendors, curators, and creators	• Data curation • Data sourcing • Dataset curation • Data labeling, annotating, and enriching • Dataset documentation	• Insufficient or misleading dataset documentation • Failure to engage with downstream applications • Misalignment between dataset and downstream use • Failure to clean biased data • Unrepresentative dataset curation • Poor data labeling guidance • Ethical remuneration and working conditions for data labelers, annotators, and creators
Foundation Model Developers Developers of LLMs that power genAI systems	• Data procurement and creation • Model pretraining • Model evaluation • Model fine-tuning • Decisions about model release	• Insufficient or inappropriate safeguards • Limitations in model training techniques • Incomplete or overconfident model evaluation • Choices pertaining to model or system design • Choices pertaining to model release • Insufficient transparency
Downstream Developers Developers of B2C or B2B genAI systems, such as LLM-powered apps	• Data procurement • System evaluation • Model fine-tuning • Development of technical safeguards	• Ineffective technical mitigations • Choices pertaining to model fine-tuning • System design or intended deployment context • Lack of feedback channels for deployers and individual users • Invalid or insufficient evaluation
Deployers User-facing entitles that deploy B2C and / or B2B genAI systems	• Integration • Capacity-building • Oversight of use	• Limited genAI literacy and/or an absence of safety culture • Choices pertaining to deployment and integration • Ineffective human oversight • Ineffective scalable oversight • Limited awareness pertaining to developer safeguards
Individual Users Users of B2C genAI systems; Employees of the deployers of B2B genAI systems	• Prompt creation • Use case-related operations	• Intentional manipulation, misuse, or abuse • Unintentional misuse or harm • Limited genAI literacy • Overreliance • Automation bias

4.2 Suppliers

In the context of genAI, suppliers are upstream actors that supply the necessary components that enable the development and deployment of genAI systems. These components generally include data, software, hardware, and infrastructure such as data centers. Because of the immediate and pressing need for vast quantities of continuously refreshed data to train genAI systems, this HRA covers only those suppliers involved in processes related to data procurement. However, it is important to remember that there are salient human rights risks associated with hardware, software, and infrastructure components of genAI systems, including the environmental costs of model training and inference. See the Appendix of this assessment for more information.

Foundation models use supplied datasets for three broad purposes:

- **Model pretraining:** The pretraining stage entails the model ingesting huge amounts of data to learn natural language capabilities (see the below section on foundation model developers for more info). This data is typically publicly available and/or hosted on open source platforms. The data suppliers involved in this phase include content creators of original works or thought, such as users who provide user generated content (UGC) to online platforms like Wikipedia, Reddit, or social media platforms; or writers and artists whose works are used to train genAI systems, with or without their knowledge and consent.
- **Fine-tuning:** Following pretraining, smaller datasets are used to optimize model performance for a specified purpose or application in a particular domain. This is known as model "fine-tuning" (see below sections on foundation model developers and downstream developers for more information about fine-tuning). Data suppliers also include those involved in the creation of datasets for the purpose of fine-tuning.
- **Mitigations such as model evaluation and red teaming[8]:** Model safety testing also requires datasets, such as of adversarial prompts or unsafe outputs, to test a model's safety performance. For instance, a dataset of adversarial prompts may be fed into a model, and red teamers will assess how often the model refuses those prompts or returns an unsafe response as part of a pre-deployment risk assessment.

While suppliers can be independent data vendors, there is often overlap in the actors involved in supplying data and those involved in the development or deployment of a genAI system. For example, after the initial pretraining of an LLM, foundation model developers may themselves also act as suppliers, curating datasets for model fine-tuning (see Example A above). Additionally, downstream developers may work closely with the intended deployers of a genAI system to curate training datasets that meet use case-specific needs. This could involve sourcing proprietary data from deployers (e.g., if the deployer is a healthcare organization, this could involve sourcing data from medical records; if the deployer is an actor in the legal system, it could involve sourcing data from case records; etc.).

8 Red teaming refers to a range of impact assessment methods for AI systems that involves using adversarial techniques and approaches to test the security, robustness, and resilience of AI systems. Red teaming aims to identify vulnerabilities, weaknesses, and potential threats in AI models, algorithms, and systems by simulating both normal / expected user behavior and adversarial attacks and scenarios. This knowledge can then be used to strategize to address identified gaps.

The creation and curation of datasets typically involves the following activities:

- **Data creation:** A variety of actors create data, including those who created it for purposes other than model training and fine-tuning (e.g., the Vox Media journalists who initially wrote the articles that comprise the dataset they now supply to OpenAI). Additionally, in-house red teamers or human data labelers may also create datasets to inform safety mitigations. AI companies are also increasingly using genAI models to create synthetic data for model training and fine-tuning.
- **Data sourcing:** This could happen through a crowdsourced or participatory data collection process, web scraping, or other means.
- **Dataset curation:** This may include filtering out undesirable data from a dataset such as hate speech, pornographic content, etc., and review of data for relevance and representativeness.
- **Data labeling, annotating, and enriching:** These are processes that classify, categorize, and provide explanation and context for data within a dataset. This may also include the addition of metadata to establish data provenance, a record of the origin and processing of the data in a dataset.
- **Dataset documentation:** This is a process of documenting properties about a dataset including its composition and intended uses.

Risk Pathways

Suppliers' decisions can have cascading human rights impacts downstream, which makes it especially important for foundation model developers and downstream developers to influence supplier practices to be human rights-respecting. Model training is a resource-intensive process because of the size of pretraining datasets and associated compute costs. This makes it costly to retrain a model that has encoded patterns in unrepresentative or privacy-invasive datasets. The suppliers that control these datasets have direct commercial relationships with developers, who consequently have significant leverage to influence human rights-respecting data curation, documentation, and annotation practices. A proactive approach to supplier engagement mitigates downstream human rights risks from the beginning of the value chain. Examples of risk pathways that may occur at the supplier level include the following:

- **Insufficient or misleading dataset documentation:** Data suppliers may provide ineffective dataset documentation that does not adequately capture the potential uses and limitations of a given dataset. Developers may then use datasets for training that are not fit for their intended purposes, which could result in gaps in model performance. Data suppliers may also omit data provenance in their dataset documentation, which may make it difficult for developers to identify the source of issues related to model outputs. The lack of provenance signals may lead to developers inadvertently training their models on synthetic data, which can decline model performance (known as "model collapse").
- **Failure to engage with downstream applications:** Data suppliers may overlook the practical contexts in which their datasets are eventually deployed, resulting in a mismatch

between the content provided and the real-world needs of developers. Without a clear "know-your-customer" framework or a robust understanding of how the data will be used, suppliers risk enabling the blind purchase or use of data, which can amplify biases, limit model functionality, or degrade overall performance.

- **Misalignment between dataset and downstream use:** Some datasets may be excessively broad or contain material irrelevant to a specialized domain, such as medical applications. In these contexts, incorporating data sources like social media content can unnecessarily introduce noise and confuse model outputs. By failing to tailor datasets to the scope of the intended application, developers may encounter reduced efficiency, compromised model accuracy, and an overall decline in the reliability of the final system.
- **Failure to clean biased data:** Data suppliers may fail to recognize and rectify entrenched biases in their foundational datasets, particularly for sensitive fields like healthcare. For example, historical pain scale datasets that underestimate women's experiences of pain can generate skewed outputs when integrated into modern models. If suppliers do not proactively identify and mitigate such issues, they risk perpetuating inequitable outcomes and undermining the reliability of downstream applications.
- **Unrepresentative dataset curation:** Data suppliers may, inadvertently or otherwise, curate datasets that are not sufficiently representative of communities, regions, languages, or other relevant features. They may include content in a dataset that could result in harmful model outputs, such as personally identifiable information, copyrighted materials, or toxic or otherwise harmful content.
- **Poor data labeling guidance:** Data labelers and annotators may be operating with insufficient guidance, which may result in datasets that are not useful for developers' purposes.
- **Ethical remuneration and working conditions for data labelers, annotators, and creators:** Data suppliers may not take sufficient measures to ensure safe and fair working conditions for data labelers and annotators. Suppliers may also unfairly compensate, or not compensate, data creators.

4.3 Foundation Model Developers

Foundation model developers are actors involved in building the underlying models that power genAI systems. Examples of foundation models include OpenAI's GPT-4, Google's Gemini, Meta's LlaMa, Anthropic's Claude, Cohere AI's Command, and Stability AI's Stable Diffusion, among others. Foundation models are general purpose in their nature and may serve as base models for a wide range of applications.

Foundation model developers sit downstream of suppliers and rely on suppliers to provide the inputs necessary to build and train foundation models. These actors engage in the following activities and operations:

- **Data procurement and creation:** Foundation model developers obtain datasets for

pretraining, fine-tuning, and technical mitigations through a variety of sources. One source is whatever is already available in the data ecosystem, such as publicly available online text corpora or evaluation datasets. They may also procure data directly from suppliers through partnerships or one-off sales. In these cases, foundation model developers provide suppliers with specified guidelines for their data needs, including requirements pertaining to data structure, scope, annotation, or other aspects.

Foundation model developers will also procure data from datasets they already have access to as part of a pre-existing business. For instance, social media platform X's foundation model is trained on X user data. In addition, specialized teams at these developers may create datasets for specific purposes. For example, red-teaming staff may create datasets of adversarial prompts to test the model. Increasingly, AI companies are using models to generate synthetic datasets for these and other model training and fine-tuning purposes.

- **Model pretraining:** This is a process through which the base model is created. It requires selecting the training dataset(s), model architecture, and training algorithms.
- **Model evaluation:** These are tests that foundation model developers conduct to assess model performance, capabilities, limitations, and risks. Foundation model developers make decisions about what to test for, which often include considerations for model safety, fairness, and accuracy. In this process, they may work closely with data suppliers to procure additional datasets to fill gaps in model performance that surface during the evaluation process.
- **Model fine-tuning:** This is a process through which the pretrained model is modified to optimize performance or safety, or for particular tasks or domains of knowledge. The fine-tuning process relies on smaller, more specialized datasets. Fine-tuning includes supervised fine-tuning (SFT) and reinforcement learning (RL). For more information, see the bullet on "Training methods" in Section 3.2 "Characteristics of Generative AI and LLMs."
- **Decisions about model release:** Once a foundation model is built, developers make decisions about how to release the model and to whom. This encompasses decisions about making the model open source or closed source, enabling API access, making it downloadable, and other considerations.

Risk Pathways

Unlike deployers and downstream developers, foundation model developers are several layers removed from the individual user. Thus, they may have limited visibility into the downstream applications of their models. Foundation model developers may also lack expertise in the numerous domains in which their general purpose foundation models can be used, limiting their ability to implement effective and context-specific risk mitigations.

Additionally, comprehensive risk mitigation at the foundation model level may not be possible or appropriate. Foundation models are general purpose, in that they are utilized in a vast number of downstream products, use cases, and contexts. At the foundation model development stage, use cases are likely undetermined. In some cases, foundation models may be

developed solely for research purposes, with no product integration intent (although this can still happen later). This makes it challenging to introduce targeted risk mitigations.

Introducing too many generalized mitigations at the foundation model level may also inhibit performance in downstream applications. This is because harm is often context-specific. For instance, an edtech chatbot product for children should not output the phrase "I will kill you," but that phrase may be an acceptable output for a fiction-writing product. Fine-tuning a model to refuse that phrase in general would then inhibit the model's applicability to fiction writing and possibly other use cases. Foundation model developers are therefore limited in the range of safety mitigations they can introduce without obstructing the range of downstream use cases. This is especially the case for open source foundation models, for which developers have even less ability to be aware of and control downstream use cases and to introduce post-release mitigations.

Nonetheless, decisions or omissions at the foundation model level may create risks that can lead to downstream adverse human rights impacts. Examples of risk pathways that may occur at the foundation model level include the following:

- **Insufficient or inappropriate safeguards:** Safeguards implemented at the foundation model level may not be sufficient or appropriate for mitigating downstream risks. Technical safeguards may be easily undone by downstream actors. In other cases, safeguards at the foundation model level may limit downstream model performance. For example, safeguards that block toxicity in model performance may limit use cases related to content moderation or the study of hate speech for legitimate research or academic purposes.
- **Limitations in model training techniques:** Some training techniques that foundation model developers utilize could enable downstream risks. For example, the use of a single reward function for training foundation models has been criticized for its limitations in capturing a diversity of human opinion. Lack of diversity in model outputs could in turn lead to other downstream risks.
- **Incomplete or overconfident model evaluation:** Model evaluation, the process of measuring model performance or impact, is a core safety risk mitigation workflow for foundation model developers. However, there are numerous pathways to risk for model evaluation; developers may not conduct model evaluations that cover the range of model risks, or they may conduct evaluations but fail to integrate results to meaningfully change the model. Model evaluation itself has limitations as a mitigation technique—evaluations often cannot capture the complex real-world impacts of LLMs, and many evaluations are focused on capabilities rather than impacts on people and society. If these limitations are not grasped by foundation model developers internally or adequately conveyed when disclosing evaluation results, developers and downstream actors may develop a false sense of security. A more detailed examination of the limitations of model evaluation may be found here.
- **Choices pertaining to model or system design:** When building a machine learning model or system, developers make various design choices about the model, such as its archi-

tecture, the choice of the loss function, or hyper-parameters. These choices affect model behavior and, accordingly, may have downstream impacts. For instance, model design choices may impact algorithmic fairness: loss function choice for a facial analysis model can influence its accuracy at tasks such as recognizing visual attributes like skin or hair color. Impacts on human rights beyond freedom from discrimination, such as privacy or personal security, may also be impacted depending on the use case and context.

- **Choices pertaining to model release:** Foundation model developers make decisions about the gradient of release for LLMs. Models may be released fully open, open source with modifiable model weights, downloadable, available through API access, or other degrees of openness. Choices pertaining to how a model may be accessed, and by whom, once it is released impacts the potential for that model being leveraged in harmful ways. There are unique challenges for monitoring downstream use of open source models. Models with released model weights may be modified, particularly by bad actors who can change or remove safeguards.
- **Insufficient transparency:** Foundation model developers may not release sufficient information about the model that would enable downstream developers to identify and mitigate risks.

4.4 Downstream Developers

Downstream developers are actors that leverage foundation models to build apps, programs, products, and systems, or to embed them into existing software and systems. One example of a downstream developer is a builder of an LLM-powered app that accesses the foundation model through an API, then sells it to a telecom company, which deploys it to consumers. Another example would be the creators of a new LLM-powered feature for integration into existing enterprise software, such as the Zoom product teams that built the AI Companion feature.

This value chain category **does not** include the user-facing deployment of the genAI model or genAI-powered product for general use (this is captured in the "deployer" category). Instead, downstream developers focus on building products or features on top of foundation models and fine-tuning for specific use cases. They may directly sell or provide products or features to enterprise customers, which then act as deployers. The line between deployer and downstream developer is blurry, and some entities are both downstream developers and deployers, especially in the B2C context. For instance, OpenAI is a foundation model developer, downstream developer, and deployer:[9]

- **Foundation model developer:** Develops the GPT and o1 series of foundation models.
- **Downstream developer:** Builds applications on top of those models, such as the ChatGPT chatbot and DALL-E image generator.
- **Deployer:** Deploys ChatGPT and DALL-E directly to consumers.

9 OpenAI is also a supplier, as noted above in the section on suppliers.

Downstream developers will sometimes interact directly with upstream suppliers in the context of fine-tuning or prompt engineering. While downstream developers may be able to directly fine-tune through APIs, others may go directly to data suppliers for datasets for model fine-tuning, model evaluations, or other workflows that require data.

Downstream developers are upstream of deployers and individual users. They may build technologies intended directly for consumer use, or for public sector or enterprise deployers. Downstream developers may have varying levels of proximity to deployers. Downstream developers could offer off-the-shelf tools for purchase through platforms that require little direct engagement with the purchasing deployer. Alternatively, downstream developers may work closely with deployers to customize genAI tools for the deployer's specified purpose. In other cases, downstream developers may work with distributors or resellers that conduct some or all of their sales. In these cases, the downstream developer may have no contact with deployers at all.

Whereas foundation models are domain-agnostic and general purpose, products built by downstream developers are generally intended for a specific use case and/or for use in a specified domain. Exceptions include the integration of foundation models, such as GPT-4o as well as Gemini 1.5 Pro and Gemini 2.0 Flash, into general purpose chatbots such as OpenAI's ChatGPT and Google's Gemini. Downstream developers' proximity to individual use cases and application domains give them greater insight into the risks associated with a given use case and their potential mitigations. However, downstream developers may not always have the resources necessary to invest in risk mitigation efforts.

Downstream developers engage in the following activities and operations:

- **Data procurement:** As with foundation model developers, downstream developers engage data suppliers to procure data for fine-tuning, testing, and evaluating models for a given domain or use case. They may provide suppliers with specified guidelines for their data needs including requirements pertaining to data structure, scope, annotation, or other aspects.
- **System evaluation:** As with foundation model developers, downstream developers should ideally evaluate their genAI systems to ensure optimal performance. Evaluations may include considerations for system safety, fairness, and accuracy. In this process, downstream developers may (but not always) collaborate with data suppliers to procure additional datasets to fill gaps in system performance that surface during the evaluation process.
- **Model fine-tuning:** As with foundation model developers, downstream developers will likely choose to fine-tune the foundation model to optimize its performance for a given domain or use case. This requires downstream developers to procure additional datasets for the fine-tuning process. The datasets may either be public or, if specialized data is required, procured from suppliers. There is no limit to the number of times a model may be fine-tuned.
- **Development of technical safeguards:** This entails the use of product-level mitigations such as filters and classifiers to limit model outputs. Technical safeguards are likely specific

to the specific application domain, or use case. However, in general, downstream developers are concerned with limiting a model's ability to produce outputs that are factually incorrect, discriminatory, biased, contain toxicity, or are otherwise harmful.

Risk Pathways

Downstream developers may have varying degrees of proximity to foundation model developers. Where there are direct relationships, this may enhance risk mitigation efforts because foundation model developers and downstream developers are able to work closely on implementing mitigations. However, downstream developers may not always have direct relationships with foundation model developers. That, in turn, could hinder risk mitigation efforts because it may be harder for downstream developers to get necessary information about the foundation model to implement appropriate and effective mitigations.

Actions and omissions by downstream developers may create risks that can lead to downstream adverse human rights impacts. Examples of risk pathways that may occur at the downstream development level include the following:

- **Ineffective technical mitigations:** Downstream developers may implement ineffective technical safeguards, such as output filters that do not adequately cover the range of potentially harmful outputs.
- **Choices pertaining to model fine-tuning:** Downstream developers often fine-tune foundation models to optimize them for a specified purpose. Choices pertaining to fine-tuning could lead to downstream impacts, particularly if there is not sufficient consideration given to how models will perform across demographic groups, representation in model outputs, or other safety issues.
- **System design or intended deployment context:** Downstream developers leverage general purpose foundation models to build technologies for a specified purpose or for use in a specific domain. Downstream developers may choose to build technologies that directly or indirectly impact human rights. For example, following the release of open source image-diffusion models, there was a sharp increase in the availability of apps that market the ability to "undress" women based on clothed photos. In other cases, technologies built on foundation models may be more likely to impact human rights due to the application domain. For example, genAI tools intended for use in sensitive domains—such as healthcare, finance, legal, justice, or other high-stakes domains—may be more likely to be associated with adverse human rights impacts due to inaccuracies or discriminatory outputs, even if used as intended. The business model B2C or B2B may also have human rights impacts, which may vary depending on decisions to design products for consumers or enterprise/public sector customers.
- **Lack of feedback channels for deployers and individual users:** Downstream developers may not enable a means to receive feedback, including grievances, from deployers or individual users of their technologies. This means that harmful system behavior may go unreported and potentially unaddressed.

- **Invalid or insufficient evaluation.** When downstream developers modify upstream foundation models, they change the nature and behavior of these models. The contexts for which downstream deployers have adapted models may not have been anticipated by the foundation model developer, and therefore not evaluated by the foundation model provider. If the downstream developer does not sufficiently adapt their own evaluations to their intended model context or expand the evaluations to encompass sufficient coverage, evaluations may fail to detect or adequately reflect the full risk landscape.

4.5 Deployers

The core distinguishing element of deployers is that they deploy genAI tools with a direct interface to users. The deployment business model (B2C vs. B2B) introduces nuances to this value chain category:

- In the B2C context, a deployer is whichever entity directly interfaces with consumer end users of genAI products. This includes companies like OpenAI, Google, and Anthropic.
- In the B2B context, deployers are often separate entities that 1) integrate genAI tools into internal systems or processes to increase operational efficiency, or 2) use externally-facing genAI tools to assist in the delivery of products or services to customers, clients, patients, constituents, etc. For example, a financial institution may deploy a genAI chatbot to assist employees in retrieving information relevant to their work (internal), or a healthcare organization may deploy a patient-facing chatbot to assist patients with understanding and interpreting their medical records (external).

Deployers have varying degrees of proximity to the downstream developers that develop the systems they deploy. In some cases deployers may belong to the same organization as downstream developers. For example, an IT team within a company may develop an LLM-powered tool for in-house deployment. In other cases, a company may deploy a genAI tool that it purchased externally, such as from an app marketplace or third-party developers.

Deployers engage in the following activities and operations:

- **Integration:** Making choices about how and where to integrate genAI systems into existing processes, workflows, and operations.
- **Capacity-building:** Providing training and guidance for users of genAI systems (e.g., employees) to understand genAI tools and use them effectively and responsibly and/or to operationalize a culture of safe genAI deployment.
- **Oversight of use:** Human oversight of the use of genAI systems to ensure they are functioning as intended and to protect against misuse or abuse.

It's important to note that entities can deploy genAI tools in different contexts. For example, the deployer of a general purpose, widely available, consumer-facing tool—such as OpenAI's ChatGPT—may reach millions of users across any application domain for innumerable use

cases. By contrast, some actors deploy genAI tools with clearly defined use cases, application domains, and user bases (e.g., a healthcare company using a genAI chatbot to respond to customer queries). The specificities of the deployment context impact the potential human rights impacts associated with a genAI tool's deployment.

Risk Pathways

Because deployers are closest to the use of genAI systems, they may be best positioned to identify risks related to their specific use case and/or issues that arise during deployment. However, deployers may not have the requisite information or resources to implement technical mitigations for observed issues in system performance. Technical mitigations at the model level are more appropriately addressed by foundation model developers and by downstream developers at the application level.

For this reason, the level of proximity a deployer has to the downstream developer may impact the effectiveness of ongoing risk mitigation. When there is less proximity between downstream developers and deployers, it may be difficult for deployers to alert downstream developers when issues arise in a specific deployment context, potentially including instances of misuse or abuse. Downstream and foundation model developers often do not have mechanisms to monitor how their tools are used, and such mechanisms may also be challenging to implement in practice. This can preclude their capacity to implement proactive mitigations. The dynamic nature of genAI risk may further complicate this issue because deployers may not have the resources or technical knowledge to understand and communicate rapidly evolving risks upstream.

Actions and omissions at the deployment level may create risks that can lead to downstream adverse human rights impacts. Examples of risk pathways that may occur at the deployment level include the following:

- **Limited genAI literacy and/or an absence of safety culture:** Deployers may fail to establish a deployment safety culture because they are unaware of the risks associated with genAI technologies. Limited awareness about how genAI works and its shortcomings may contribute to an inaccurate sense that the technology is infallible or may lead deployers to implement the technology in ways or into work streams that are inappropriate. Some deployers may also refuse to dedicate resources to safety and human rights assessment and mitigations due to a belief that upstream developers bear that responsibility. Just as with developers, deployers should have an AI governance system to ensure safe deployment of the technology.
- **Choices pertaining to deployment and integration:** Deployers make choices about how and where to integrate or deploy genAI tools. These choices may impact the likelihood of those tools being connected to downstream adverse human rights impacts. For example, deployers may choose to integrate genAI technologies into processes for which they are not fit-for-purpose or which limit the ability to implement human oversight. In one such instance, a tribunal found Air Canada liable for damages caused when the airline's chatbot provided inaccurate information to a customer about a discount.

Given that genAI models may convey inaccurate information through hallucinations, deployers should carefully consider how inaccuracies may impact end users when making decisions about how and when to deploy genAI tools. Deployment contexts may also lead to downstream harms; for instance, deploying powerful genAI models in conflict or high-risk affected areas may increase the likelihood that bad actors may utilize those tools for severe human rights violations.

Ineffective human oversight: Deployers may not ensure appropriate and effective human oversight of the end use of genAI tools. For instance, deployers may not establish usage monitoring workflows or choose not to use the data from those workflows to maximize safer product deployment. While all value chain actors should enact some form of human oversight, human oversight is especially salient for deployers because they are user-facing. This means that they can more nimbly act on information derived from human oversight to deploy risk mitigations, such as updating a classifier, banning an account, or temporarily shutting down a product.

In some cases, effective human oversight may be challenging to achieve due to a deployer's choices about how to integrate or deploy a genAI tool. Deploying genAI tools at scale creates intrinsic problems for human oversight because it is impossible for humans to review the output of millions of users. Scalable methods—which have their own challenges (see next bullet)—are necessary.

Additionally, when deployers launch externally facing genAI tools with model outputs that go directly to the user, human oversight may only be reactive rather than proactive. For example, if a public benefits office launches a genAI chatbot that is directly accessed by applicants to assist them in retrieving information about their public benefits cases, there is no opportunity for a caseworker to review the chatbot's outputs for accuracy before they reach the applicant. Inaccurate outputs may be associated with varying degrees of harm depending on the application domain.

- **Ineffective scalable oversight:** Deploying genAI tools at scale will often necessitate scalable oversight mechanisms. Core elements of these oversight mechanisms generally include a content and usage policy; trust and safety teams that help develop, maintain, and enforce those policies; and classifiers that detect and filter harmful generated content. These systems may negatively impact human rights for a variety of reasons, including poor policy design or classifiers that are over- or under-inclusive in content filtering and detection.
- **Limited awareness pertaining to developer safeguards:** Deployers may have limited knowledge about developer safeguards, including what they are, how they work, their limitations, and available mechanisms for information exchange and reporting failures to upstream actors. This may make it such that deployers are unaware of when safeguards are not performing as intended.

4.6 Individual Users

Individual users of genAI systems can include any consumer who interacts with genAI-powered products (such as chatbots, image generation software, or other apps built on foundation models), workers across domains who use genAI tools as a function of their jobs, or students or researchers who rely on genAI technologies in their research or studies.

Individual users of genAI systems are the final point in the value chain. As such, they have the least insight into systems design, evaluation processes, and risks, and the least leverage to mitigate potential harms that stem from upstream risk pathways.

Individual users may have varying degrees of proximity to the value chain categories upstream. For instance, users may have direct relationships with deployers if the deployed system is designed with direct input from users. This often occurs with domain-specific applications of genAI, such as medical note-taking AI apps that are designed with input from doctors and nurses. However, in many cases, individual users have distant or indirect relationships to deployers and developers in the form of communication channels such as grievance mechanisms, user communities, and stakeholder engagement workflows. The impact that these relationships have depends substantially on the degree of investment that the developer or deployer places in these communication channels. Sufficient investment can be challenging given the scale of the user base of most online applications.

This issue takes on a unique spin for open source models, which in effect permit users to become downstream developers since they may freely access and modify the model for their own use. While this has benefits (see the open source bullet in this section), it also opens the door to users removing safeguards and then proliferating those riskier models. One example of this is when bad actors modified open source image-generation models, such as Stable Diffusion, to remove safeguards against CSAM-generators and "undressing" apps.

Individual users engage in the following activities and operations:

- **Prompt creation:** This is the process of drafting prompts to models across modalities such as text, image, audio, and video to elicit desired outputs from a genAI system. Choices of prompt influence outputs, since users may explicitly ask for harmful outputs that may be connected with adverse human rights impacts.
- **Use case-related operations:** To engage with specific genAI systems, the user will participate in relevant system workflows. This may include preparing documents to be processed by a genAI system or acting on the system's outputs (e.g., an employee-facing chatbot surfaces information that the employee then uses to execute a job function).

Risk Pathways

Decisions by individual users of a genAI system may create risks that can lead to adverse human rights impacts. Examples of risk pathways that may occur at the level of individual use include the following:

- **Intentional manipulation, misuse, or abuse:** Individual users may intentionally manipulate, misuse, or abuse genAI systems by overriding their programmed safeguards to get the system to produce disallowed outputs such as incitements to violence or dangerous instructions.
- **Unintentional misuse or harm:** While some impacts will stem from abuse, many impacts will arise from a user using a system without the intent to cause harm or beyond the developer's intended uses. Some of these may be unintentional harm to others, while some will be harm to the user. One example of the latter is users who commit suicide following conversations with companion chatbots.
- **Limited genAI literacy:** Individual users may be unaware of the limitations of genAI systems. This may lead to overreliance, which may in part be driven by the confident tone that LLMs strike when they hallucinate, automation bias (see below), or otherwise using the technology in ways that are not appropriate for its capabilities or which may potentially adversely impact the user or others.
- **Overreliance:** Individual users may come to over-rely on genAI systems. This may be harmful if genAI systems' limitations may not be fit-for-purpose, such as detecting plagiarism, or where the genAI system is effective but overreliance may harm to the user, such as when students wholly rely on genAI for all their homework.
- **Automation bias:** Individual users may be more likely to believe information, decisions, or assessments presented by a technology product regardless of accuracy.

5. Human Rights Risks and Opportunities

GenAI systems are being deployed across sectors, and the associated human rights impacts—both risks and opportunities—are impacted by the nature of the sector and the context in which a genAI tool is deployed. The impact is especially high when genAI systems are deployed in sectors that provide essential services or which have significant impact on individuals' human rights. These include, but are not limited to education, social security, public benefits (e.g., welfare agencies), the criminal legal system, immigration, military and intelligence services, finance, health care, legal services, and critical infrastructure, such as the supply of water and energy. This is especially the case when those systems intersect with the core operations of the sector, such as when their outputs are directly used in decision-making processes.

However, there are some rights that are more likely to be impacted regardless of the use case or deployment context due to features that are common in genAI systems. For instance, the creation of convincing text, audio, image, or video is core to genAI. If this capability lowers the barriers to creating disinformation, that may impact access to information by complicating the information environment, thereby making it harder for people to have confidence in the information they are accessing. Additionally, genAI's ability to create content on the basis of natural language prompts may make genAI tools susceptible to enforcing user biases. Prompts that may imply discrimination on the basis of sensitive personal characteristics—such as gender, race, or sexuality (e.g., "create an image of your typical oversensitive woman")—may elicit discriminatory outputs. There may be many more examples.

The tables below highlight **six** categories of human rights impacts that BSR identifies as especially salient in the context of genAI. They are based on our experience conducting human rights assessments of a variety of genAI products, services, and associated business operations. We note that many or all of the listed impacts have already been identified by risk taxonomies published by other entities, such as the United Nations, AI companies, and academic institutions. The contribution of this section is that these risks are phrased in human rights terms, illustrating the flexibility and breadth of the human rights-based approach to identifying and mitigating genAI impacts. Additionally, the "risk pathways" component of each table connects these impacts to the decisions of each value chain actor, drawing from the risk pathways identified for each actor in the previous section.

However, because of the great number of possible use cases for genAI, there could be impacts that extend beyond these six identified areas of human rights. Therefore, genAI value chain entities should assess specific contexts (use case, application domain, deployment contexts, etc.) to identify which other rights may be relevant. In addition, since human rights are interdependent and interrelated, the impacts that are identified below could have knock-on impacts on other human rights. These are identified where relevant. It is also important to remember that the space is highly dynamic, meaning that ongoing assessment is needed to identify evolving harms.

The tables below illustrate how actors across the value chain may be connected to salient human rights risks associated with genAI through their actions of omissions. While both risks and opportunities are covered, the tables analyze only the connection between value chain actors and risks. This is in alignment with the UNGPs, which focus on human rights risks and do not require consideration of opportunities.[10] Note also that the actions or omissions in the examples provided are not necessarily failures or mistakes; they could be necessary for the creation of a certain product or service, but nonetheless can be connected with downstream human rights impact.

The below risks and opportunities will evolve along with the evolution of the technology and the development of new use cases. This list should therefore be considered as illustrative of the most salient risks and opportunities associated with genAI at the time of writing.

1. **Equality and Nondiscrimination**—the right to be free from differential treatment on the basis of attributes such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.
2. **Access to Information**—the right to have general access to information of public interest from a variety of sources.
3. **Privacy**—the right to control information about oneself and who may access it.
4. **Economic, Social, and Cultural Rights**—a range of rights on socioeconomic issues, including the right to work, to an adequate standard of living, and to property.
5. **Bodily Integrity**—the right to freedom from physical nonconsensual acts.
6. **Freedom of Thought and Opinion**—the right to hold or change all forms of opinions and thoughts without interference.

10 See the commentary to Principle 11 of the UNGPs: "Business enterprises may undertake other commitments or activities to support and promote human rights, which may contribute to the enjoyment of rights. But this does not offset a failure to respect human rights throughout their operations." For more reading on this topic, please see: https://www.bsr.org/en/reports/the-shared-opportunity-to-promote-a-second-decade-priority-for-the-ungps.

5.1 Equality and Nondiscrimination

The right to be free from differential treatment on the basis of attributes such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Relevant Human Rights Instruments:

- UDHR 1, 2, & 7
- ICCPR 2, 3, & 26
- ICESCR 2, & 3
- CEDAW 2
- CERD 2, 4, & 5
- CRPD 5
- DRIP 17
- ILO C111

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Outputs that reinforce stereotypes or encourage discrimination** GenAI model outputs may contain toxic content that reinforces stereotypes based on attributes such as gender, ethnicity, or nationality.	**Example**: A publicly available text-to-music tool produces song lyrics that contain hate speech and incitement to violence against women. • **Example connection to suppliers:** A dataset vendor does not adequately clean the dataset of misogynistic language. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** The developer utilizes a model evaluation technique that is focused on Western contexts, failing to detect toxicity in non-Western contexts or languages. **RISK PATHWAYS:** Insufficient or inappropriate safeguards, inappropriate model evaluation choices. • **Example connection to downstream developers:** A downstream developer is unaware of harmful content trends because it did not introduce feedback channels. The developer does not update output filters for recent misogynistic terminology. **RISK PATHWAYS:** Ineffective technical mitigations, lack of feedback channels. • **Example connection to deployers:** A deployer fails to evaluate the model for their specific deployment context. **RISK PATHWAYS:** Limited genAI literacy and an absence of safety culture. • **Example connection to individual user:** A user deliberately attempts to produce misogynistic lyrics that incite violence against women through prompts designed to jailbreak safety protections.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Over-, under-, mis-representation** There may be over-, under-, or mis-representation of certain groups in the outputs of genAI models. These outputs may reinforce and exacerbate stereotypes against groups that are under-represented or misrepresented.	**Example:** An image-generation tool prompted to produce pictures of doctors and CEOs disproportionately produces images of white men. • **Example connection to suppliers:** A dataset labeling vendor does not ensure demographic balance in its dataset, such that many of the male CEOs and doctors in the dataset are white. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** The developer does not release transparency notes that warn developers of the risk of racial bias in model outputs. **RISK PATHWAYS:** Insufficient transparency. • **Example connection to downstream developers:** The developer that creates the tool does not fine-tune the model to reduce racial bias that may be present in the dataset. **RISK PATHWAYS:** Ineffective technical mitigations, model fine-tuning choices. • **Example connection to deployers:** An advertising agency purchases the tool to create ad campaigns. However, it does not provide bias training to employees, who use the tool to create campaigns containing discriminatory images. **RISK PATHWAYS:** Limited genAI literacy and an absence of safety culture. • **Connection to individual users:** A creative lead at the ad agency uses the tool, producing a public campaign that reflects biases against ethnic minorities.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Process discrimination** GenAI may facilitate or exacerbate discrimination in decision-making processes that impact human rights, such as in employment, housing, applications for financial products, or education.	**Example:** An LLM chatbot used for resume screening favors applicants from certain ethnicities or genders over other similarly qualified candidates (example). • **Example connection to suppliers:** A supplier of CVs for datasets provides more qualified male resumes than female resumes, leading the LLM to associate accomplishment with maleness. This disparity in the data is not reflected in accompanying documentation. **RISK PATHWAYS:** Poor dataset curation, insufficient or misleading dataset documentation. • **Example connection to foundation model developers:** The developer does not subject its model to debiasing techniques, causing its outputs to favor names that are associated with maleness. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** A downstream developer uses the model to build a resume screening tool. The developer does not evaluate and mitigate for bias and sells it to the public sector. **RISK PATHWAYS:** Model fine-tuning choices, human rights-impacting use cases or application domains. • **Example connection to deployers:** The developer included a "manual mode," where operators may manually screen resumes stripped of identifying information such as names or countries of origin, and the genAI tool does not score the resumes. However, deployers fail to incorporate user education during rollout, causing this mode to be underused. **RISK PATHWAYS:** Limited genAI literacy and an absence of safety culture, limited awareness pertaining to developer safeguards. • **Example connection to individual users:** Operators are overconfident about the app's inferences, leading many to accept its resume-screening results without further consideration. This results in the tool disproportionately screening out female candidates.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Inaccessibility** GenAI models may perform less effectively or be less accessible for certain groups because, for example, they speak low-re-sourced languages or live in regions that are under-represented in the training data.	**Example:** Speakers of Palestinian Arabic experience an especially high rate of hallucinations in responses from a globally released general purpose genAI chatbot. When these chatbots are deployed to provide information on governmental services, they perform poorly for speakers of Palestinian Arabic, obstructing their access to public services. • **Example connection to suppliers:** Dataset curation companies include some Standard Arabic text, but under-index on Palestinian Arabic when putting together training datasets. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** The foundation model developer evaluates and fine-tunes for performance across English, European languages, and Standard Arabic, but neglects to do so for Palestinian Arabic. **RISK PATHWAYS:** Inappropriate model evaluation choices, insufficient or inappropriate safeguards. • **Example connection to downstream developers:** The developer of the chatbot advertises it as being "intelligent in any language you can imagine," misleading potential buyers as to its linguistic capabilities. **RISK PATHWAYS:** Insufficient transparency, human rights-impacting use cases or application domains. • **Example connection to deployers:** Public entities in Palestine deploy the chatbot in sensitive domains, such as advising on government benefits application procedures and eligibility, without testing for accuracy in Palestinian Arabic. **RISK PATHWAYS:** Choices pertaining to integration, limited genAI literacy, and an absence of safety culture. • **Example connection to individual users:** Users query the chatbot in Palestinian Arabic, asking for instructions on how to access health insurance or government benefits. The app regularly hallucinates, confusing users on application procedures and delaying their access to these services.

Impacts on Vulnerable Groups:

- Groups that are the likeliest targets of discrimination in society, such as women, children, older adults, or ethnic minorities, may face disproportionate impacts from genAI outputs that are discriminatory or feature under- or mis-representation.
- The concentration of genAI infrastructure and developers in the US and China may deepen the global digital divide with other countries.

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Educating against bias** GenAI outputs that systematically educate users about diversity, equity, and inclusion may create more advocates for equality and shift societal attitudes.	**Example**: GenAI assists users to understand the multifaceted impacts of systematic bias and how they may manifest in society in nonobvious ways.
OPPORTUNITY: **Equitable and representative datasets** Synthetic data may be used to create datasets that are more representative or equitable.	**Example**: LLMs are used to create synthetic datasets that are more representative of user populations. These datasets become industry benchmarks that are utilized by AI companies to create equitable and bias-reducing products or features, such as more diverse skin-tone scales.

5.2 Access to Information

The right to have general access to information of public interest from a variety of sources.

Relevant Human Rights Instruments:

- UDHR 19
- ICCPR 19 & 20
- CRC 13 & 17
- CRMW 33, 37, & 38

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Hallucination** GenAI models may "hallucinate"[11] by producing false or misleading outputs.	**Example**: A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures. • **Example connection to suppliers:** Suppliers provide a training dataset that is missing key information about taxpayer dates and procedures. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** Developers make errors during model training that result in lower levels of general accuracy, regardless of domain. **RISK PATHWAYS:** Limitations in model training techniques. • **Example connection to downstream developers:** Developers do not conduct adequate fine-tuning for accuracy in tax due to commercial pressures to release the app. **RISK PATHWAYS:** Choices pertaining to model fine-tuning, human rights-impacting use cases or application domains. • **Example connection to deployers:** Public bodies integrate the chatbot into their tax filing tools without warnings to users about the need to fact-check outputs. **RISK PATHWAYS:** Choices pertaining to integration, limited genAI literacy and/or an absence of safety culture. • **Example connection to individual users:** Users utilize the chatbot's guidance on tax filing without fact-checking, resulting in erroneous filings.

11 Hallucination is the tendency for genAI models to output factual errors. Hallucination may be *unfaithful* or *unfactual*. *Unfaithful* hallucination occurs when an LLM is asked to summarize an existing corpus of text and does not represent it accurately in its outputs. *Unfactual* hallucination occurs when an LLM is asked a question and makes up information. For more on this, please see: https://dl.acm.org/doi/pdf/10.1145/3703155.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Disinformation** GenAI may facilitate the creation of increasingly convincing synthetic disinformation.	**Example:** During an election, bad actors use genAI tools to flood the information environment with synthetic disinformation, making it difficult for voters to understand what is true or false. • **Example connection to suppliers:** Data annotation workers do not label data accurately, reducing the effectiveness of safety fine-tuning that would prevent model generation of content of this type. **RISK PATHWAYS:** Poor data labeling guidance. • **Example connection to foundation model developers:** A global tech company releases an open source foundation model, permitting anyone—including bad actors—to access its source code and model weights. **RISK PATHWAYS:** Choices pertaining to model release. • **Example connection to downstream developers:** App stores are flooded with dual purpose content-generation apps that are lacking safety guardrails and are adept at creating disinformation narratives. **RISK PATHWAYS:** Ineffective technical mitigations, human rights-impacting use cases or application domains. • **Example connection to deployers:** Bad actors aligned with foreign interests deploy the generative apps to produce and distribute convincing audio deepfakes of opposition leaders. **RISK PATHWAYS:** Choices pertaining to deployment and integration. • **Individual users:** Many people believe the deepfakes to be real, reducing voter turnout.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Flooding of synthetic content** GenAI tools lower the barriers to creating and distributing synthetic content. This pollutes the information environment, making it difficult for people to find accurate information.	**Example:** A direct-to-consumer image-generation app contains a feature that allows users to post images directly to their social media. Synthetic, photorealistic images flood the internet at scale, making it difficult for users to trust what they see online. • **Example connection to suppliers:** Suppliers implement a system that aims to filter out synthetic images from training datasets, but it does not catch new and evolving types of synthetic content. **RISK PATHWAYS:** Data curation. • **Example connection to foundation model developers:** Foundation model developers do not adequately fine-tune the model to refuse excessive requests to produce photorealistic images. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** The developer of the image-generation app creates a one-click button that permits users to post generated images directly to the social media accounts of their choice. **RISK PATHWAYS:** Human rights-impacting use cases or application domains. • **Example connection to deployers:** N/A (developer is also deployer in direct-to-consumer products) **RISK PATHWAYS:** N/A • **Example connection to individual users:** Users utilize the app's feature to create and widely distribute misleading images about political or historical topics, making it hard for people to distinguish between accurate and misleading information online.

Impacts on Vulnerable Groups:

- People with lower levels of technological literacy—such as children, older adults, or communities affected by digital divide—may be less able to identify and fact-check genAI hallucinations or synthetic disinformation.
- Access to poor quality information may have disproportionate impacts on right-sholders who live in countries with a weak information environment, such as countries lacking a free press or other reliable sources of information.

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Rapid and accurate access to information** Users may utilize genAI tools to sift through large quantities of data and rapidly receive accurate, complete, and understandable information.	**Example**: Journalists utilize a genAI chatbot to search through a large government database and obtain a summary of key information they need. **Example:** Users utilize genAI-powered search engines to access complex information, including advanced medical knowledge, educational methods, and explanations for advanced technical topics.

5.3 Privacy

The right to control information about oneself and who may access it.

Relevant Human Rights Instruments:

- UDHR 12
- ICCPR 17
- General Comment No. 16 on Article 17 ICCPR
- CRC 16
- CRPD 22
- ICRMW 14

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Personal data leaked in model outputs** Model outputs may contain sensitive personal data for various reasons, including dataset leakage or the user inputs containing sensitive information.	**Example**: A GenAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs. • **Example connection to suppliers:** Data suppliers supply datasets that contain sensitive personal information. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** Developers do not apply privacy techniques during model training, such as differential privacy, which would ensure that no personal information from the dataset becomes encoded in the model. **RISK PATHWAYS:** Insufficient or inappropriate safeguards, limitations in model training techniques. • **Example connection to downstream developers:** Developers do not create a product policy against queries that might leak personal information, such as "tell me about Jane Doe" or "tell me where Jane Doe lives." **RISK PATHWAYS:** Ineffective technical mitigations. • **Example connection to deployers:** (developer is also deployer in direct-to-consumer products) **RISK PATHWAYS:** N/A • **Example connection to individual users:** Users submit repeated queries to the chatbot to obtain private information about people they know.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Personal data revealed through invasive analytics** GenAI models may be used to analyze data to produce privacy-violative insights about individuals.	**Example:** A genAI-powered data analytics tool is sold to a public body in an authoritarian country, which uses it to review CCTV footage to identify outlawed minority religious practices. • **Example connection to suppliers:** A supplier does not de-identify a dataset of images it provides to downstream actors, allowing for individuals within the dataset to be identified. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** The foundation model developer does not conduct due diligence before granting API access to the downstream developer, which has a history of sales to authoritarian governments. **RISK PATHWAYS:** Insufficient or inappropriate safeguards, choices pertaining to model release. • **Example connection to downstream developers:** The developers use the model to build the surveillance-oriented analytics tool. They then sell the tool to a public security arm of an authoritarian state. **RISK PATHWAYS:** Human rights-impacting use cases or application domains. • **Example Connection to Deployers:** The public security arm procures the analytics tool for the purpose of oppressive surveillance of minorities and/or perceived enemies of the state. **RISK PATHWAYS:** Human rights-impacting use cases or application domains, limited genAI literacy, and an absence of safety culture. • **Example connection to individual users:** Employees at the public security entity use the application to conduct indiscriminate surveillance and save the data to their personal devices.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Generation of nonconsensual explicit imagery** GenAI models are used to alter images or videos to create synthetic nonconsensual explicit imagery.	**Example:** An image-generation tool is utilized to create images that combine the faces of real children with nude bodies. • **Example connection to suppliers:** Suppliers do not clean the data of child sexual abuse material or use datasets that combine image data of children with other data of sexualized behavior, creating the possibility of models making associations between both types of data. **RISK PATHWAYS:** Poor dataset curation. • **Example connection to foundation model developers:** Foundation model developers undertake safety fine-tuning but do not red team for determined attempts to produce AI-generated child sexual abuse material (CSAM). **RISK PATHWAYS:** Insufficient or inappropriate safeguards, inappropriate model evaluation choices. • **Example connection to downstream developers:** Developers of the image-generation tool do not include prompt filters for requests that may produce CSAM. **RISK PATHWAYS:** Ineffective technical mitigations. • **Example connection to deployers:** The deployer of the image-generation tool does not include a reporting channel for generated CSAM outputs. **RISK PATHWAYS:** N/A • **Example connection to individual users:** Bad actors utilize the app to generate synthetic CSAM that depicts real people at scale for distribution.

Impacts on Vulnerable Groups

- People with lower levels of technological literacy—such as children, older adults, or communities affected by digital divide—may be less capable of identifying and securing remedy for privacy breaches, such as the inclusion of their personal data in a dataset or AI-generated CSAM.
- GenAI-enabled privacy breaches may have more significant consequences for over-policed or surveilled groups. For instance, human rights defenders and other perceived dissidents who live in authoritarian states may experience especially severe consequences connected to privacy infringement.

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **GenAI-powered privacy tools** GenAI powers product features or tools that empower users to make informed choices about their privacy.	**Example:** Users utilize retrieval-augmented generation (RAG), a form of genAI that focuses on information retrieval and summarization, to quickly understand lengthy privacy policies. This enhances user knowledge of the options they may take to exercise privacy when utilizing digital services. **Example:** A genAI-powered privacy settings dashboard dynamically suggests optimal privacy settings for an enterprise, enabling users to dynamically observe and adjust data permissions in one convenient location.

5.4 Economic, Social, and Cultural Rights

A range of rights on socioeconomic and cultural issues, including the right to work, to an adequate standard of living, to science, and to property

Relevant Human Rights Instruments:

Right to desirable work, dignified conditions, and union
- UDHR 23
- ICESCR 6, 7, & 8
- CEDAW 11

Right to rest and leisure
- UDHR 24
- ICESCR 7

Right to health and an adequate standard of living
- UDHR 25
- ICESCR 12
- CRC 24
- CEDAW 12

Right to culture and science
- UDHR 15 & 27
- ICESCR 15

Right to education
- UDHR 26
- ICESCR 13 & 14
- CRC 28 & 29
- CEDAW 10

Right to personal property
- UDHR 17

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Decrease in employment opportunities** The use of genAI by workplaces may be associated with a decrease in employment opportunities for a range of professions across all income levels, or may affect the profitability of certain types of work, leading to the loss of economic opportunity and impacts on standard of living for workers.	**Example:** An insurance company integrates a genAI customer support chatbot to their website to assist customers with purchasing and renewing policies, filing claims, and other processes. Due to the introduction of the chatbot, the insurance company hires fewer call center workers. • **Example connection to suppliers:** N/A **RISK PATHWAYS:** N/A • **Example connection to foundation model developers:** Consistent improvements to natural language processing in LLMs enables the design of tools that effectively replace humans in customer service applications. **RISK PATHWAYS:** Choices pertaining to model release. • **Example connection to downstream developers:** Downstream developers develop apps for commercial use which perform tasks previously performed by humans, enabling the replacement of human labor in the workplace. **RISK PATHWAYS:** Human rights-impacting use cases or application domains. • **Example connection to deployers:** Deployers are likely to be most closely connected to this risk. Employers across sectors integrate genAI tools into workflows in ways that

either decrease the need for human labor or replace human labor entirely.

RISK PATHWAYS: Choices pertaining to integration.

- **Example connection to individual users:** N/A—individual users are the ones being replaced by the genAI tool, and their actions have no connection to the human rights impact.

  RISK PATHWAYS: N/A

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Infringement on intellectual property rights** GenAI models may be trained on the original works of others and are used without compensation to the owners or authors. Furthermore, genAI models may directly reproduce the original works of others in ways that negatively impact their intellectual property rights.	**Example:** A genAI image-generation program trained on the original works of a visual artist produces outputs in the style of that artist or directly reproduces the artist's work. Although this may or may not violate domestic copyright laws, it impacts the artist's intellectual property rights and potential economic opportunity. • **Example connection to suppliers:** Suppliers deliberately scrape or fail to screen out copyrighted content, intellectual property, or other forms of original and protected work from training datasets. **RISK PATHWAYS:** Poor dataset curation, poor data labeling guidance. • **Example connection to foundation model & downstream developers:** Both foundation model and downstream developers may have a similar connection to this risk. Developers neglect to instruct data suppliers to ensure that datasets are filtered for content that may be copyrighted, intellectual property, or otherwise protected and may not review training datasets to screen for such content prior to using them for model training purposes. Furthermore, developers may neglect to integrate safeguards to prohibit genAI tools from producing outputs in the likeness of original works. **RISK PATHWAYS:** Insufficient or inappropriate safeguards, ineffective technical mitigations. • **Example connection to deployers:** Deployers, intentionally or unintentionally, use genAI tools to produce outputs that mimic the likeness of original works for purposes related to advertising, product design, or other operations. **RISK PATHWAYS:** Lack of effective human oversight.

- **Example connection to individual users:** Individual users, intentionally or unintentionally, use genAI tools to produce outputs that mimic the likeness of original works.

  RISK PATHWAYS: Intentional manipulation, misuse, or abuse.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Facilitation of cybersecurity attacks** GenAI models may be used to produce content (including code) for cybersecurity attacks, including scams, fraud, spam, phishing, and data breaches, resulting in loss or damage of property.	**Example:** Bad actors disguise malicious inputs inside prompts ("prompt injection attack") to bypass the model's safeguards and trick it to produce harmful outputs, such as leaking personal data from the dataset. • **Example connection to suppliers:** N/A **RISK PATHWAYS:** N/A • **Example connection to foundation model developers:** The foundation model developers' red-teaming exercises do not catch new and evolving forms of prompt injection. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** Downstream developers may neglect to integrate appropriate technical limitations to prevent their products from being used in association with scams, fraud, spam, phishing attempts, or data breaches. Affected stakeholders have no way of flagging the misuse to the downstream developers. **RISK PATHWAYS:** Ineffective technical mitigations, lack of feedback channels for deployers and individual users. • **Example connection to deployers:** Deployers are bad actors, such as organizations that engage in illegal activity like prompt injection attacks to obtain sensitive user data for the purposes of sale. **RISK PATHWAYS:** Choices pertaining to deployment and integration. • **Example connection to individual users:** Individual users are bad actors, such as individuals who engage in illegal activity like using prompt injection attacks to obtain sensitive user data for the purposes of sale. **RISK PATHWAYS:** Intentional manipulation, misuse, or abuse.

Impacts on Vulnerable Groups:

- Workers from diverse sectors may be adversely impacted by widespread enterprise integration of genAI.
- Artists, writers, and other content creators, particularly those who are small-scale and/or are not protected by or do not have the resources to enforce copyright laws, are more likely to be adversely impacted by genAI's ability to mimic the likeness of the original works in its training data.
- People with lower levels of technological literacy—such as children, older adults, or communities affected by digital divide—are more likely to be victims of genAI-enabled scams.

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Improved workplace conditions** GenAI may be integrated into workflows in ways that improve conditions for workers and/or their work experience.	**Example:** An employee-facing, organization-specific genAI chatbot is integrated into internal systems to help employees quickly and efficiently retrieve information necessary to perform their job functions. This creates net improvements to employees' work conditions, reducing job-related stress.
OPPORTUNITY: **Expanded access and improved quality of education** GenAI may increase access to education or improve quality of education through general consumer products or via integration into education systems.	**Example:** GenAI tools optimized for high-quality performance in subjects—such as mathematics, history, literature, the sciences, or others—provide new, effective, and accessible educational resources for students, improving the learning experience and expanding access to education.

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Expanded access and improved quality of healthcare** GenAI may be applied to use cases that expand access to and quality of healthcare services.	**Example:** GenAI chatbots are integrated into patient-facing platforms to assist with interpreting test results, scheduling appointments, or other operations.
OPPORTUNITY: **New means of cultural expression** GenAI may provide new tools for cultural expression.	**Example:** GenAI tools are leveraged by artists, writers, and others to facilitate the creation of art, media, literature, poetry, or other creative cultural content for public enjoyment.
OPPORTUNITY: **Facilitation of scientific innovation** GenAI may accelerate or facilitate processes related to scientific innovation.	**Example:** GenAI tools are integrated into various research and development processes, such as data analytics and visualization. This may accelerate scientific innovation or enable types of innovation that were previously not possible, facilitating the right to benefit from scientific advancement.

5.5 Bodily Integrity

A range of rights related to freedom from physical, nonconsensual acts

Relevant Human Rights Instruments:

Right to life, liberty, and security of person
- UDHR 3
- ICCPR 3, 6, & 9
- CRC 2
- ILO Occupational Safety and Health Convention (No.155)

Freedom from arbitrary arrest and detention
- UDHR 9
- ICCPR 3, 6, & 9
- CRC 2

Freedom from torture, cruel, inhumane, and degrading treatment
- UDHR 5
- ICCPR 7
- UNCAT

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Incitement of societal violence** GenAI model outputs may be created and used to incite violence, harassment, or other bodily harm, either intentionally or unintentionally.	**Example:** Bad actors leverage publicly available genAI products to produce realistic incendiary content at scale to sew division along religious lines, leading to violent social unrest. • **Example connection to suppliers:** Datasets are not filtered to remove hate speech or violent, graphic, or otherwise harmful content. **RISK PATHWAYS:** Poor dataset curation, poor data labeling guidance. • **Example connection to foundation model developers:** Foundation model developers do not provide guidelines for data suppliers, which require datasets used for fine-tuning to be cleaned of violent, graphic, hate speech, or other harmful content; and neglect to implement technical and/or policy safeguards to prevent outputs containing such content. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** Downstream developers neglect to implement technical safeguards to prevent their products from being used to create content that is violent or graphic or which contains hate speech or other potentially harmful content that is relevant to the deployment context. Affected stakeholders have no way of flagging the misuse to the downstream developers. **RISK PATHWAYS:** Ineffective technical mitigations, lack of feedback channels for deployers and individual users.

	• **Example connection to deployers:** Deployers—such as political campaigns or news outlets—do not adhere to ethical or journalistic standards when using genAI, or fail to implement violent content detection workflows. **RISK PATHWAYS:** Limited genAI literacy and/or an absence of safety culture, lack of effective human oversight. • **Example connection to individual user:** Individual bad actors intentionally use genAI to produce content meant to sew division and spread it via social media channels. **RISK PATHWAYS:** Intentional manipulation, misuse, or abuse.
Human Rights Risks	**Examples and Connection to Value Chain Actors**
RISK: **Facilitation of violence or harassment** GenAI systems may be used to facilitate violence or harassment toward individuals.	**Example:** Bad actors leverage genAI products to produce speech in the likeness of a rightsholder's loved one requesting to meet them in a location where they may be mugged or assaulted. • **Example connection to suppliers:** Suppliers do not filter datasets to remove personally identifiable information, such as voice data from real individuals. **RISK PATHWAYS:** Poor dataset curation, poor data labeling guidance. • **Example connection to foundation model developers:** Foundation model developers neglect to implement technical and/or policy safeguards, such as post-training the model to refuse to comply with requests to imitate the voice of specific persons. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** Downstream developers neglect to implement technical safeguards to prevent their products from being used to create content in the likeness of real individuals. Affected stakeholders have no way of flagging the misuse to the downstream developers. **RISK PATHWAYS:** Ineffective technical mitigations, choices pertaining to model fine-tuning, lack of feedback channels for deployers and individual users. • **Example connection to deployers:** N/A **RISK PATHWAYS:** N/A

- **Example connection to individual user:** Bad actors intentionally use genAI to produce content meant to deceive and harm.

  RISK PATHWAYS: Intentional manipulation, misuse, or abuse.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Connection to critical system failures** GenAI models integrated into high-risk systems or critical infrastructure may be connected to system failures that result in bodily harm or loss of life or security (e.g., unintended shutdown of energy infrastructure or discharge of weapons, etc.).	**Example**: A genAI interface integrated into a control panel for the delivery of essential utility services, such as electricity, misinterprets a user's prompt, resulting in unintended power shutoff. • **Example connection to suppliers:** N/A **RISK PATHWAYS:** N/A • **Example connection to foundation model developers:** Foundation model developers do not create acceptable use policies for the application of their models in high-risk or critical systems. **RISK PATHWAYS:** Insufficient or inappropriate safeguards. • **Example connection to downstream developers:** Downstream developers make choices to integrate genAI into systems in ways that are inappropriate for the technology's capabilities and limitations. Affected stakeholders have no way of flagging the misuse to the downstream developers. **RISK PATHWAYS:** Human rights-impacting use cases or application domains, lack of feedback channels for deployers and individual users. • **Example connection to deployers:** Deployers make inappropriate decisions about how to integrate genAI tools into systems and workflows and/or do not provide sufficient capacity building to employees responsible for operating genAI tools. **RISK PATHWAYS:** Limited genAI literacy and/or an absence of safety culture, choices pertaining to integration, lack of effective human oversight • **Example connection to individual users:** Individual users of the genAI tool, such as employees, do not review genAI outputs for relevance, accuracy, and safety. **RISK PATHWAYS:** Limited genAI literacy, overreliance, automation bias.

Impacts on Vulnerable Groups:

- People in regions with weak rule of law, armed conflict, or heightened social unrest may be more likely to be impacted by the spread of misinformation created using genAI.
- Members of vulnerable groups, such as racial, ethnic, or religious minorities; women and LGBTQ+ persons; as well as activists, journalists, and other human rights defenders, may be more likely to be the victims of bodily integrity harms facilitated or incited by genAI tools (e.g., the subject of misinformation leading to real-world harm against individuals).

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Rapid access to information on treating violations of bodily integrity** GenAI may provide users with access to information on how to maintain their bodily integrity against incursions such as wounds or other injuries.	**Example:** A user sustains a severe cut in a rural area. They obtain step-by-step instructions from their mobile genAI chatbot app on how to treat and dress the wound.
OPPORTUNITY: **Facilitation of activities that protect the bodily integrity of others** GenAI features may facilitate activities that may protect the bodily integrity of themselves and other rightsholders.	**Example:** A nonprofit utilizes genAI tools to help them design and execute a successful campaign against domestic violence.

5.6 Freedom of Thought and Opinion

The right to hold or change all forms of opinions and thoughts without interference

Relevant Human Rights Instruments:

- UDHR 18 & 19
- ICCPR 18 & 19
- CRC 14

Human Rights Opportunities	Examples and Connection to Value Chain Actors
OPPORTUNITY: **Increased opportunity to develop informed opinions** GenAI technologies may expose users to information that they have not previously been exposed to, expanding their ability to exercise their right to freely form thoughts and opinions.	**Example:** A user accesses a genAI chatbot which presents information that expands their worldview and enables them to form opinions about a number of topics.

Human Rights Risks	Examples and Connection to Value Chain Actors
RISK: **Decline in critical analysis skills** Users may become over-reliant on genAI systems, may be unaware of their limitations, and/or may overestimate the quality of information they present. This in turn could decrease users' independent critical analysis when engaging with genAI outputs in ways that impact their right to freely form thought and opinions.	**Example:** Some users rely on genAI chatbots as their primary source of news and information without regard to the tool's limitations, accuracy, or representativeness, adversely impacting their processes of freely forming opinions. • **Example connection to suppliers:** N/A **RISK PATHWAYS:** N/A • **Connection to foundation model developers:** Foundation model developers are not transparent about the limitations of the model and its appropriate uses (i.e., that it should not be integrated into products where it will be used as a comprehensive and authoritative source of general information). **RISK PATHWAYS:** Insufficient transparency, limitations in model training techniques. • **Connection to downstream developers:** Downstream developers do not take measures to ensure transparency about how information featured in the chatbot's outputs is sourced and the limitations in the chatbot's performance. **RISK PATHWAYS:** Insufficient transparency, limitations in model training techniques. • **Connection to individual user:** Users lack genAI literacy and become over-reliant on the chatbot in ways that shape their thoughts and opinions over time. **RISK PATHWAYS:** Limited genAI literacy, overreliance, automation bias.
RISK: **Hyper-personalized behavioral nudging** GenAI systems' advanced data analytics may enable increasingly targeted	**Example:** A user accessing a genAI virtual companion as part of their mental healthcare services becomes increasingly reliant on the assistant. The virtual companion draws from its memory of conversations with the user, including personal health data, to nudge the user to buy pharmaceutical products. • **Example connection to suppliers:** N/A **RISK PATHWAYS:** N/A

Human Rights Risks	Examples and Connection to Value Chain Actors
nudging of user behavior, such as through advertisements or suggested actions. Users that have become over-reliant on GenAI systems may have their thoughts substantially shaped by these systems; in extreme cases, this may impact their ability to freely form thoughts and opinions.	• **Connection to foundation model developers:** Foundation model developers do not integrate policies to prevent the use of foundation models in ways that may be related to anthropomorphization, manipulation, or persuasion. **RISK PATHWAYS:** Insufficient or inappropriate safeguards, limitations in model training techniques, insufficient transparency. • **Connection to downstream developers:** Downstream developers neglect to implement technical safeguards to prevent the genAI tool from exhibiting persuasive or manipulative behavior. Affected stakeholders have no way of flagging the misuse to the downstream developers. **RISK PATHWAYS:** Ineffective technical mitigations, human rights-impacting use cases or application domains, choices pertaining to model fine-tuning, lack of feedback channels for deployers and individual users. • **Connection to deployers:** Deployers—such as mental health clinics—do not perform human rights due diligence prior to the integration of new features, such as advertisements, into genAI tools, and risks introduced by upstream actors may go undetected. **RISK PATHWAYS:** Limited genAI literacy and/or an absence of safety culture, choices pertaining to integration, lack of effective human oversight, limited awareness pertaining to developer safeguards. • **Connection to individual users:** Users lack genAI literacy and develop a tendency to anthropomorphize and become over-reliant on genAI chatbots. **RISK PATHWAYS:** Limited genAI literacy, overreliance, automation bias.

Impacts on Vulnerable Groups:

- People with lower levels of technological literacy—such as children, older adults, or communities affected by digital divide—are most likely to be impacted by overreliance, problematic nudging, or receptiveness to synthetic misinformation in ways that impact freedom of opinions

5.7 Attribution and Remedy

5.7.1 Attribution

When conducting human rights assessments, BSR leverages the following definitions outlined in the UNGPs to determine an entity's level of connection to a human rights impact. The UNGPs also offer guidance on appropriate actions for the actor to take based on the extent of connection to a human rights impact.

- **Cause:** An entity is considered to "cause" a harm when its actions alone are sufficient to create harm. One example would be if a company designs and uses a biased genAI recruiting tool that discriminates against minority job applicants. When an entity causes a harm, the UNGPs state that it should take the necessary steps to cease or prevent the impact.
- **Contribute:** An entity is considered to "contribute" to a harm when it enabled or facilitated another party to create harm. One example would be if an app developer did not introduce safety filters on its GenAI chatbot, and users used the app to create CSAM. When a company contributes to a harm, the UNGPs state that it should take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining impact to the greatest extent possible.
- **Directly linked:** The entity's connection is less than causing or contributing, but the company is nevertheless linked in some way to the harm. GenAI value chain actors will generally be at least directly linked to harm, because the point of a value chain is that each actor's actions or contributions to genAI are essential to the final product or service. When a company is linked to a harm, it should determine action based on factors such as the extent of leverage over the entity concerned and the severity of the harm.
- **Not linked:** If the business has no link to the harm, it need not act. All businesses that do not fall into the "cause," "contribute," or "directly linked" categories fall into this category.

The UNGPs state that when companies are found to have "caused" or "contributed to" adverse impacts, they should provide for or cooperate in remediation processes. A company that is "directly linked" to harm by their business relationships is not required to provide or cooperate in remediation, though it may take a role in doing so.

In practice, the cause, contribute, or directly linked categories operate on a spectrum and often require contextual analysis. For companies in the genAI chain, a key factor in determining that position will often be **the extent to which they have enacted adequate human rights mitigations and safeguards and the extent to which risks/impacts are reasonably foreseeable.** Companies with adequate risk mitigation efforts are more likely to be "directly linked" to harm than to "contribute" to it, removing their subsequent obligation to provide or cooperate in remedy. Simply put, companies that have acted proactively to identify and mitigate risks will subsequently face less harms that they will have the duty to remediate.

Determining where actors in the genAI value chain fall on this spectrum is uniquely challenging because there are often multiple actors involved in the creation of an eventual genAI product that may be connected to adverse human rights impacts. For example, when genAI tools produce outputs that unintentionally perpetuate racial stereotypes, there are a variety of actions or omissions at different points of the value chain by suppliers, foundation model developers, and downstream developers that are connected to that impact, and it may be impossible to identify one single cause. Because of this, suppliers and foundation model developers are less likely to "cause" adverse human rights impacts on their own. An exception to this is the inclusion of intellectually protected content in datasets, which is an action by suppliers that causes impacts to property rights.

For the adverse human rights impacts associated with the use of genAI technologies, deployers or users are most likely to cause these impacts. Developers are more likely to cause an impact when they develop a tool that on its own is connected to harm. This would be the case for technologies designed for a purpose that is inherently violative of human rights. An example of this are genAI-powered apps designed to "undress" individuals, which even when used as intended violate the rights to privacy and human dignity.

Most attribution analyses will revolve around ascertaining whether value chain actors are "contributing to" or "directly linked" to the adverse human rights impacts of genAI. This is a consequential distinction because "contributing" actors must directly provide or cooperate in remedy, while "directly linked" actors only need to use their leverage to mitigate human rights harms. An actor is more likely to be considered "contributing to" adverse human rights impacts, rather than "directly linked" to those impacts, if its activities (including both actions or omissions):

- **Facilitate or enable** another entity to "cause" an adverse impact, where a company's actions add to the conditions that make it possible for use of a product by a third party to "cause" harm.
- **Incentivize or motivate** another entity to "cause" an adverse impact, where a company's actions make it more likely that a product or service will be used in ways that "cause" harm.

Therefore, a key question for actors at all points in the genAI value chain is whether their actions or omissions may facilitate, enable, incentivize, or motivate others to cause harm.

An example of this would be if a downstream developer created a general-purpose image-generation app, but did not implement safeguards to prohibit users from using the app to produce nude images of real individuals. In this case the downstream developer may be considered to be "contributing" to the adverse human rights impact due to the reasonable foreseeability of the risk and a failure to implement effective safeguards that would prevent the technology from being used to cause harm in this way.

In most cases where an actor may be considered to be "contributing" to harm, this connection may decrease to "directly linked" if human rights due diligence is conducted and mitigations are put in place to prevent the facilitation, enablement, incentivization, or motivation of others to cause harm using the data or technology.

For more information on attribution in the context of AI and digital services more broadly, please see the "Taking Action to Address Human Rights Risks Related to End Use" paper published by the B-Tech Project at the United Nations Office of High Commissioner for Human Rights. The B-Tech Project has many other useful resources on issues related to the intersection of human rights, AI, and other digital technologies.

5.7.2 Remedy

Remedy refers to the redressing of impacted rightsholders for harms suffered. Remedy for genAI's harms can take many forms, including apologies, restitution, rehabilitation, financial or nonfinancial compensation, or guarantees of non-repetition. These will often require action by multiple actors, including entities in the genAI value chain, nonprofit organizations, and public entities. This is known as the "remedy ecosystem."

Direct remediation by businesses is important, but it is not the only tool in the human rights-based remedy toolkit. Given the speed and scale of potential harm from genAI systems, many harms will be cumulative or societal in nature, such as the harms associated with disinformation or job loss. Effectively addressing these will necessitate an ecosystem approach that relies on cooperation between state and non-state actors, including courts, governmental bodies, and private/nonprofit actors.

Collaboration across the value chain is needed for effective risk mitigation and remedy. For instance, effectively preventing genAI chatbot from generating hate speech may require updates to the application's safety filters, new fine-tuning of the foundation model, and the cleaning of datasets of harmful content and associations. Value chain actors should establish effective means of communication with one another in order to collaborate on remedy and risk mitigation.

However, where direct remediation by businesses is required, and that remedy requires collaboration across the genAI value chain, the duty to coordinate that remedy should lie with a single point of contact. A single point of contact ensures that remedies are accessible and adequate for impacted rightsholders and prevents "finger-pointing" by value chain entities that may frustrate the effective resolution of grievances. This single point of contact should coordinate different elements of remedy across the value chain and communicate those actions to impacted stakeholders.

For more information about remedy in the genAI context, please see the Guide 8: Remedy for Generative AI-Related Harms.

6. Recommendations

There are many existing frameworks for understanding how to address potential harms that arise from genAI. This section does not aim to collate an exhaustive list of all existing research on genAI risk mitigation; rather, it provides an ecosystem approach for how value chain actors may influence or cooperate with each other to address human rights risks.

Because this HRA is grounded in the UNGPs and the international human rights instruments upon which the UNGPs are based, and because the UNGPs focus on the roles of governments and companies, this HRA does not include recommendations for individual users. However, individual users are part of the genAI value chain and should avoid misuse of genAI models and opt to increase their genAI literacy.

The following recommendations are measures that enterprise actors may take to address human rights risks related to genAI. Other value chain-adjacent actors, such as public policymakers and NGOs, also have key roles to play in managing the impacts of generative AI. However, these are beyond the scope of the current paper.

The first section below, Section 6.1 Ecosystem Recommendations, contains recommendations that apply to all actors in the genAI ecosystem. The sections that follow—Section 6.2 to 6.5—are recommendations targeted at specific value chain categories.

Summary of Recommendations

Cross-Ecosystem	• Human rights due diligence (HRDD) of genAI products or features should incorporate risk assessments of all value chain actors. • Increase transparency between value chain actors to improve risk mitigation efforts. • Communicate actions taken to address risks and changes to "safety" approaches across the value chain. • Collaborate with value chain and value chain-adjacent actors to establish safety standards. • Increase collaboration between value chain actors and affected stakeholders to improve risk mitigation efforts.

Suppliers	• Institutionalize workflows that ensure responsible and human rights-respecting data collection, curation, and provision. • Collaborate with developers to establish unified templates for data documentation, such as datasheets for datasets. • Collaborate with foundation model and downstream developers to create standards and best practices for data annotation / labeling guidelines.
Foundation Model Developers	• Create data procurement requirements to ensure training data is fit for purpose and supports privacy, equity, and safety; and evaluate datasets according to that criteria prior to procurement. • Create publicly available transparency documentation that accompanies the release of foundation models. • Create responsible use guidance containing recommendations and considerations for how downstream developers should responsibly build on foundation models. • Ensure that labor considerations for model training, labeling, or annotation activities are incorporated into human rights assessments • Make informed choices about model access and availability. • Work closely with downstream developers to take a collaborative approach to addressing risks that arise at the use case level. • Invest in research on the impacts of using synthetic data to train or fine-tune models both to model performance and the data ecosystem.
Downstream Developers	• Ensure that human rights due diligence considers the impacts of B2C or B2B business models, where applicable. • Engage deployers and individual users throughout the full development lifecycle of a genAI system or product. • Design products to have easy-to-use reporting mechanisms for deployers, end users, and other stakeholders to flag issues.
Deployers	• Proactively and regularly report issues to the downstream developer. • Invest in capacity-building to promote cross-organizational genAI literacy and safety culture. • Consider labor rights implications prior to integrating genAI-powered tools into workflows. • Engage downstream developers to inform the design of systems that support workers.

6.1 Cross-Ecosystem Recommendations

This section contains recommendations that are applicable to all actors in the genAI ecosystem.

› Human rights due diligence (HRDD) of genAI products or features should incorporate risk assessments of all value chain actors.

HRDD should be conducted on relevant upstream and downstream actors prior to the development and deployment of a genAI system and on an ongoing basis. This may include targeted analysis of a downstream customer's human rights risk profile, human rights conditions in the market of deployment, and intended or unintended use cases for the model by the customer or third parties who may gain access to it. This due diligence should go beyond sanctions list assessments that are standard in "know your customer" and consider the full spectrum of human rights risks (see for more information on downstream HRDD). Practitioners may also use the value chain mapping in this paper as a guide. See Practitioner's Guide 3: A Human Rights-Based Approach to Impact Assessment and Practitioner's Guide 4: A Human Rights-Based Approach to Risk Mitigation for applied guidance on this topic.

› Increase transparency between value chain actors to improve risk mitigation efforts.

The development and deployment of safe genAI systems must be a shared effort. Increasing transparency and communication between value chain actors will enable more effective risk mitigation efforts. For data suppliers this could entail providing datasheets, or similar documentation, for training datasets. For foundation model and downstream developers this may involve releasing system cards, model cards, and factsheets; producing regular reports for system evaluations and/or risk assessment processes; or enabling pre- and post-deployment incident reporting mechanisms for developers, deployers, and end users/operators. These should go beyond capabilities / performance reporting to include easy-to-digest safety disclosures. These transparency mechanisms should also be carefully designed to avoid enabling bad actors or overwhelming other value chain actors with too much information. Specific transparency recommendations for value chain actors are included in the following sections. See Practitioner's Guide 7: Aligning Transparency and Disclosure Practices with Human Rights Responsibilities for applied guidance on this topic.

› Communicate actions taken to address risks and changes to "safety" approaches across the value chain.

Model safety evaluations should happen on an ongoing basis both proactively and if/when issues surface during deployment. As safety issues are identified and addressed, these changes should be communicated to upstream and downstream value chain actors to enable holistic improvements to genAI safety throughout the entire ecosystem. Notification of safety actions could happen through existing reporting channels between value chain actors, specialized incident reporting channels, company announcements, and/or in-app notifications. These communication flows are likely to be iteratively improved over time.

› Collaborate with value chain and value chain-adjacent actors to establish safety standards.

Value chain actors should seek to work closely with researchers, academics, policy-makers, and domain-specific experts to establish industrywide, and domain-specific safety standards for the development and deployment of genAI systems. Standardized processes can ensure a uniform approach to ethical, responsible, and safe genAI development and deployment. Standardized safety reporting processes and formats may also increase transparency and comparability. Minimum thresholds for model safety should be determined by the application domain and individual use case with more conservative requirements for higher-stakes domains, such as public sector applications that may impact the realization of human rights.

› Increase collaboration between value chain actors and affected stakeholders to improve risk mitigation efforts.

Stakeholder engagement is a requirement under the UNGPs, and collaborative efforts—such as participatory design approaches that center affected communities in processes of problem formulation, data collection and storage, or model evaluation design—can increase meaningful involvement of relevant stakeholders. Additionally, stakeholder convenings, such as open- or closed-door industry sessions, should be held for the candid discussion of risk and mitigation for a genAI system or use case. Such sessions could invite data vendors, the foundation model developer, downstream developers, prospective sector-specific deployers and experts to collaborate on the safe and equitable design of the system. (For example, for a healthcare specific genAI application, the design process should engage healthcare organization deployers and experts in public health and healthcare inequality). For applied guidance on this topic, see Practitioner's Guide 5: Conducting Stakeholder Engagement.

6.2 Recommendations for Suppliers

› Institutionalize workflows that ensure responsible and human rights-respecting data collection, curation, and provision.

Supplied datasets are used for model training and development workflows, which means that they can have cascading human rights impacts downstream. This makes it especially important for developer practices to respect the full spectrum of human rights. Such practices may be commensurate with the size and resources of the supplier. Smaller entities may use simpler checklists that capture common risks, such as ensuring that sexual imagery and pictures of children are not comingled in a dataset, personal data is filtered out, and that datasets are representative of user populations. Larger entities may have more detailed risk identification, mitigation, and seller due diligence processes. Responsible practices for suppliers should also include adequate compensation and favorable working conditions for labor, such as data labelers or annotators. Owners of data that is repurposed into supplied datasets should also be adequately compensated for their time.

› Collaborate with developers to establish unified templates for data documentation, such as datasheets for datasets.

Data documentation methods like datasheets may allow developers to understand important details about a dataset prior to using it for training or fine-tuning. Those details should include composition, how it was collected, modality, format, purpose (where applicable), and methods of testing or evaluating key properties of the dataset. Creating a template and using it would reduce fragmentation and business costs associated with understanding each dataset's value and potential limitations. These standards may be established through industry associations and/or standards-setting bodies and can be included in contractual requirements by customers who purchase from these developers.

› Collaborate with foundation model and downstream developers to create standards and best practices for data annotation / labeling guidelines.

Separate guides or instructions may be created for general model training, training different types of models with defined downstream use cases, or for fine-tuning tasks. These should be established and standardized through industry associations and/or standards-setting bodies, and foundation model developers should require their use as a contractual condition.

6.3 Recommendations for Foundation Model Developers

› Create data procurement requirements to ensure training data is fit for purpose and supports privacy, equity, and safety; and evaluate datasets according to that criteria prior to procurement.

Source datasets that have accurate and transparent documentation and which are designed to account for crucial considerations such as privacy, equitable representation, and safety. This may include collaborating with dataset vendors, to create standardized dataset transparency artifacts and/or evaluation processes. Where datasets do not meet minimum requirements, work with data vendors to understand data procurement requirements and how to meet them. Foundation model developers may nuance these data procurement requirements to recognize that there are trade-offs involved with cleaning datasets at different stages of the pipeline. For instance, failing to include toxic language in the pretraining set may inhibit a foundation model's ability to recognize that type of content downstream.

› Create publicly available transparency documentation that accompanies the release of foundation models.

These documents, known as "model cards" or "system cards," should include key information regarding the model, including the purpose of the model, training and evaluation methods, safety mitigations, identified risks, and performance across various. The types of information that should be disclosed in the model card may vary according to context or purpose. This Model Card Toolkit is a useful resource for foundation model developers trying to create such

documentation. Examples of model cards from frontier model developers include the GPT-4 System Card and the Gemma Model Card. Foundation model developers should solicit feedback on these documents from external stakeholders to improve their utility over time.

› Create responsible use guidance containing recommendations and considerations for how downstream developers should responsibly build on foundation models.

Information about risks should draw from the developer's human rights and/or other risk assessments of their foundation models. The guidance should provide downstream developers with clear instructions on how to build applications with human rights risk mitigation integrated into design. Developers may consider incorporating alignment with these instructions as a contractual condition of downstream model provision. One example of a responsible use guide is Meta's Responsible Use Guide for LLaMa.

› Ensure that labor considerations for model training, labeling, or annotation activities are incorporated into human rights assessments.

Dataset curation and model training and development are often conducted by low-wage workers or contractors who may be exposed to poor working conditions, low wages, and distressing content. They may be hired in countries with poor labor rights protections or where the government may persecute them on the basis of their affiliation with US tech companies. Foundation model developers should address these human rights risks in their own operations, such as by only working with data enrichment providers that pay fair wages and provide tools to reduce the psychological impact of distressing content, like "greyscaling" for distressing images.[12] This can be done through contractual terms or questionnaires for their data suppliers.

› Make informed choices about model access and availability.

Where a model lies on the spectrum from closed source to open source may alter its risk profile. Once a model is made open source, it becomes irreversibly accessible to all actors, including bad ones, who may make copies of and modify the model. Strategies such as phased release can help address risks. Foundation model developers should use downstream HRDD findings to inform systematized processes for controlling model access. This may include implementing a screening or gating process for downstream developers to access foundation models. For additional recommendations for addressing the risks associated with open source foundation models, see Partnership on AI's guide on the topic.

› Work closely with downstream developers to take a collaborative approach to addressing risks that arise at the use case level.

Because foundation model developers are seldom aware of the specific applications of their model for individual use cases, it is not possible to mitigate all potential risk at the foundation model level. However, once risks related to a given use case are identified, some mitigation efforts may be most effective through direct adjustment to the foundation model. For this reason, foundation model developers should maintain open lines of communication with

12 A process of converting color images to only shades of grey or no color.

downstream developers to provide guidance on best practices for implementing technical safeguards. This should be a two-way line of communication where foundation model developers provide direct consultation and receive feedback from downstream developers. Furthermore, foundation model developers may help alleviate some of the road blocks that emerge due to resource discrepancies by supporting downstream developers to meet their responsibilities to mitigate risk.

› Invest in research on the impacts of using synthetic data to train or fine-tune models both to model performance and the data ecosystem.

As developers explore the use of synthetic data for training and fine-tuning genAI models, there should be a collaborative effort to identify and understand the impacts this has on issues such as model performance, in addition to existential impacts on the data ecosystem and data workers.

6.4 Recommendations for Downstream Developers

› Ensure that human rights due diligence considers the impacts of B2C or B2B business models, where applicable.

B2C and B2B products and services present different human rights risks, and the human rights due diligence processes for each should be context specific. For example, downstream developers may wish to engage different value chain, or value chain-adjacent actors, when conducting stakeholder engagement for B2C vs B2B product or service offerings. Developers should seek to understand the unique risks that come with the deployment of their product or service in a B2C vs B2B context and tailor their human rights due diligence accordingly.

› Engage deployers and individual users throughout the full development life cycle of a genAI system or product.

When developing a domain-specific genAI tool, downstream developers should proactively seek input from the intended users, AI safety and human rights professionals, and experts in that domain to ensure that the tool will be fit for purpose. For example, developing genAI tools for use in the clinical care setting should involve engaging healthcare professionals throughout the product development life cycle and using their insights and expertise to inform design choices. When developing genAI tools that are not domain-specific, downstream developers should conduct user research to ensure helpfulness and to understand how the tool may be misused or abused. Engagement with target users throughout ideation, product development and design, and post deployment should inform risk mitigation efforts.

› Design products to have easy-to-use reporting mechanisms for deployers, end users, and other stakeholders to flag issues.

The UNGPs stipulate that companies should implement effective grievance mechanisms for reporting human rights issues. Reportable issues might include inaccuracies in model

performance; outputs that are biased, discriminatory or otherwise harmful, or when a tool is being misused or abused. Reporting channels should be simple to use, prominently featured—ideally within the product interface itself—and include feedback loops that notify users of the status of their reports. To maximize the efficiency and uptake of these reporting channels, downstream developers should collaborate closely with deployers in channel design. These may be incorporated into any existing "report an issue" workflows.

6.5 Recommendations for Deployers

› Proactively and regularly report issues to the downstream developer.

If issues in the use or performance of a genAI tool emerge, deployers should ensure all users and operators of the tool are aware of how to report them to the developer. Users and operators should be particularly vigilant about reporting issues related to misuse or abuse, discrimination, toxicity, harmful content, and dangerous capabilities.

› Invest in capacity-building to promote cross-organizational genAI literacy and safety culture.

When deploying AI systems internally or as customer-, patient-, client-, or constituent-facing tools, deployers should ensure all users and operators at the organization understand how genAI works, the tool's capabilities and limitations, and application domain-specific risks. This could involve working closely with the downstream developer that created the tool and/or the entity that assisted with its set up or integration into existing systems (e.g., a reseller) in order to leverage their expert understanding of the tool's unique risks.

› Consider labor rights implications prior to integrating genAI-powered tools into workflows.

While some labor displacement is likely inevitable when integrating new technologies into workplaces, consider the severity of impact on workers. Work with downstream developers to inform the development of tools based on analysis of work produced from automation alone, human work augmented with automation, and human labor alone.

› Engage downstream developers to inform the design of systems that support workers.

Deployers have the greatest insight into the potential use cases and applications of genAI. Therefore they can helpfully inform developers on the features and functionalities that would be most useful in a genAI system for a given application. Additionally, they can inform the design of genAI systems which would assist and compliment workers rather than substituting them. This may in turn alleviate some of the potential labor disruption associated with widespread integration of genAI across sectors.

7. Appendix

Glossary of Terms

AI alignment: The process of encoding human values and goals into AI models.

AI ethics: The moral principles utilized by policymakers or companies to guide the design and deployment of AI systems.

AI governance: An umbrella term used to describe laws, principles, or processes used to direct AI design and deployment; this may occur within companies or by public authorities.

AI safety: A broad interdisciplinary field concerned with understanding and preventing harmful outcomes that arise from AI systems.

application domain: The area in which an AI system is integrated; this could refer broadly to a sector such as healthcare, finance, legal, or retail.

artificial intelligence (AI): A field of computer science concerned with designing complex computer systems capable of mimicking human cognition.

B2B: "Business to Business" refers to products and services intended for enterprise or public sector use.

B2C: "Business to Consumer" refers to products and services intended for direct consumer use.

foundation model: AI models trained on a broad set of unlabeled data that can be adapted for different tasks (such as conversing in natural language or generating text and images) with minimal fine-tuning.

frontier model: Large-scale AI models that exceed the capabilities of the most advanced existing models and can perform a wide variety of tasks.

generative AI (genAI): A type of AI that, in addition to recognizing patterns in training data, is capable of producing content in various modalities, including natural language text and speech, image, audio, video, and computer code.

genAI model vs. genAI system: A genAI model is the technology upon which generative AI systems are built. Users do not typically interface directly with a genAI model, rather they interface with a generative AI system, of which the generative AI model is a component. For example, ChatGPT is a generative AI system built on the GPT-4 model architecture.

hallucination: The output of false or misleading information by LLM-powered products.

large language model (LLM): A category of foundation models trained on immense amounts of data, making them capable of generating natural language and other types of content in response to user prompts to perform a wide range of tasks.

LLM-powered product: Consumer- or business-facing applications that are built on LLMs; this includes chatbot products such as ChatGPT, Google Gemini, or Claude.

machine learning: A subfield of artificial intelligence related to the development of algorithms and statistical models that enable computers to perform tasks without explicit instruction.

model safety: A non-normative assessment of an AI model's likeliness to contribute to harmful outcomes related to bias and discrimination, representation, toxicity, hate speech, or other criteria.

parameter: Values that control how a model generates a response; they are adjusted during training to optimize model performance. The more parameters a model has, the more generally capable the model is likely to be.

reinforcement learning through AI feedback: A process of fine-tuning a generative AI model that relies on an AI assistant to evaluate and label model outputs based on a defined set of principles; this process is used to improve model performance.

reinforcement learning through human feedback: A process of fine-tuning a genAI model that relies on human evaluation and categorization of model outputs, which is used to improve model performance.

Responsible AI: Voluntary company principles or processes intended to ensure that the design and deployment of AI is safe and ethical.

rightsholder: Under international human rights law, all individuals are entitled to exercise the full spectrum of universal human rights and fundamental freedoms and, therefore, are rightsholders.

token: Strings of characters that are combined to form words, which are used in the training of large language models.

user generated content (UGC): Content produced by individual users and posted to an online platform in the form of text, audio, image, or other formats.

Additional Resources

Business and Human Rights

- UN Guiding Principles on Business and Human Rights (UNGPs): A series of principles that set out the human rights responsibilities of companies.
- UNGPs Interpretive Guide: Provides guidance on how to interpret the UNGPs, alongside examples.
- OECD Guidelines for Multinational Enterprises on Responsible Business Conduct: OECD standards for all companies related to "responsible business conduct."
- OECD Due Diligence Guidance for Responsible Business Conduct: A sub-component of the OECD Guidelines that describes how companies should carry out due diligence. These guidelines are in line with the UNGPs.
- The Corporate Human Rights Benchmark's themes on governance and policy commitments (Theme A) and embedding respect for human rights and human rights due diligence (Theme B). The Corporate Human Rights Benchmark assesses the human rights disclosures of companies across industries to rate their human rights policies, processes, and practices. Indicators are grounded in the UNGPs and international human rights standards and can therefore be considered best practice.
- BSR's FAQ on Stakeholder Engagement: Answers common questions about stakeholder engagement for companies, including defining key terms and providing high-level best practices.
- DIHR's guidance on stakeholder engagement during human rights assessment processes: Provides guidance about how stakeholder engagement should be conducted as part of human rights assessments.

Responsible AI / Technology and Human Rights

- OECD Due Diligence Guidance for Responsible AI (forthcoming): Builds on OECD Due Diligence Guidance for Responsible Business Conduct to provide guidance for companies developing and using AI.
- The UN B-Tech Project: A UN initiative to provide authoritative guidance and resources for implementing the UNGPs in the technology sector.
- Taking Action to Address Human Rights Risks Related to End-Use (UN B-Tech Project): Provides guidance to companies about how to address human rights risks related to the use of technology.
- Risk Mitigation Strategies for the Open Foundation Model Value Chain (Partnership on AI): Provides guidance on risk mitigation across the value chain of open source foundation models.

- Guidance for Safe Foundation Model Deployment (Partnership on AI): A framework for model providers to responsibly develop and deploy foundation models.
- The Foundational Model Development Cheat Sheet: Describes best practices in open source foundation model development, including risk mitigations related to dataset procurement, model training and evaluation, and model release.
- The Importance of Data Quality (Hugging Face): Describes what constitutes "high-quality" data, why prioritizing data quality from the outset is crucial, and how organizations can utilize AI for beneficial initiatives while mitigating risks to privacy, fairness, safety, and sustainability.
- A Human Rights-Based Approach to Content Governance (BSR)**:** Describes how online platforms can take a human rights-based approach to content governance. The key points in this paper provided the foundation for this guide.
- Responsible Product Use in SaaS Sector (BSR)**:** Explores how SaaS companies should promote responsible use of their product and services, including through policies. It provides examples and analysis of different approaches to acceptable use policies, terms of service, and other policies.
- The Santa Clara Principles on Transparency and Accountability in Content Moderation**:** A set of human rights-based principles for content moderation devised by a broad coalition of organizations, advocates, and academic experts.

Remedy

- UN B-Tech Project's paper on access to remedy in the technology sector: An authoritative resource on the human rights principles behind providing remedy in the tech industry. The key points in this paper formed part of the foundation for this guide.
- UN B-Tech Project's paper on the remedy ecosystem in the technology sector: An authoritative resource on the various actors that comprise the remedy ecosystem in the tech industry. The key points in this paper formed part of the foundation for this guide.
- UN B-Tech Project's paper on designing and implementing effective company-based grievance mechanisms: An authoritative resource on how value chain entities can design grievance mechanisms. The key points in this paper formed part of the foundation for this guide.
- BSR's guide on Access to Remedy: An examination of the various components of remedy mentioned in this guide, generalized beyond the tech sector.

Conflict Minerals

- European Union Conflict Minerals Regulation (2021)
- Government Accountability Office—Report on Conflict Minerals (2023)
- SEC Conflict Mineral Disclosure

Environmental Toll of AI & Related Infrastructure

- Harvard Business Review (2024)
- MIT Press (2022)
- UN Environment Program (2024)

BSR™ is an organization of sustainable business experts that works with its global network of the world's leading companies to build a just and sustainable world. With offices in Asia, Europe, and North America, BSR™ provides insight, advice, and collaborative initiatives to help you see a changing world more clearly, create long-term business value, and scale impact.

www.bsr.org