A Human Rights-Based Approach to Policies and Enforcement

Guide 6 of the Responsible AI Practitioner Guides for Taking a Human Rights-Based Approach to Generative AI

February 2025

BSR®

Summary

This paper provides guidance to Responsible AI practitioners about how to integrate human rights into the development and enforcement of policies related to generative AI (genAI). It includes the following sections:

1. **Policies and Enforcement as GenAI Risk Mitigations:** Discusses the role that product policies and employee-facing or organizational policies play in mitigating impacts to people and society associated with genAI.
2. **Integrating Human Rights into Policies and Enforcement:** Highlights the benefits and key elements of a human rights-based approach to policies and enforcement.
3. **Key Resources**

Key Points

- Product policies relevant for mitigating genAI risks include terms of service, acceptable use policies, model policies, and content policies.
- Organizational or employee-facing policies relevant for mitigating genAI risks include Privacy / Data Protection Policies, Security / Data Access Policies, Quality Assurance Policies, Compliance Policies, and Intellectual Property Policies.
- Taking a human rights-based approach to policies and enforcement enables organizations to take a consistent global approach across borders and to carry out enforcement in a way that is thoughtful, consistent, and rights-informed.
- Key elements of a human rights-based approach to policies and enforcement include:
  - Policies should reference human rights when relevant
  - Policies should address impacts to all human rights

ACCOMPANYING RESOURCES

A HRA of the GenerativeAI Value Chain
Overview of the Practitioner Guide
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 Disclosures
Guide 8: Remedy for GenAI Related Harms

– Policies and enforcement approaches should adhere to the human rights principles of legitimacy, necessity, proportionality, and nondiscrimination

– Policy development and enforcement should be informed by stakeholder engagement

– Policy development should address the most severe impacts first

– Policies and enforcement should honor the principles of internationally recognized human rights when faced with conflicting requirements

– Policies enforced in conflict-affected areas should receive enhanced attention

– Reporting and appeals channels should be established to help identify policy violations and correct enforcement errors

– Review enforcement and update policies on an ongoing basis

1. Policies And Enforcement as Generative AI Risk Mitigations

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 therefore play an important role in addressing the impacts 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. As such, actions related to product policies are commonly recommended in human rights assessments (HRAs) of any technology-related product or service.

Product policies relevant for the genAI value chain include:

- **Terms of Service (ToS):** ToS are a legal agreement between the provider and the user of a product or service that outlines the rules users must adhere to when accessing and using the product or service. They are also known as "terms and conditions," "terms of use," or "end user license agreements." They often cover the entirety of an entity's products and services or a category of products and services, and as such tend to be higher level. One example is Google's ToS.
- **Acceptable Use Policies (AUPs):** AUPs are a set of rules that govern how a specific product or service may be used, such as by users or developers. At their most basic, AUPs typically include a description of prohibited uses of the product or service and a description of how the AUP is enforced. One example of an AUP is OpenAI's usage policies.
- **Model Policies:** In contrast to AUPs which dictate safe user behavior, model policies govern safe behavior by the model. These policies are usually written by foundation model developers and may include rules that instruct the model to avoid outputting child sexual abuse material or dangerous instructions. Layering model policies with AUPs and other policies ensures differentiated protection from different threats, such as poorly phrased user instructions, jailbreak attempts, and model errors. Model policies are usually kept

confidential, as bad actors may attempt to jailbreak them if they were aware of their precise wording. One example of publicly released model policies are Anthropic's system prompts for its large language model, Claude.

- **Content Policies:** Content policies outline what type of content is or is not permitted on a product or service and how those policies are enforced. Content policies are most associated with online platforms oriented toward user-generated content, and are often called "community guidelines." In the context of genAI products and services, content policies might outline content a genAI chatbot tool cannot be used to generate, or a social media platform's approach to flagging synthetic content (e.g., Meta's approach to labeling AI-generated content).
- **AI Principles:** These are a high-level set of ethical commitments made by companies to govern their use of AI. Examples include "fairness," "accountability," or "transparency." AI principles are often operationalized by companies in different workflows, including processes that require reviews of products, business decisions, or other policies for alignment with the principles. More information about integrating human rights into AI principles can be found in Guide 2: A Human Rights-Based Approach to Governance and Management.

Organizational / employee-facing policies can also be relevant for mitigating risks associated with the development and deployment of a genAI system. These include, but are not limited to:

- **Privacy / Data Protection Policies**—Govern the collection, use, storage, and protection of personal data to ensure user privacy when interacting with genAI products.
- **Security / Data Access Policies**—Define protocols for controlling and safeguarding access to genAI systems and data, preventing unauthorized use or breaches.
- **Quality Assurance Policies**—Establish standards and procedures to ensure that genAI outputs are accurate, reliable, and meet the organization's quality criteria.
- **Procurement Policies**—Outline the procedures an entity must follow when purchasing goods and services, such as data labeling services or a genAI tool.
- **Compliance Policies**—Ensure that the use of genAI products adheres to all applicable laws, regulations, and industry standards.
- **Intellectual Property Policies**—Govern the ownership, usage rights, and protection of content generated by genAI products, respecting both the organization's and third parties' intellectual property rights.

What a given policy contains and how it is enforced will depend on where the entity is located in the genAI value chain, the nature of the product or service, the intended use case, and the customer / context in which it is used. For example:

- An AUP or model policy for a foundation model is likely to be much more high-level and general than an AUP for a genAI product designed to be used in a healthcare context.

- Content policies for a genAI chatbot designed for use by children are likely to be much more restrictive than content policies for a general purpose genAI chatbot.
- Content policies can typically be enforced by the deployer of the genAI system via automated and human review systems, whereas terms of service violations by an end user of an enterprise genAI product may need to be reported to the deployer or developer to be enforced.

2. Integrating Human Rights Into Policies and Enforcement

The Benefits of a Human Rights-Based Approach to Policies and Enforcement

In order to effectively address human rights impacts, it is important that human rights be deliberately integrated into policy development and enforcement. Human rights should shape how genAI products and services are developed, how they may be used, and how policies are enforced. There are three key benefits to this:

- **Global approaches across geographic borders:** Human rights-based approaches enable consistent approaches to be taken across international borders, including jurisdictions with no relevant regulations and jurisdictions or where laws and regulations conflict with international human rights standards. A human rights-based approach can be a counterweight against problematic local laws, and provides a strong foundation from which to push back against the actions of governments that restrict human rights. The commentary to Principle 23(b) of the United Nations Guiding Principles on Business and Human Rights (UNGPs) states that "where the domestic context renders it impossible to meet [the responsibility to respect human rights] fully, business enterprises are expected to respect the principles of internationally recognized human rights to the greatest extent possible in the circumstances."
- **Enforcement that is thoughtful, consistent, and rights-informed:** A human rights-based approach emphasizes process as much as the enforcement decision; international human rights norms provide an overall framework for decision-making and action rather than a "copy and paste" set of rules. Decisions should be intellectually consistent, defensible on human rights grounds (i.e., tensions are worked through using counterbalancing—see Guide 4: A Human Rights-Based Approach to Risk Mitigation), and conveyed transparently to those they impact.

- **Flexibility to adapt to evolving risks over time:** Grounding policies and enforcement in human rights also enables policies and enforcement to adapt over time. Policies can articulate the impacts the entity is trying to prevent rather than attempting to list every prohibited scenario or use of a product or service.

Key Elements of a Human Rights-Based Approach to Policies and Enforcement

Human rights can be integrated into the development and enforcement of policies related to genAI via the key elements below:

- **Policies should reference human rights when relevant.** In some cases, referring to an organizational human rights policy or simply stipulating that use of a product must not adversely impact human rights may be sufficient. For example, Microsoft's genAI code of conduct prohibits uses of genAI "to make decisions without appropriate human oversight as part of an application that may have a consequential impact on any individual's legal position, financial position, life opportunities, employment opportunities, or human rights, or may result in physical or psychological harm to an individual." In others, it may be important to include explicit references to the fact that the policy is grounded in international human rights standards, such as the International Bill of Human Rights. Other relevant thematic human rights instruments may also be referenced when relevant (e.g., the Child Rights Convention for products designed for use by children).
- **Policies should address impacts to all human rights.** While impacts on certain rights may be more salient than others, policies can preempt the evolution of emerging threats or rights impacts by being inclusive of the full spectrum of human rights. To do this, organizations can undertake a gap analysis between their product policies and all internationally recognized human rights (see Guide 3: A Human Rights-Based Approach to Impact Assessment) to make sure all impacts are appropriately covered. For an example of what this looks like in practice, see Annex 1 of BSR's HRIA of Twitch.
- **Policies and enforcement approaches should adhere to human rights principles.** When a policy and/or enforcement decision has the potential to impact human rights in some way (e.g., loss of access to a key service, a restriction on free expression) it is important that they be grounded in UN principles for restricting human rights:
  - **Legitimate**—The policy and enforcement decisions address clearly defined threats and/or violations.
  - **Necessary**—The policy is necessary and the goal cannot be achieved by other means.
  - **Proportionate**—The penalties are commensurate with the nature of the policy-violation (e.g., they take into account the severity of the violation).

  - **Nondiscriminatory**—The enforcement approach is applied equally across all policy violations (e.g., an AUP is not only enforced for certain types of customers or in certain geographies).

These principles may be formalized into Trust and Safety or other frameworks that govern how enforcement decisions are made or policies are developed. For an example of how these principles can be applied in practice, see the Oversight Board's case decisions, which analyze whether Meta's content policies and their enforcement align with these principles in specific cases.

- **Policy development and enforcement should be informed by stakeholder engagement.** Recall that stakeholder engagement is an important part of a human rights-based approach (see Guide 5: Conducting Stakeholder Engagement). Rather than being created in a vacuum, policies should be informed by the perspectives of affected stakeholders and by relevant experts who understand the context. These stakeholders can surface potential unintended impacts of a policy or unidentified impacts that it should address. It is important to emphasize the risks and needs of vulnerable groups (see Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI). It is therefore also important to ensure engagement includes representatives of vulnerable groups, especially for such a rapidly evolving technology as genAI, where impacts may be cumulative and evolve over time.
- **Policy development should address the most severe impacts first.** The UNGPs state that where it is necessary to prioritize actions to address adverse human rights impacts, companies should first seek to prevent and mitigate those that are most severe.[1] This means prioritizing developing policies for products, services, use cases, or other situations that have the highest risk of harm to people and society. The notion of prioritization is especially important for policy and enforcement coverage of genAI products that produce new content (such as chatbots and image- or video-generation products), given the sheer volume of content that can be produced.
- **Policies and enforcement should honor the principles of internationally recognized human rights when faced with conflicting requirements.** The UNGPs recognize that in some cases, domestic context, such as restrictive local laws, may make it impossible for companies to fully meet their responsibility to respect human rights. In these cases, companies are expected to respect human rights to the greatest extent possible given the circumstances.[2] To address this risk, companies can set out their approach to local legal requirements that conflict with human rights norms, and refer to them when faced with conflicting requirements in practice (e.g., overbroad government requests for user data).
- **Policies enforced in conflict-affected areas should receive enhanced attention.** The risk of severe adverse human rights impacts is heightened in places experiencing armed conflict, contentious elections, or societal upheaval, and operations or business relationships may increase the risk of an organization being complicit in gross human rights abuses

1 See Principle 24 of the UNGPs.
2 See Principle 23 of the UNGPs.

committed by other actors.[3] Policies that may apply or be enforced in conflict contexts should therefore address these risks, and be informed by enhanced due diligence that examines the organization, product, or service's potential impact on the conflict (See Guide 3: Human Rights Based-Approach to Impact Assessment).

- **Reporting and appeals channels should be established to help identify policy violations and correct enforcement errors.** A channel to report potential violations of policies (e.g., an employee grievance mechanism to report violations of a data access policy, or a product misuse reporting channel) are important for enabling an organization to enforce its policies. Similarly, there will inevitably be enforcement mistakes that need to be remediated, and there should therefore be an appeals channel that is appropriate to the nature of the policies and the products they apply to. It is important that these channels adhere to the UNGPs effectiveness criteria for operational-level grievance mechanisms: legitimacy, accessibility, predictability, equitability, and transparency.[4] For more information, see Guide 7: Remedy for Generative AI Related Harms.
- **Policies should be updated and their enforcement reviewed on an ongoing basis.** Just as human rights due diligence is an ongoing process (see Guide 1: Human Rights 101), policies should be updated as needed, and their effectiveness and enforcement should be monitored. Regular policy updates, informed by ongoing human rights due diligence, are important to account for the evolution of human rights impacts over time.

3 See UNGPs Principles 17, 18, 21, and 23, the UNGPs Interpretive Guide, and the OECD Due Diligence Guidance for more information on the role of context in human rights assessments.

4 See Principles 22, 29, and 31 of the UNGPs.

3. Key Resources

For further information on taking a human rights-based approach to policies and enforcement, see the following resources:

- **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 Software as a Service (SaaS) companies should promote responsible use of their products and services, including through policies. It provides examples and analysis of different approaches to AUPs, ToS, 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.

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