A Human Rights-Based Approach to Risk Mitigation

Guide 4 of the Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI

February 2025

BSR®

Summary

This paper provides guidance on how Responsible AI practitioners can take a human rights-based approach to mitigating risks to people and society associated with generative AI (genAI). It includes the following sections:

1. **Human Rights Foundations for Risk Mitigation:** Describes where risk mitigation fits into the human rights due diligence process, how to utilize the UNGPs concepts of connection to impacts and leverage to inform risk mitigation, and other human rights principles that are important for risk mitigation.
2. **Addressing Tensions and Trade-Offs:** Describes how a human rights methodology called "counterbalancing" can be utilized when there are tensions or trade-offs among risk mitigations.
3. **Integrating Human Rights into GenAI Risk Mitigation Approaches:** Provides a categorization of mitigations that can help Responsible AI practitioners think expansively about risk mitigation, and describes how common genAI mitigations can address human rights risks.
4. **Human Rights Risk Mitigation Across the GenAI Value Chain:** Illustrates how a human rights-based approach can be applied to existing efforts to address upstream and downstream risk by each genAI value chain category.
5. **Key Resources**

Key Points

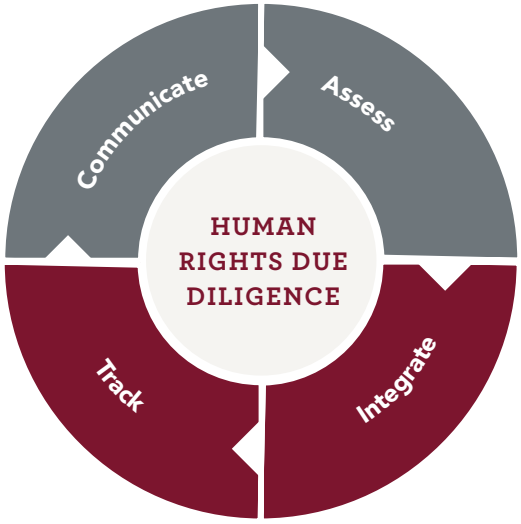
- Key aspects of taking a human rights-based approach to risk mitigation are:
  - Understanding how the entity is connected to / involved with impacts, such as whether it is causing, contributing to, or linked to an adverse impact
  - Identifying the leverage the entity has to address impacts, to inform what mitigation measures to pursue and how to implement them most effectively.

ACCOMPANYING RESOURCES

  - Not allowing for "risk tolerance" or "offsetting" risks against benefits
  - Identifying risks that arise from mitigations
  - Integrating mitigations across the organization
- When dealing with tensions and trade-offs (aka "competing equities" or rights in tension), a counterbalancing exercise can be used to identify a rights-respecting path forward.
- When identifying potential mitigations, it can be helpful to think expansively across the following categories: technical/product actions, policies, processes, transparency/communication, and collaborative actions.
- There are many established mitigation measures to address risks associated with genAI that also address some human rights impacts; however, they can be improved upon to more effectively address risks to all internationally recognized human rights.

1. Human Rights Foundations for Risk Mitigation

After identifying the risks to people and society associated with a genAI system, the next step is to mitigate those risks. Identifying mitigations is often part of broader AI impact assessment processes, and can occur any time risks are identified. In most cases, standalone assessments of AI impacts, such as those conducted by external parties, include a series of recommended risk mitigations.



The UN Guiding Principles on Business and Human Rights (UNGPs) describes risk mitigation as taking "appropriate action" to "avoid, cease, prevent, or mitigate" (or collectively, "address") adverse human rights impacts.[1] Because "mitigation" is the dominant term utilized in the Responsible AI field for any action that addresses a risk, this guide uses that term broadly as well.

Risk mitigation occurs as part of the third and fourth steps of human rights due diligence—**integrating and acting upon the findings from impact assessments** across relevant internal functions and processes to address adverse impacts and **tracking the effectiveness of the response**.[2]

1 Actions include preventing an impact, ceasing an impact, avoiding an impact, or mitigating an impact. The term "address impacts" encompasses all actions. The UNGPs Interpretive Guide defines mitigation as actions taken to reduce the extent of an actual human rights impact or reduce the likelihood of a risk occurring, with any residual impact then requiring remediation.

2 See Principles 19 and 20 of the UNGPs.

Understanding Attribution and Leverage to Inform Mitigations

The UNGPs outline two factors that should inform "appropriate action" to mitigate risks: 1) attribution—how the entity is connected to/involved with the impact, and 2) leverage—the extent of the leverage the entity has to address the impact. Each factor is explained below.[3]

- **Attribution reflects how closely connected the entity is to the impact.** Attribution is a spectrum, and companies may fall under four points on the spectrum. In decreasing order of the closeness of the connection, these points are:

  Cause the impact:
  The entity is considered to cause an impact when its **actions or omissions (i.e., failure to act) alone are sufficient to have caused the human rights impact.**

  – EXAMPLE: If a private security company designed a genAI surveillance tool and used it to surveil political dissidents, it would cause the impacts to privacy.

  – APPROPRIATE ACTION: The entity must take the necessary steps to **cease or prevent the impact**, as well as **provide remedy to affected individuals** if the impact occurs. Remedy is discussed in Guide 8: Remedy for Generative AI-Related Harms.

  Contribute to the impact:
  The entity is considered to have contributed to the impact if it has **facilitated, enabled, incentivized, or motivated another party** to cause the human rights impact through its actions or omissions.

  – EXAMPLE: If an AI data supplier knowingly provided a developer with datasets containing significant amounts of toxic and discriminatory content, the supplier would contribute to the downstream impact of discrimination.

  – APPROPRIATE ACTION: The entity 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. It should also **provide remedy to affected individuals** if the impact occurs.

  Linked to the impact:
  An entity is considered to be linked to an impact when it is involved solely because **the impact is directly linked to its operations, products, or services by a business relationship**.

3 For more information, see Principle 19 of the UNGPs, the UNGPs Interpretive Guide, the UN B-Tech Project Paper, Taking Action to Address Human Rights Risks Related to End-Use, and BSR's Seven Questions to Determine a Company's Connections to Human Rights Abuses.

– EXAMPLE: If a developer offers a general purpose genAI chatbot that has robust safety protections, but bad actors jailbreak it to create disinformation campaigns, the developer would be linked to the impact.

– APPROPRIATE ACTION: The entity should **utilize its leverage to address the impact**.

Not linked to the impact:
Companies fall into this category if their operations, business relationships, or products and services have **no link to any adverse human rights impacts associated with genAI**. Most companies that do not use genAI technology fall into this category.

DETERMINING WHETHER AN ENTITY IS "CONTRIBUTING" TO AN IMPACT

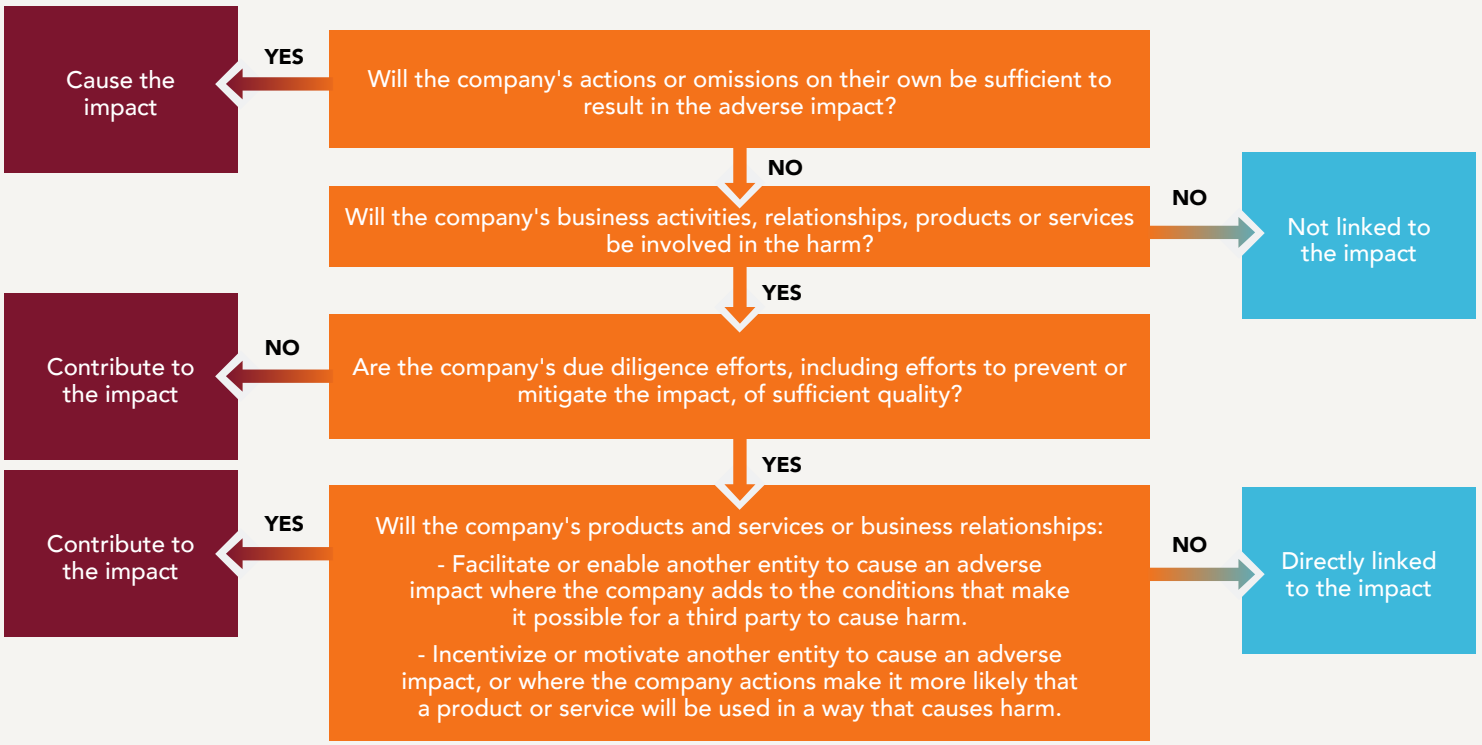
Determining whether an entity is contributing to vs. directly linked to an impact can be challenging. The UN B-Tech Project's paper on Taking Action to Address Human Rights Risks Related to End-Use describes the following factors that can be used to determine whether a technology company is contributing to an impact, which are also relevant for genAI impacts:

- **Facilitating or enabling:** Occurs when a company's actions or decisions add to the conditions that make it possible for the use of a product by a third party to cause harm if it is inclined to do so. For example, by customizing products in ways that increase the possibility of misuse, or by providing a component or input into an existing technology that ends up being a critical factor in that technology's use for harm.
- **Incentivizing or motivating:** Occurs when a company's actions or decisions during product design, promotion, and marketing exacerbate or perpetuate behaviors that lead to harm. For example, designing consent processes in ways that make it more challenging for users to make informed privacy choices, or the known use of biased training data that ends up perpetuating social patterns of discrimination.
- **The effectiveness of a company's human rights due diligence:** If a technology company knows or should have known about a given human rights impact, but fails to take any action to address it, it may find itself in a situation of contribution rather than directly linked. For example, failure to develop policies that prohibit certain uses, or utilizing certain design features that could have been reasonably foreseen to increase the risk of adverse impacts.

- **Leverage reflects the ability of the entity to influence/address the practices of the entity causing the harm.** The UNGPs state that if a business has leverage to address impact, it should exercise it, and if it lacks leverage, it should seek to increase it. There are a wide variety of ways in which entities across the genAI value chain can exercise leverage on other entities, both upstream and downstream. For example, foundation model developers have leverage over both suppliers and app developers. There are many mechanisms for exercising leverage that are also relevant risk mitigations, such as contractual requirements, due diligence processes, training, UX design features, direct engagement with the responsible party, or collaborative efforts with other relevant actors.

As mentioned above, the "cause, contribute, directly linked, not linked" framework operates on a continuum. The distinction between "contribute" and "directly linked" can be blurred, will often depend on a range of contextual factors, and can also shift over time.[4] The flowchart below can be used to help illuminate connection of an entity to different impacts, although it is illustrative rather than fully determinative. Responsible AI practitioners do not need to do this analysis for every single risk or impact identified in an impact assessment, however, **thinking through the connection to impacts and the leverage that exists to address them can usefully inform what mitigation measures to pursue and how to implement them most effectively.**

"Cause-Contribute-Directly Linked" Framework



4 See the UN B-Tech Project's paper on Taking Action to Address Human Rights Risks Related to End-Use.

Other Human Rights Principles for Risk Mitigation

- **Risk tolerance and "offsetting" risks and benefits is not acceptable in a human rights context.** One component of risk mitigation commonly found in some impact assessment approaches is the notion of "risk tolerance." This is the idea that an entity can choose to accept or internalize a certain amount of risk, and therefore will take no action or limited action to address it. This is often justified via "offsetting," which is the idea that adverse impacts can be canceled out or outweighed by benefits.

  While perfect risk mitigation is not possible, and trade-offs may need to be made due to rights in tension, the UNGPs are clear that businesses should take steps to address all potential human rights impacts.[5] The UNGPs also recognize that while companies may undertake activities to support and promote human rights, positive impacts do not "offset a failure to respect human rights.[6] This means the concepts of "risk tolerance" or "offsetting" as understood in other approaches to risk management are not appropriate when it comes to risks to human rights.

- **It is important to also identify risks that arise from mitigations.** Sometimes efforts to address human rights impacts can themselves come with human rights risks. For example, model safety efforts may overly constrain a genAI model's ability to provide desired outputs to users, thereby restricting their access to information. When identifying and deploying risk mitigations, it is therefore important to consider what risks those mitigations may have and take action to address them. This often requires an iterative, ongoing process as mitigations necessarily evolve alongside the evolution in genAI products and how they are used.

- **Integrate risk mitigation across the organization.** It is important not to view risk mitigation of a given genAI product or service in isolation. Oftentimes, mitigations identified for a specific product or service are best implemented as part of broader organizational processes (for example, via development of a new policy or tweaks to a due diligence process). This is supported by the UNGPs, which recommend horizontal integration across the business of specific findings from human rights assessments.[7]

- **Tracking the effectiveness of risk mitigation over time.** The UNGPs state that in order to verify whether adverse human rights impacts are being effectively addressed, it is important to track the effectiveness of risk mitigation efforts. Tracking can be based on both qualitative and quantitative indicators, and should draw on feedback from both internal and external sources, including affected stakeholders. This can be done via existing internal processes if relevant (e.g., as part of regular monitoring of model performance metrics or analyzing data from user reporting channels or other grievance mechanisms.[8]

5 Although the UNGPs expects companies to address all impacts, when necessary they can prioritize actions based on severity. See Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI and Guide 2: Human Rights-Based Approach to Risk Assessment to review this concept

6 See Principle 11 of the UNGPs.

7 See Principle 19 of the UNGPs.

8 See Principle 20 of the UNGPs.

2. Addressing Tensions and Trade-Offs

Because human rights are interconnected, they can come into tension with one another for legitimate reasons (see Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI to review this concept). Within the Responsible AI field this is often referred to as "competing equities," and it can create challenges for risk mitigation. One example of this is the tension between helpfulness and harmfulness—mitigations that seek to reduce the harmfulness of genAI outputs can also reduce their helpfulness. Simply put, model safety measures that reduce harmful outputs will result in the refusal of more input requests, which users may perceive to be less helpful. For example, a genAI chatbot might misclassify a question about why a racial slur is considered to be offensive to be harmful because it contains a slur. In human rights terms, this example of the tension between helpfulness and harmfulness is a tension between users' right to access information versus the right to nondiscrimination.

When such tensions occur, human rights principles can be used to define a rights-respecting path forward. BSR refers to this approach as "**counterbalancing**," which is a methodology inspired by the structured reasoning of human rights courts when dealing with cases involving competing rights. The goal of counterbalancing is to help identify ways to secure the fullest possible expression of the competing rights without unnecessarily limiting them. This methodology is consistent with the notion that most human rights are not absolute and can be limited in certain legitimate circumstances, such as to keep people safe.[9]

Counterbalancing can be done by considering the following international human rights principles when thinking through risk mitigation involving competing rights:

- **Reverting to principle:** Can the core principle of the restricted right still be upheld in different ways?
- **Legitimacy:** Is there a legitimate aim in pursuing the restriction of this right?
- **Necessity:** Is the restriction of the right necessary or can the legitimate goal be achieved through other means?

9 The situations in which human rights can be restricted are outlined in the various international human rights instruments. For example, Article 4 of the ICCPR allows states to limit certain rights during public emergencies; Article 17, which outlines the right to privacy, describes infringements on privacy as those that are "arbitrary or unlawful;" and Article 19, which outlines the right to free expression and opinion, states that restrictions must be provided by law and necessary "for respect of the rights or reputations of others" or "for the protection of national security, public order, public health, or morals."

- **Proportionality:** Is it the least intrusive way to restrict the right?
- **Nondiscrimination:** Can the restriction of the right be done in a nondiscriminatory manner?

Carrying out a counterbalancing exercise can sometimes reveal a single clear solution to a given tension, but may also reveal a range of possible solutions that are "rights-respecting." In the context of risk mitigations, this means there may be a range of potential mitigations that uphold the principle of the restricted right/s, and are legitimate, necessary, proportionate, and nondiscriminatory. In such cases, stakeholder engagement to gather feedback on different solutions can help inform a path forward. See the call out box below for an example of counterbalancing in the context of mitigating risks associated with genAI. For other examples of counterbalancing in practice, see BSR's HRIA of the Tech Coalition's Lantern Program and BSR's HRIA of Meta's Expansion of End-to-End Encryption.

COUNTERBALANCING HUMAN RIGHTS IN TENSION EXAMPLE: ACCESS TO INFORMATION AND POLITICAL PARTICIPATION

Context: GenAI chatbots, such as those released by companies like Google and OpenAI, are increasing in popularity among users as a resource to access information and are being used in ways similar to search engines. However, for a number of reasons—such as limitations in training methods and their tendency to "hallucinate"—LLMs are not always capable of presenting accurate, unbiased, and representative information in response to user queries. For this reason, most developers place restrictions on the topics about which chatbots are able to produce responses. In these cases, developers must weigh the trade-offs that come with limiting access to information. A common restriction is political/elections-related content, due to the possible harms that may result from presenting information that is inaccurate, biased, and/or nonrepresentative. This results in a tension between the right to access information and the right to political participation.

The below points are things that developers should consider when determining the best approach for navigating the trade-off between the rights to access information (Article 19 of the ICCPR) and political participation (Article 25 of the ICCPR):

- **Reverting to principle:** The underlying principle of the right to access information in the context of political participation and elections is that people should be able to access accurate and unbiased information about political candidates, electoral processes, policy proposals, or other topics related to a political figure's campaign. Within this context, the question developers could ask is "can access to information be preserved in ways that ensure accurate, unbiased, and pluralistic representation of elections-related topics?"

Given the known challenges related to LLMs (e.g., inaccuracies in model outputs are common, current training methods do not support diversity of opinion in model outputs, models tend to support rather than challenge users' beliefs, etc.) and the pressing need for unbiased and authoritative information about elections-related topics, it may not be possible—in the context of chatbots—to uphold the underlying principle of access to information as it relates to political participation. Additionally, users have other options to access elections-related information, which means the core principle of the right is upheld in other contexts (e.g., through news outlets, traditional internet searchers, etc.).

- **Legitimacy:** Limiting access to information to protect against the spread of false, biased, unrepresentative, or otherwise potentially harmful elections-related content that could adversely impact the right to political participation is a legitimate aim. Democracies around the world have long emphasized the need for free and fair elections and transparent political processes. Actions in service of those goals are legitimate to a larger democratic process.
- **Necessity:** Given the current capabilities and limitations of genAI, restricting all content related to elections may or may not be necessary. Developers should assess whether their individual proprietary models can surface reliable, unbiased information related to elections in some contexts. If not possible now, developers may consider how this could be achieved through advancements in model capabilities, technical and/or policy safeguards, or other means.
- **Proportionality:** Restricting all content related to an election also may or may not be proportionate. For example, it may be possible to allow some elections-related content to preserve access to information (e.g., basic facts about election dates, polling locations, who the candidates are, etc.). The proportionality of the proposed restrictions may vary depending on the context. For example, more restrictions may be justified during an election month, where hallucinations could have the most significant impact on the right to political participation.
- **Nondiscrimination:** A decision to restrict elections-related content could impact all users, including both those who are and are not eligible to vote. A full restriction on election-related content may prevent potential cases of the chatbot presenting some candidates more favorably than others.

3. Integrating Human Rights Into Generative AI Risk Mitigation Approaches

Effective mitigation to address the human rights risks associated with genAI will ultimately differ depending on the system/product, the use case, and an entity's place in the genAI value chain. However, it can be helpful to think across the following categories when identifying mitigations. In BSR's experience, Responsible AI practitioners often focus primarily on the first two categories without sufficiently considering the others:

- **Technical / product actions** include any data, engineering, or product design mitigations (e.g., data quality efforts, model training / fine-tuning techniques, automated detection and enforcement, UX design choices). This can also include user empowerment via tools, controls, and education.
- **Policies** include product, user or customer policies (e.g., acceptable use policies, contractual clauses), as well as internal or employee-facing policies (e.g., data access policies) and processes to enforce them. See Guide 6: Human Rights Based-Approach to Policies and Enforcement for more information.
- **Processes** include any process designed to identify and address risks over time (e.g., ongoing impact assessment processes, ongoing model evaluation approaches, sales due diligence process, internal escalation channels and review committees, employee / customer training).
- **Transparency / Communications** includes any information sharing about the model / system / product that could help address impacts, such as how it functions, the dataset it was trained on, risks, mitigations, and best practices for use (e.g., via dataset documentation, white papers, model or system cards, developer guidance, and other types of documentation). For more information, see Guide 7: Aligning Transparency and Disclosure Practices with Human Rights Responsibilities.

- **Collaborative actions** include activities that the entity can pursue in collaboration with other stakeholders that address systemic or industrywide impacts (e.g., participating in multi-stakeholder fora such as the Frontier Model Forum or standards-setting processes, engaging with policymakers, such as the White House Voluntary AI Commitments). Collaborative actions are also important for addressing cumulative impacts connected to multiple companies.[10]

There are a wide variety of common mitigations for genAI risks known in the Responsible AI field. The section below discusses how a few of the common mitigations at different points in the genAI value chain address some human rights impacts, and how they can be improved to more effectively address impacts all human rights. This is illustrative, and is not intended to be an exhaustive list of genAI risk mitigations. A table summarizing additional common mitigations is in the following section.

Suppliers and Foundation Model Developers:

- **Privacy best practices and privacy preserving approaches** for data collection and model training and operation are designed to address the many privacy impacts associated with training and deploying genAI models. Privacy impacts include the addition of personal and/or sensitive data in training datasets and the use of genAI models in sensitive domains, which may involve data collection without informed consent or the exposure of personal information by genAI tools.

  Privacy preserving approaches are an evolving area of research and practice across the Responsible AI field. Established approaches include data cleaning (e.g., removing personal information and metadata from a dataset that could be used to re-identify people), model training with differential privacy techniques (the model learns from large amounts of data without remembering or outputting the data of individual users), on-device processing (aka "edge AI"), and federated learning, which can also be applied to genAI products and services.

  Privacy is known as an "enabling right," meaning that having privacy enables people to feel comfortable exercising other rights, such as freedom of expression, and adverse impacts on privacy can also lead to adverse impacts on other rights. For example, genAI video surveillance tools could lead to physical safety impacts for those whose information is exposed, especially if used to support state surveillance of perceived critics. Therefore, mitigations that address privacy impacts often address connected human rights impacts as well.

  However, there can be tension between addressing privacy impacts via privacy preserving approaches and addressing impacts related to misuse or abuse of genAI tools that require the ability to monitor user activity. When this tension arises, counterbalancing can be used to identify the best approach.

Foundation Model Developers and Downstream Developers:

10 For more information about collaboration to address cumulative impacts, see Part 3.1 of the OECD Due Diligence Guidelines; for more about cumulative impacts, see Guide 3: A Human Rights-Based Approach to Impact Assessment.

- **Model alignment techniques** are designed to reduce the likelihood that models will produce harmful outputs, such as instructions to create a bioweapon, biased and discriminatory text or videos, hallucinations, etc. Harmful outputs from genAI models may be associated with impacts to most, if not all, human rights. A few common examples of model alignment techniques include Reinforcement Learning from Human Feedback (RLHF), Reinforcement Learning from AI Feedback (RLAIF), and rule-based rewards (RBR).

  Many human rights impacts associated with genAI stem from bias and inaccuracy in datasets, and therefore model alignment techniques that seek to mitigate those issues also mitigate associated human rights impacts. The effectiveness of these approaches for addressing impacts can be expanded by utilizing them to directly mitigate human rights impacts identified in an impact assessment. This could include using human rights language to steer model behavior. One example of this is Anthropic's use of the Universal Declaration of Human Rights to ground which values are imbued into its model, Claude, through RLAIF. However, each of these approaches also come with limitations that can lead to human rights impacts, and therefore it is important to anticipate and address these to the extent possible.

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

  Privacy and safety-by-design approaches often involve building in established mitigations (e.g., data minimization, model safety measures) early in the product development process. However, they can be expanded to be "human rights-by-design" by also integrating mitigation of known human rights impacts. This is also helpful for surfacing any tensions and figuring out how to best address them before they are "baked in" to the product and become more complicated to address.

  Human rights-by-design approaches should include consideration for how a given mitigation may impact downstream performance, and upstream actors should work closely with downstream actors to implement human rights by design at the most effective point in the genAI tool's lifecycle.

Deployers:

- **Developing and enforcing policies** for the use of genAI systems. Policies related to the development, deployment, procurement, sale, and use of genAI play an important role in addressing adverse human rights impacts. Product policies are oriented toward the customer or end-user of a product or service, and are therefore an important mitigation for addressing the risks associated with use of that product or service by outlining the intended use of a product, limitations on use, and guidance on actions to be taken when a product is used outside of its intended purpose. Relevant policies include terms of service, acceptable use policies, model policies, and content policies. For more information, see Guide 6: A Human Rights-Based Approach to Policies and Enforcement.

4. Human Rights Risk Mitigation Across The Generative AI Value Chain

The below table[11] illustrates how a human rights-based approach can be applied to existing efforts to address upstream and downstream risk using an example risk mitigation from each genAI value chain category. (See Section 4 of the HRA of the Generative AI Value Chain for a detailed description of the value chain.)

Note that the UNGPs identify companies as the primary duty bearers of respecting human rights. For this reason, individual users are not included below; however, they are part of the genAI value chain and should be aware of how their use of genAI systems may be associated with adverse human rights impacts and avoid deliberate misuse.

Suppliers	
Example Mitigation	**Data documentation**
Description	The process of recording and describing the characteristics and properties of data to increase its usability and provide transparency Examples include: data statements, datasheets, data nutrition labels, dataset cards, and dedicated research papers

11 Key sources for the table include:
- The Foundational Model Development Cheat Sheet
- The Importance of Data Quality—Hugging Face
- Risk Mitigation Strategies for the Open Foundation Model Value Chain—Partnership on AI
- Overview of Responsible AI Practices for Azure OpenAI Models—Microsoft
- Responsible AI Practices—Google
- Meta Llama Responsible Use Guide—Meta.

Risks It Addresses	**Downstream**: compromised model performance due to lack of data relevance / inappropriate applications of data, which may lead to adverse impacts on people during deployment **Downstream**: representational harms in model performance, which may lead to adverse impacts on people during deployment
How to Integrate a Human Rights-Based Approach	• Consider how inappropriate application of the dataset could result in adverse impacts on people (i.e., use in unintended / sensitive domains or for potentially discriminatory purposes) • Stipulate "no go's" in the data documentation (i.e., uses that could lead to adverse impacts on people) • Consider where the dataset may have gaps in representation (e.g., unequal or imbalanced representation of individuals or communities) • Where representational gaps cannot be addressed, flag them in the data documentation

Foundation Model Developers

Example Mitigation	**Managed model rollout**
Description	The process of making informed and calculated decisions about model release on a gradient from closed to open source, which may include receiving feedback from relevant stakeholders to inform iterative model fine-tuning
Risks It Addresses	**Upstream:** gaps in training data that lead to model performance issues, such as unequal performance across demographic groups and communities or outputs that are not representative of or relevant for the user base **Downstream:** applications of foundation models associated with adverse impacts on people
How to Integrate a Human Rights-Based Approach	• Consider how human rights could be adversely impacted by the application of foundation models into technologies across a range of use cases • Consider how human rights could be adversely impacted by limiting access and availability of the foundation model • Identify any tensions in the above scenarios (i.e., adverse impacts from closed vs. open source) • Consider what leverage the foundation model developer has to mitigate and provide remedy for harms resulting from downstream applications of the model • Establish decision-making processes for navigating human rights trade-offs related to types of model release that secure the fullest possible expression of the competing rights without unduly limiting them

Downstream Developers	
Example Mitigation	Sales due diligence
Description	The process of assessing a prospective customer and/or use case prior to the sale of a product or service
Risks It Addresses	**Downstream:** intended or unintended use of a genAI tool associated with adverse impacts on people
How to Integrate a Human Rights-Based Approach	• Consider how human rights could be adversely impacted by a prospective customer's operations generally (e.g., law enforcement agencies have the power to restrict many rights.) • Consider how the downstream developer could cause, contribute, or be directly linked to the potential human rights impacts related to proposed use of a genAI tool. • Perform more in-depth due diligence on potentially high-risk customers
Deployers	
Example Mitigation	**Human oversight**
Description	The process of human verification and/or approval of AI-generated predictions, decisions, or other outputs
Risks It Addresses	**Upstream:** Upstream: instances in which programmed safeguards fail to prevent inaccurate, inappropriate, biased, or otherwise harmful system outputs **Downstream**: unintended adverse impacts on people associated with AI outputs
How to Integrate a Human Rights-Based Approach	• Consider how the integration of genAI tools into operations, processes, or workflows could cause, contribute to, or be directly linked to potential adverse human rights impacts • Be aware of issues that could increase the likelihood of adverse human rights impacts (e.g., inaccurate or biased outputs, harmful behavior specific to the use case and application domain, etc.) • Ensure users/operators are adequately trained on how genAI works, aware of potentially harmful outputs and behavior, and review outputs accordingly • Establish processes for flagging and escalating harmful outputs • Establish decision-making processes for navigating human rights trade-offs that secure the fullest possible expression of the competing rights without unnecessarily limiting them

5. Key Resources

The following resources contain more detailed information about how to take a human rights-based approach to risk mitigation, as well as more information about common genAI risk mitigations. There are a wide variety of resources related to genAI risk mitigation, and the list below is therefore just a starting point.

- **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.
- **OECD Due Diligence Guidance for Responsible AI** (forthcoming): Builds on OECD Due Diligence Guidance to provide guidance for companies developing and using AI, including on taking action to address impacts and tracking results.
- **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.

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

Copyright © 2025 by Business for Social Responsibility (BSR)
All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical reviews and certain other noncommercial uses permitted by copyright law.