A Human Rights-Based Approach to Impact Assessment

Guide 3 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 integrate human rights principles and human rights assessment (HRA) methodology into other forms of impact assessments of generative AI (genAI) and how they can effectively scale impact assessment processes across their organizations. It includes the following sections:

1. **Human Rights Assessment Methodology:** Describes what human rights assessments are, their core methodological components, how they are conducted in practice, and their benefits and limitations for assessing risks to people and society associated with AI.
2. **Integrating Human Rights into Other AI Impact Assessments:** Describes how human rights principles and assessment methodology can be integrated into algorithmic audits/risk/impact assessments, model evaluations, fairness testing, data quality reviews, and red teaming.
3. **Scaling Impact Assessment Processes:** Describes the role of standalone vs. integrated impact assessments, and how AI impact assessment processes can be effectively scaled across organizations.
4. **Integrating Human Rights Assessment Across the Value Chain:** Provides guidance for how each entity of the value chain can integrate human rights into its impact assessment processes.
5. **Key Resources**
6. **Appendix 2:** Detailed List of Human Rights
7. **Appendix 2:** Sample HRA Prioritization Criteria

ACCOMPANYING RESOURCES

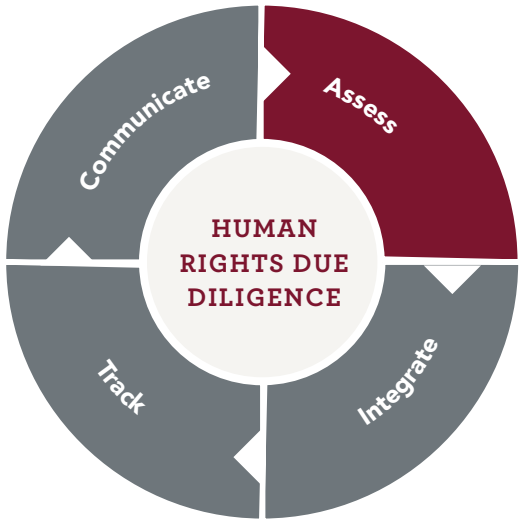
Key Points

- HRAs related to AI products and services identify impacts of the product or service to all relevant internationally recognized human rights.
- Key methodological elements of HRAs include identifying impacts to human rights using all internationally recognized human rights as a reference point; assessing and prioritizing risks based on severity; emphasizing vulnerable and marginalized groups; engaging affected stakeholders; assessing the connectivity between rights; and accounting for context.
- HRAs for assessing AI-related risks have several benefits. They focus on impacts to people; they enable more comprehensive identification of impacts than to other AI assessment approaches; they are adaptable to a variety of contexts; they provide an approach to prioritizing impacts; and they have an established, internationally accepted methodology. HRAs can also help meet emerging regulatory requirements related to AI risk assessment.
- HRAs also have some limitations. While most impacts on people and society are also human rights impacts, there may be relevant risks that fall outside the scope of human rights (e.g., those not related to people). HRAs are also more qualitative than quantitative, and they do not replace technical assessments.
- Human rights principles and assessment methodology can be integrated into other AI impact assessment processes to help identify impacts, assess their severity, and prioritize action to address them. HRAs can also utilize the results of other AI impact assessment processes as inputs.

1. Human Rights Assessment Methodology

Within the Responsible AI field there are a wide variety of impact assessment models that are used to identify and assess risks to people and society. These assessments can be both technical and non-technical or issue based, and at varying levels of depth. They can also be standalone assessments or integrated into existing organizational processes (e.g., as part of product development).

As introduced in Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI, identifying and assessing actual and potential adverse human rights impacts is one step of human rights due diligence. In other words, carrying out some form of human rights impact assessment.



This section describes the methodology and process for conducting human rights assessments. It provides the foundation for Section 2, which describes different AI impact assessment approaches and how they can integrate human rights principles and methodology.

Note: This guide uses the term "impact assessment" broadly to include both risk assessment (i.e., assessment of potential impacts) as well as assessment of actual impacts.

What Are Human Rights Assessments?

Human Rights Assessments (HRAs) identify and assess actual and potential human rights impacts. They utilize a methodology based on guidance in the UN Guiding Principles on Business and Human Rights (UNGPs) to assess the severity of identified impacts and what

appropriate action should be taken to mitigate them.[1] While "human rights assessments" and "human rights *impact* assessments" are very similar, the latter are generally understood to encompass direct engagement with affected stakeholders and tend to go into more depth.

HRAs can have a variety of scopes, from an area of technology (e.g., genAI broadly), a specific product, service or feature, an application area (e.g., genAI in healthcare) or use case, a customer, a geography, or other business decisions such as a merger or acquisition. HRAs related to AI products and services identify the human rights impacts of a product or service using all internationally recognized human rights as a reference point, since a given product or service could impact any of these rights depending on its design, functionality, use case, and the context in which it deployed.

Large companies have been conducting both standalone HRAs and integrated HRAs (i.e., as a part of organizational processes) of AI products and services for the last several years.[2] These assessments are largely qualitative in nature, although relevant quantitative data and insights from other assessment and testing processes (see below) is often considered as an input.

Key methodological elements of human rights assessments

The UNGPs, accompanying interpretive guide, and OECD Due Diligence Guidance establish several key elements that are important for identifying and assessing human rights impacts. Each is described below, along with what they mean in the context of genAI.

- **Identifying human rights impacts of a product or service using all internationally recognized human rights as a reference point:** A human rights assessment involves identifying the ways in which human rights could be impacted by a genAI system. To do this, the UNGPs state that companies should consider all human rights in the international bill of human rights (the UDHR, ICCPR, and ICESCR) and the ILO Core Conventions, while considering other rights areas when relevant to the scope of the assessment (e.g., the rights of persons with disabilities, children's rights).[3] As described in Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI, these instruments encompass a broad and diverse range of rights, including civil and political rights such as the right to freedom of association and to privacy, as well as economic and social rights such as the right to science and culture.
- **Assessing and prioritizing impacts based on severity:** A key component of HRAs is that impacts are assessed and prioritized based on their impact to people, not to the organization or business. Human rights impacts are assessed based on the severity of the impact and the likelihood or frequency of its occurrence. It is often necessary to prioritize actions to address human rights impacts when it is not possible to do everything at once. This is

1 Principle 14 of the UNGPs states that "severity of impacts will be judged by their scale, scope and irremediable character." These three factors are defined in the UNGPs interpretive guide as the gravity of the impact (scale), the number of people affected (scope), and the possibility of restoring those affected to the same or equivalent situation before the impact (remediability).

2 BSR and others have conducted many standalone HRAs for companies related to AI. However, at the time of writing, none have been fully published. An executive summary of one such assessment BSR conducted in 2019 for Google on a celebrity recognition tool can be found here. To get an idea of what standalone HRAs can look like, see other fully published HRAs related to tech products and services, such as this HRA of end-to-end encryption for Meta, this HRA of Twitch, this HRA of the Tech Coalition's Lantern Program, this Ericsson HRA of 5G, and this HRA of Wikimedia Free Knowledge Projects.

3 See Principle 12 of the UNGPs.

often the case with genAI-related impacts given limited resources and the broad range of human rights impacts that may be associated with a genAI system (See BSR's Human Rights Assessment of the Generative AI Value Chain for more detail). When prioritization is needed, the UNGPs state that companies should first address those that are most severe and likely or where a delayed response would make them irremediable.[4]

THE CONCEPT OF CUMULATIVE IMPACTS

Cumulative human rights impacts are successive, incremental, or combined impacts that occur over time.[5] A given human rights impact may not on its own be especially severe, but as impacts continue to occur over time the severity can significantly increase. Cumulative impacts can result from the actions of one entity over time or the combined actions of multiple entities.

To effectively identify and assess impacts to people and society associated with genAI, it is important for Responsible AI practitioners to think ahead about potential cumulative impacts that stem from their own products and services, as well as the interaction of their products and services with the broader technology ecosystem. For example:

- If an image generation tool reproduces harmful gender stereotypes in one output, the adverse impact is likely to be minor. However, if the image generation tool does this consistently over time, that kind of content is likely to be repeatedly viewed by more people and therefore have broader reaching, incremental impacts over time that become more severe, such as contributing to a rise in gender-based violence.
- Individual genAI systems may have safety features and limitations designed to prevent misuse and abuse. However, if those systems may be used in combination with other tools, they could be misused or abused in unanticipated ways that bypass those risk mitigations. For instance, a user could bypass a safety filter against producing a malicious program by asking a genAI chatbot for segments of code that, individually, would be harmless. The user could then input those segments into a coding platform to assemble a malicious program.

- **Emphasis on vulnerable and marginalized groups:** A key aspect of the corporate responsibility to respect human rights according to the UNGPs is an emphasis on the rights and needs of marginalized or vulnerable groups. This is because vulnerable groups generally

4 See Principle 14 of the UNGPs and the UNGPs Interpretive Guide for more information about assessing and prioritizing human rights risks / impacts.

5 For more information about cumulative impacts, see DIHR's HRIA Toolkit.

face heightened risks, or different risks, compared to others, and are less likely to have their needs represented in decision-making processes. For example, many of the discrimination risks associated with genAI systems (e.g., perpetuating harmful stereotypes) primarily impact vulnerable or marginalized groups. Therefore, when carrying out a human rights assessment, it is important to pay special attention to identifying risks to vulnerable groups and consider how those from marginalized or vulnerable groups may experience different or disproportionate impacts.[6]

HOW TO IDENTIFY VULNERABLE OR MARGINALIZED GROUPS

Vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. Vulnerability can change across geographies, and in relationship to different genAI products and applications.

BSR identifies vulnerable groups based on four dimensions:

1. FORMAL DISCRIMINATION—laws or policies, and/or their application, that favor one group over another

2. SOCIETAL DISCRIMINATION—cultural or social practices that marginalize some and favor others

3. PRACTICAL DISCRIMINATION—marginalization due to life circumstances, such as poverty or disability

4. HIDDEN GROUPS—people who might need to remain hidden and consequently may not speak up for their rights

- **Importance of engaging affected stakeholders:** The UNGPs emphasize that effective human rights due diligence requires meaningful engagement with affected stakeholders (aka "rightsholders") or those who can act as representatives of affected stakeholders, such as independent experts, human rights defenders, or others from civil society. Therefore, when possible, HRAs should involve stakeholder engagement in the form of one-on-one or group consultations, as well as make use of insights gathered from previous stakeholder engagement.[7] See Guide 5: Conducting Stakeholder Engagement for more information about how to do this in practice.
- **Considering interconnectivity between rights:** As discussed in Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI, all human rights are indivisible, interdependent, and interrelated, and therefore the improvement of one right facilitates

6 See the "general principles" section of the UNGPs and the UNGPs Interpretive Guide for more information.

7 See UNGPs principles 16, 18, 20, and the UNGPs Interpretive Guide for more information, and the definition of "meaningful stakeholder engagement" in the Annex of the OECD Due Diligence Guidance.

advancement of the others; while the deprivation of one right adversely affects others. A common example of this is privacy. If a genAI chatbot exposes personal data in an output (a privacy impact), for instance, the personal data could be used to dox someone who then may be physically attacked (an impact to their right to life, liberty, and security). When conducting an HRA, it is important to consider how risks could impact multiple rights or have cascading impacts.

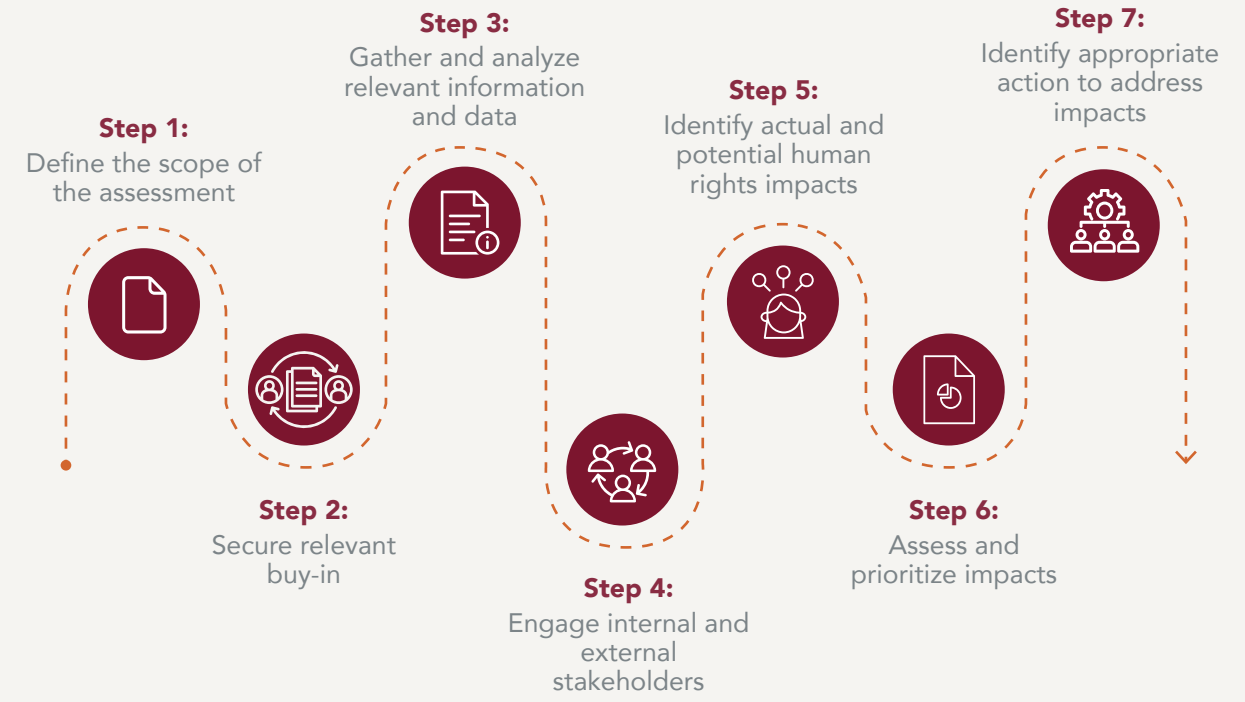
- **Accounting for context:** The context in which a company operates or a genAI tool is deployed (e.g., the industry, the geographic context) can significantly influence the nature and severity of human rights impacts. On the industry side, for example, genAI deployed in healthcare settings or the criminal legal system can involve especially severe impacts due to the life-nature of services and decisions in those contexts. On the geographic side, the often Western-centric nature of genAI training data and known challenges related to performance across languages can result in impacts of varying severity, depending on the geographic context in which it is deployed. Therefore, context should always be considered when identifying and addressing human rights impacts.[8]
- **Conflict-affected areas:** The risk of severe adverse human rights impacts is especially heightened in places experiencing armed conflict, contentious elections, or societal upheaval. In conflict contexts it is therefore important to carry out enhanced and conflict-sensitive human rights due diligence to identify both human rights impacts and conflict-related impacts. More information and a toolkit for conducting conflict-sensitive human rights due diligence in the technology sector can be found here.

How to Conduct a Human Rights Assessment

There is no "right" way to conduct a standalone HRA. As long as the assessment includes the methodological elements outlined in the previous section, it is a human rights assessment. However, below is a description of the basic process that BSR and other organizations that regularly conduct HRAs follow. Understanding this process can help Responsible AI practitioners think about the relevance of standalone HRAs in their work, as well as how aspects of HRA methodology can be integrated into other types of AI impact assessments. Resources that describe this process in more depth can be found at the end of this section.

8 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.

Human Rights Assessment Process



Step 1: Define the scope of the assessment

Define what is being evaluated. Some scopes relevant for genAI include a research project or publication decision, a model, a product or feature, a use case, a category of technology (e.g., image generation), a market or geography, a business partner, or a sale.

Step 2: Secure relevant buy-in

Secure buy-in among relevant teams and individuals who will need to participate in the assessment or utilize its findings. For example, ensure they know what an HRA is, the scope and the purpose of the HRA, and how they will be involved (e.g., in interviews, to provide feedback on drafts).

Step 3: Gather and analyze relevant information and data

Carry out the background research required for the assessment by reviewing any relevant internal and external sources and any important data. This might include academic research, civil society reports, media reports, government disclosures, information about how specific products or features function, where they are or will be available, how they are intended to be used, and the results of any technical evaluations or any related impact assessment process.

Step 4: Engage internal and external stakeholders

Conduct interviews with relevant internal stakeholders, utilizing the background research to create questions and avenues for inquiry. If feasible and appropriate, engage with relevant external stakeholders such as subject matter experts, civil society groups, or others who can speak to the topic and represent the views of potentially affected stakeholders. Direct engagement with affected stakeholders (e.g., users) is challenging but may be feasible via existing user engagement mechanisms, such as UX research. External partners, such as consultants, can help facilitate stakeholder engagement and may be especially important for vulnerable user groups, such as children, where informed ethical engagement is especially important. Guidance on external stakeholder engagement can be found in Guide 5: Conducting Stakeholder Engagement.

Step 5: Identify actual and potential human rights impacts

Utilize everything learned via research and consultation to identify how all internationally recognized human rights could be impacted. To do this, it is helpful to start with a "long list of human rights," like in the table below, as a basis for brainstorming. Walk through each right, one by one, and identify specific impacts. See, for example, the following impact statements from the Human Rights Assessment of the Generative AI Value Chain, which describe how genAI can impact different human rights:

- **Right to nondiscrimination:** GenAI model outputs may contain toxic content that discriminates against people based on attributes such as gender, ethnicity, or nationality.
- **Right to privacy:** User data from the training dataset may be leaked through model outputs.

Because the HRA of the genAI value chain is high-level, so are the impact statements. The narrower the scope of the assessment (e.g., a specific product, a specific use case), the more specific the impact statements should be.

If the assessment is backward-looking (i.e., examining something that has already happened), this step will involve identifying actual impacts found to have occurred. If the assessment is forward looking, this step will involve considering both actual and potential impacts (i.e., risks). Not every single human right will be impacted in all cases, so the resulting list should be narrowed to only those rights that are relevant to the scope of the HRA.

Summary List of Internationally Recognized Human Rights

The following list of human rights is a summarized version of the rights listed in all the international human rights instruments listed above. It includes those most relevant for AI/technology more broadly. Some rights have been combined for simplicity.

Civil and Political Rights

- Right to life, liberty, and security (including both physical and psychological security)
- Freedom from torture, inhuman or degrading treatment
- Freedom from slavery
- Freedom from arbitrary arrest or detention
- Right to a fair trial and the presumption of innocence
- Right to equality and nondiscrimination
- Freedom of movement
- Right to seek asylum
- Right to privacy
- Freedom of expression and opinion
- Freedom of thought
- Freedom of religion
- Access to information
- Freedom of assembly and association
- Right to marry and have a family
- Right to personal property
- Right to political participation (e.g., to vote)

Economic, Social, and Cultural Rights

- Labor rights (e.g., just working conditions, fair wages, right to form unions)
- Right to social security
- Right to an adequate standard of living (e.g., adequate food, housing)
- Right to health
- Right to education
- Right to participate in cultural life
- Right to benefit from scientific advancement
- Right to a healthy environment
- Right to internet access

Thematic Rights

- Children's rights (e.g., to free expression, access to age-appropriate information, freedom from exploitation)
- Disability rights (e.g., accessibility)
- Indigenous people's rights

Step 6: Assess and prioritize impacts

Once all impacts have been identified, the next step is to "assess" them. Human rights impacts are assessed utilizing the following criteria from the UNGPs and OECD Due Diligence Guidelines:

- The **severity** of the impacts, which is comprised of three factors:
  - **Scope**—the number of people who could be impacted
  - **Scale**—the seriousness or gravity of the impacts for those affected
  - **Remediability**—the extent to which the harm could be remediated or "made good"
- The **likelihood** of the impacts occurring or the frequency with which they occur
- These criteria together are known as "salience." Severity (scope, scale, remediability) + likelihood = salience.

In in-depth HRAs, each of these criteria are often broken down into a scale appropriate for the scope of the assessment (e.g., very small, small, medium, large). (See the sample HRA criteria in the appendix.) Each identified impact is evaluated against these criteria, sometimes with the aid of a spreadsheet. This helps identify the impacts that are most "salient" (i.e., most severe and likely), which is important for prioritizing action to address them.

For example, a biotech company conducting an HRA on its use of genAI for drug discovery might find that it could affect a wide range of human rights, but that its potential impacts on the rights to health, life, science, and access to information are especially severe. It would then prioritize taking action to address those impacts.

While quantitative scoring is common and can be helpful (e.g., the sample HRA criteria provided in the appendix can be signed a numerical score), it is important to note that severity and likelihood cannot always be precisely or quantifiably assessed, especially for high-level or broadly scoped HRAs. For example, the severity of a given impact of a widely available, general purpose genAI chatbot will vary significantly depending on the context in which it occurs. In these cases, it is appropriate to consider severity and likelihood at a high level. Quantitative data to support assessments should be used when relevant, but should not constrain the professional judgment of the analysis.

Step 7: Identify appropriate action to address impacts

After identifying and prioritizing impacts, the next step is to identify actions to address them. In HRAs this should be informed by two factors underscored in the UNGPs:

- **Attribution**—how the entity is involved in the impact
- **Leverage**—the leverage the entity has to address the impact

This step is further described in Guide 4: A Human Rights-Based Approach to Risk Mitigation.

A standalone HRA will typically include analysis of these factors for the identified impacts, as well as a list of recommendations for addressing the impacts. Some HRAs will also include an assessment of how each impact is currently being managed and make recommendations to fill any gaps.

The Benefits and Limitations of HRAs for Assessing AI-Related Impacts

HRAs provide several benefits for assessing AI-related impacts, including for genAI:

› A focus on impacts to people

By focusing on how different rights could be impacted, HRAs enable an identification of risk that prioritizes impacts to people, as opposed to impacts to the organization. Engagement with affected stakeholders and emphasizing vulnerable groups also helps ensure a focus on impacts to people rather than more nebulous high-level impacts.

› Comprehensiveness of impact identification

Using all internationally recognized human rights as a reference point for identifying impacts on people helps lead to comprehensive identification of impact. AI impact assessment approaches that rely only on pre-established risk taxonomies, already documented risks, or no starting point at all for identifying risks to people are more susceptible to inherent biases of the assessors (e.g., geographic or cultural context, professional background) and more likely to fail to identify relevant impacts.

› Adaptability to a variety of contexts

In the decades of conducting HRAs since the establishment of the UNGPs and OECD guidelines, the human rights framework has proven to be adaptable enough to assess impacts arising from a wide range of products, services, and business operations. It can be utilized for high-level, broad impacts assessment or as a basis for identifying the precise pathways by which a right can be impacted, which is particularly helpful when assessing impacts for a defined product or service with a clear use case.

› An approach to prioritizing impacts

One common challenge Responsible AI practitioners have raised with BSR is a feeling that there are too many potential impacts to address and they don't know where to start. The UNGPs provides clear guidance that impacts should be prioritized based on their severity and likelihood.

› An established, internationally accepted methodology

International human rights, the UNGPs, and the OECD Guidelines constitute an established, internationally accepted framework. As Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI detailed, global companies have been conducting HRAs as part of their human rights due diligence responsibilities for decades, which has enabled the methodology and approaches to mature significantly over time. With the integration of HRDD requirements into regulation around the world, the human rights-based approach is only becoming more established.

› An approach that helps satisfy emerging regulatory requirements

Guide 1: Fundamentals of a Human Rights-Based Approach to Generative AI discussed how the principles and approaches from the UNGPs and OECD Guidelines are being incorporated into regulation, including two pieces of regulation related to AI risk: the EU Digital Services Act and the EU AI Act. Both regulations feature requirements to conduct risks assessments related to AI products and services, and anchor those requirements in human rights assessment methodology. The Digital Services Act requires companies in scope to "identify, analyze, and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems," including "any actual or foreseeable negative effects for the exercise of fundamental rights."[9] The EU AI Act requires in-scope deployers of AI in a set of defined "high risk AI categories" to carry out "fundamental human rights impact assessments" prior to deployment.[10] Taking a human rights-based approach to impact assessment can therefore help organizations in scope comply with the risk assessment provisions of those regulations.

However, HRAs do have a number of important limitations in the context of identifying and addressing AI impacts, including for genAI.

› HRAs may not cover all relevant impacts

While the vast majority of impacts on people and society are also human rights impacts, there may be relevant impacts that fall outside the scope of human rights. For example, meta-level issues about the broader role of genAI in society, or impacts that are not related to people. Additionally, because the human rights framework focuses primarily on the rights of individuals, it may not enable sufficient assessment of higher-level societal impacts.

› HRAs are more qualitative than quantitative

As discussed above, assessing human rights impacts is more of an art than a science because they are not always easy to quantify (e.g., calculating the probability of a risk can be impossible). HRAs are therefore not always well suited to contexts when risk quantification is necessary.

› HRAs are not technical assessments

While HRAs can and should be informed by the results of technical assessments (e.g. fairness testing or model evaluations), they do not replace the need for and importance of these forms of assessment. However, human rights can be integrated into these types of assessment, which is explored in the next section.

These limitations do not mean that HRAs are not relevant or valuable for Responsible AI practitioners, but rather that they should not be the only approach to identifying, assessing, and addressing AI impacts.

9 See Article 34 and Recital 47 for further information on the human rights assessment related requirements in the EU Digital Services Act.

10 See Article 27 of the EU AI Act for further information on this requirement.

2. Integrating Human Rights Into Other AI Impact Assessments

While HRAs can be conducted as standalone assessments, human rights principles and assessment methodology can also be usefully integrated into other types of AI impact assessments to help identify impacts, assess their severity, and prioritize action to address them.[11] This section describes common types of AI impact assessments, where in the value chain they are conducted, and provides guidance for effective integration of a HRBA into each. They include:

- Algorithmic impact assessments / algorithmic audits
- Model/application evaluations
- Fairness testing
- Data quality reviews
- Red teaming

SUMMARY: INTEGRATING HUMAN RIGHTS INTO OTHER AI IMPACT ASSESSMENTS

Assessment Type	Description	How to Integrate Human Rights
Algorithmic Impact Assessments / Algorithmic Audits	Systematic examination of the algorithms and data used in an AI system to assess their fairness, accountability, transparency, and ethical implications.	Utilize the list of internationally recognized human rights (see the appendix) as a foundation for brainstorming to help identify impacts or create a risk/harm taxonomy. Consider severity when assessing impacts.

11 Much of the description of assessment approaches and how they are used comes from the UN B-Tech Project Paper "Responsible AI and Human Rights: An Overview of Company Practice," which BSR helped draft.

Model/Application Evaluations	Empirical assessments of an AI system's performance or impact on people and society.	Utilize human rights as a foundation for identifying impacts/harms to evaluate.
Fairness Testing	Assessment of whether an AI system exhibits biases or discrimination against certain groups of individuals based on protected characteristics such as race, gender, ethnicity, or age. Often includes model/application evaluation.	Utilize the vulnerable groups framework to help identify groups for the basis of testing. Consider how additional human rights may be impacted as a result of identified fairness issues.
Data Quality Reviews	Examination of the data used to train AI models to look for issues such as incorrect labels, representativeness, accuracy, and bias, that may lead to inaccurate or problematic outputs.	Consider how different data quality issues could lead to human rights impacts, and consider the severity of those impacts to help prioritize corrective actions / mitigation of related impact.
Red Teaming	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.	Identify pathways to human rights impacts as part of the red-teaming process. Include red teamers with a background suited to identifying risks to people, as well as people representative of, or familiar with, risks and needs of vulnerable groups.

Note: all of the AI assessments described below can also be used as inputs to human rights assessments. For example, the red-teaming may identify adversarial techniques that are especially effective on a genAI system, such as jailbreaking prompts to produce biased content in a certain language. This information can then inform the identification of human rights impacts (in this case freedom from discrimination).

A. Algorithmic Audits / Impact Assessments

Description

- Algorithmic audits and/or algorithmic impact assessments (AIAs) involve the **systematic examination of the algorithms and data used in an AI system to assess their fairness, accountability, transparency, and ethical implications.**
- The goal is to **identify biases, discriminatory patterns, and potential harms** in the design, development, and deployment of AI systems. Algorithmic audits and impact assessments have emerged in the Responsible AI field over the past several years.
- There is currently no universally accepted definition or standardized methodology for either, and despite audits and assessments being fundamentally different processes, the terms are often used interchangeably. See the call out box on audits vs. assessments for more information.

How they are implemented

- **Algorithmic audits typically involve quantitative statistical analysis of specific issues.** For example, see Twitter's audit of its image cropping algorithm. Both Meta and Open AI conducted and documented algorithmic audit activities of their genAI models, although they were not necessarily referred to as such. Algorithmic audits often involve model evaluations (see below).
- **AIAs are designed to assess possible social impacts of AI systems,** and a variety of largely qualitative tools and methodologies have been proposed by civil society and government entities. Although companies tend not to use the term, many of them conduct AIAs in practice. This is typically done by assessing products or services against their AI principles, or against a predefined taxonomy of risk/harm/impacts. New taxonomies have been proposed to account for the particular impacts of genAI systems. For example, Open AI identified "harms of representation, allocation, and quality of service" in its assessment of GPT-4.

Value chain relevance

Algorithmic audits / impact assessments can be conducted by the following parts of the genAI value gain:

- Foundation model developers: e.g., on foundation models
- Downstream developers: e.g., on a genAI tool they develop
- Deployers: e.g., on the use of a genAI tool for a specific use case and deployment context

How to integrate human rights

- Of all the common assessment approaches, **AIAs bear the closest resemblance to human rights assessments** in that they are largely qualitative in nature and identify and assess impacts to people and society. Although they may not utilize human rights terminology

or call out human rights specifically, by identifying and assessing impacts to people and society they often de facto identify human rights impacts.

› To integrate human rights, algorithmic audits / impact assessments **can utilize the list of internationally recognized human rights (see the appendix) as a foundation for brainstorming to help identify impacts or create a risk/harm taxonomy**. They need not necessarily use human rights terminology, and other types of impacts can be included as well.

› Algorithmic audits / impact assessments **can also consider severity when assessing impacts**. While scope, scale, and remediability are robust factors for assessing the severity of impacts on people, they may be challenging to determine, especially for high-level assessments or in the early stages of product development. Considering a simpler notion of severity—how serious or grave the impacts would be—may be appropriate in these cases.

AUDITS VS. ASSESSMENTS

Assessments and audits are not the same thing. The term "audit" has established definitions in the fields of accounting and cybersecurity. However, the Responsible AI field often uses the terms interchangeably. With the onset of both audit and assessment requirements as part of regulation (see, for example, the EU Digital Services Act, which requires assessments and audits of those assessments), it is important to understand the difference and take care with terminology.

- **An audit** is an examination of something to verify its accuracy or compliance with clearly defined standards, criteria, or controls. They sometimes have legal bearing and are often performed by qualified/certified auditors. They are most often backward-looking/historical. They may include recommendations for how to reach compliance with the standard.

  In the context of Responsible AI, the term audit is correctly used for evaluations of a given system against predefined technical standards or metrics.

- **Assessments** are broader, more holistic and flexible than audits. They provide analysis on a particular topic. They are typically advisory in nature, and generally do not have any legal bearing. They are often, but not always, forward-looking. They may include a wide variety of recommendations related to scope and findings of the assessment.

  In the context of Responsible AI, the term assessment is correctly used for any form of impact assessment or evaluation that does not entail evaluating an AI system against a clearly defined set of standards.

B. Model/Application Evaluations

Description

- Model/application evaluations are **empirical assessments of an AI system's performance or impact on people and society**. Most current model evaluations focus on assessing capabilities, such as whether a model can read Amharic or play StarCraft. Some evaluations also assess model performance on issues that impact people and society.

How they are implemented

- Model evaluations that focus on identifying and addressing impacts to people and society are typically done by defining a harm (e.g., gender bias), operationalizing it into a metric (e.g., "ratio of misogynistic outputs"), obtaining data, and then judging the outcome. However, these evaluations have been critiqued as assessing models in isolation without adequately considering interaction with real users in the real world.
- To address that gap, some academics and practitioners have begun to design "sociotechnical" evaluations that can assess a model's impacts on people and society more broadly.[12] Sociotechnical model evaluations are a nascent field, and there are few current published and tested evaluations of impacts to people and society. Challenges involved with designing such evaluations include the subjectivity of defining harm to be evaluated, obtaining valid measurements of those harms, and the complexity of rigorously evaluating the real world.

Value chain relevance

Model evaluations can be performed by the following parts of the genAI value gain:

- Foundation model developers: e.g., on foundation models
- Downstream developers: e.g., on a genAI tool they develop
- Deployers: e.g., on a genAI tool fine-tuned for their specific use case and deployment context

How to integrate human rights

- Model evaluations that seek to empirically assess model performance factors that can adversely impact people and society de facto identify and address human rights impacts. One example of this might be evaluations for toxic or biased language, which shed light on impacts to the right to freedom from discrimination. Model evaluation results can often also be a helpful input into human rights assessments.
- Model evaluations **can utilize human rights as a foundation for identifying impacts or harms to evaluate**. The human rights framework may be especially useful for defining harms for sociotechnical model evaluations, since human rights are intrinsic indicators of

12 Data and Society has described a sociotechnical perspective as viewing society and technology together as one coherent ecosystem, recognizing that it is not possible to understand the "social" without the "technical," and visa versa. For more information see https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/.

social and individual impact. While human rights terminology need not be used, grounding evaluations in human rights may help address concerns about the legitimacy or subjectivity of the harms chosen for evaluation.

C. Fairness Testing

Description

- The process of assessing whether an AI system exhibits biases or discrimination against certain groups of individuals based on protected characteristics such as race, gender, ethnicity, or age. The goal is to identify, measure, and mitigate potential biases in AI systems to ensure that they treat all individuals fairly and equitably. Fairness testing typically evolves model/application evaluation.

How it is implemented

- Fairness testing is an increasingly established best practice in the Responsible AI field. It involves examining the training dataset and probing the AI model to see whether it produces unfair outputs that exacerbate existing societal biases.
- Fairness testing is typically both a qualitative and quantitative exercise, and involves technical interaction with both the dataset and the model (e.g., see Google's developer guide to fairness testing). For a genAI product, for example, fairness testing might examine whether a prompt to an image-generation tool for images of doctors returns images that are predominantly of men, and identify how to address that particular issue.

Value chain relevance

Fairness testing can be performed by the following parts of the genAI value gain:

- Foundation model developers: e.g., on foundation models
- Downstream developers: e.g., on a genAI tool they develop
- Deployers: e.g., on a genAI tool fine-tuned for their specific use case and deployment context

How to integrate human rights

- Fairness testing effectively seeks to identify and address impacts on the specific human rights of equality and nondiscrimination.
- Fairness testing can further integrate human rights by **utilizing the vulnerable groups framework described above to help identify groups for the basis of testing**.
- Nondiscrimination is often the primary impact of unfair biases in AI systems; however, other rights can be impacted as well. Fairness testing **can consider how additional human rights may be affected as a result of identified fairness issues**. Considering the range and severity of human rights impacts can also help address tensions between the different definitions and approaches to fairness and inform decisions about what to emphasize.

D. Data Quality Reviews

Description

› Data quality reviews involve the examination of the data used to train AI models to look for issues such as incorrect labels, representativeness, accuracy, and bias, that may lead to inaccurate or problematic outputs.

How it is implemented

Data quality reviews often have three different focus areas:

› **Dataset Curation:** Involves filtering out problematic data (such as toxic speech, pornographic content, etc.), and ensuring the relevance, representativeness, and comprehensiveness of the data.

› **Data Provenance:** Involves the inclusion of metadata in training datasets to establish data provenance for the purpose of identifying the origin/source and processing of all data in the dataset. Establishing data provenance may help developers identify the source of issues in model outputs.

› **Data Documentation:** Involves ensuring all data in the dataset is documented (e.g., what the data contains, how it can be used, limitations, etc.). Dataset documentation helps developers understand appropriate uses for a given dataset.

Value chain relevance

› Data quality reviews are primarily performed by data suppliers in the genAI value chain; however, they may be performed in collaboration with other value chain actors when relevant. For example, healthcare providers seeking to deploy a genAI tool in a clinical setting may be involved in dataset curation.

How to integrate human rights

› In genAI systems, poor quality data hinders the model's ability to generate accurate, meaningful, and representative outputs. These kinds of assessments are therefore key for identifying and addressing some of the root causes of human rights impacts that can stem from inaccurate or problematic genAI outputs.

› To integrate human rights, data quality reviews **can consider how different data quality issues could lead to human rights impacts, and consider the severity of those impacts to help prioritize corrective actions / mitigation of related impacts**. For example, identifying what human rights impacts might result from data quality issues in a fine-tuning dataset used for a genAI chatbot designed to provide health advice. This understanding of potential impacts can then be used to develop data labeling / annotation guidance designed to mitigate those impacts.

E. Red Teaming

Description

› 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 or expected user behavior and adversarial attacks and scenarios. This knowledge can then be used to strategize to address identified gaps.

How it is implemented

› Red teaming usually involves a group of experts from a variety of backgrounds who adversarially test an AI system to identify flaws and vulnerabilities (e.g., ways in which it could produce undesirable outputs, how safety measures can be bypassed, vectors for cybersecurity risks, etc.).

› Red teaming can be both technical (e.g., technical jailbreaking attempts) and nontechnical (e.g., adversarial prompting), and has gained prominence as a particularly helpful approach to identifying and addressing impacts associated with genAI systems.

› Both OpenAI and Meta have written publicly about their red teaming approaches and results for their respective genAI models. Red teaming for genAI thus far has involved primarily technical experts from Western contexts, and has only in some cases involved participants with human rights expertise or explicit consideration of human rights impacts (e.g., see Microsoft's "harms modeling" approach).

Value chain relevance

Red teaming can be performed by the following parts of the genAI value gain:

› Foundation model developers: e.g., on foundation models

› Downstream developers: e.g., on a genAI tool they develop

› Deployers: e.g., on a genAI tool fine-tuned for their specific use case and deployment context

How to integrate human rights

› Some of the vulnerabilities identified by red teaming could lead to adverse impacts on human rights; therefore red teaming is a process that can de facto identify human rights impacts.

› **Pathways to human rights impacts can be identified as part of the red teaming process.** For example, developers could integrate a human rights-based approach to red teaming processes, such as those that involve introducing specific prompts to a model in order to understand where guardrails are required to prevent harmful outputs. Prompts that target human rights topics could be used to ensure the model does not return content that could create human rights impacts. Prompts related to inquiries about the asylum

process, for instance, may help identify whether models could potentially offer users false or misleading information that may-impact their right to seek asylum. The results of these tests can then inform the implementation of human rights-respecting safeguards.

› Properly integrating a human rights-based approach to red teaming requires **including red teamers with a background suited to identifying impacts on people, as well as people representative of, or familiar with, the risks and needs of vulnerable groups.** The appropriate makeup of a given red team will look different for a general purpose product (e.g., red teaming of an LLM) vs. a product with a specific use case.

3. Scaling Impact Assessments

One of the challenges for practitioners in operationalizing Responsible AI is effectively scaling impact assessment processes across their organization. This includes when to pursue standalone assessments vs. integrated processes, and ensuring that impact assessments for AI are occurring across all relevant parts of the organization, at the appropriate level of depth, and with the appropriate teams involved.

The role of standalone vs. integrated assessments

An important human rights concept for genAI practitioners is ongoing human rights due diligence. This is the notion that because human rights impacts evolve alongside technologies and use cases, HRDD cannot just be conducted at one moment in time, but rather should be considered an ongoing process.[13] Both the UNGPs and the OECD Due Diligence Guidance recommend companies reassess impacts periodically throughout the lifecycle of a given business activity or relationship, and especially prior to major decisions or changes (e.g., major product launch, market entry, significant policy change) or in anticipation of major changes in operating environments (e.g., significant political shifts).[14]

Standalone assessments are a helpful foundation for identifying and understanding impacts. However, it is important to have a process for identifying new or evolving risks. This is especially important for a technology as fast-evolving as genAI.

Standalone impact assessments can take a variety of forms—from rapid and high level to in-depth and conducted over an extended period of time. In-depth HRAs can be time consuming to conduct because they require detailed research, significant stakeholder consultation, and complex analysis that take time and resources to complete. They are, therefore, often not feasible to conduct for every new product, service, feature, sale, or procurement of a product, and need to be conducted early enough to effectively influence decision-making.

13 See principle 17 of the UNGPs.
14 See principle 18 of the UNGPs and section 2.2 of the OECD Due Diligence Guidance.

Because of this, in-depth impact assessments are most valuable for:

- New or emerging areas of technology
- Significant new products or services
- Major business / organizational decisions (e.g., merger or acquisition, market entry/exit)
- High-risk contexts or decisions (e.g., a product involving high-risk technology, a high-risk use case, or a high-risk customer)

Because standalone assessments cannot be conducted in every case, and because risk assessment needs to be an ongoing process, impact assessment should be integrated into existing organizational processes and should include mechanisms to identify and escalate situations when more in-depth, standalone assessment is needed. This approach is in line with the OECD Due Diligence Guidance, which recommends carrying out iterative and increasingly in-depth assessment based on areas with more significant human rights impacts.[15]

Approaches to scaling impact assessment

In BSR's experience working with a wide variety of technology companies, there is a two-prong approaches for scaling impact assessment across organizations that can be adapted to different contexts:

1. Risk Frameworks: Based on a standalone assessment (e.g., a companywide HRA), identify the areas of the organization, types of products and/or services, use cases, customers, geographies, etc., that have the most significant risks to people and society. Utilize the results of this assessment to create risk frameworks that can identify high-risk situations as they arise in relevant parts of the organization. For example, a sales team within a downstream genAI developer might have a framework that flags high-risk characteristics related to the product and use case, the customer, and the geography. The example below provides an explanation for how these factors are relevant. Risk frameworks can range from a simple list of high-risk characteristics to more detailed assessments of factors with different risk levels (e.g., high, medium, low) and what actions should be taken depending on the factors present in a given situation.

2. Gating and Escalation Processes: The results of the risk framework can then be utilized to identify and escalate higher-risk situations for in-depth assessment. In the sales example above, a gating and escalation process might involve integrating questions designed to flag sales of genAI products that meet any of the high-risk characteristics from the risk framework into a customer due diligence questionnaire. Flagged sales would then be escalated for in-depth impact assessment, which could be carried out by a relevant team (e.g., Responsible AI team). Establishing internal committees to review impact assessments and make go/no-go decisions or require teams to implement certain mitigations can also be helpful to ensure the findings of impact assessments are appropriately integrated.

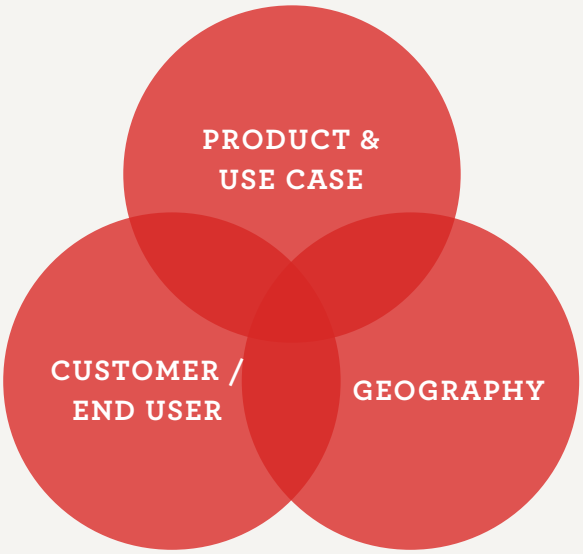
15 See Section 2.2 of the OECD Due Diligence Guidance.

In order for these approaches to be successful, it is important to have clear policies and processes, and ensure relevant teams are aware of them. Teaching relevant teams basic impact assessment skills via applied training is also important for effective adoption and implementation of impact assessment escalation processes. For example, product teams should learn how to proactively identify risks to people that may be associated with a genAI tool they are developing so that they can take action to mitigate those risks during the product development phase. Similarly, sales teams should have a basic understanding of what combinations of products, customers, use cases, and geographies are high risk and why, so that they can proactively identify potentially risky sales that may fall through the cracks of a gating and escalation process.

It is also important to introduce impact assessment processes early to ensure that there is adequate time to identify and implement risk mitigations. For example, it is generally easier to implement risk mitigations earlier in a product development process vs. immediately prior to launch or post-launch. However, while it is possible to identify risks to people during the early phases of product development, some risks do not become clear until the product's functionality is more fleshed out. Therefore, it is important to time impact assessment processes such that there is sufficient information available to effectively identify risks while also ensuring there is adequate time to implement mitigations.

END-USE RISK FRAMEWORK

The following factors interact to contribute to the human rights associated with the use of technology products and services.



The Product and Use Case: Certain products and certain uses of technology come with higher levels of human rights risk. This includes, for example, any technology used for serveillance, to inform high-stakes dicision-making processes, or in safety-critical industries (e.g., health care, transportation).

The Customer / End User: Certain customers also come with higher levels of human rights risk, particularly when they utilize technology to aid in activities that significaintly affect people's lives– for example, government entities involved in law enforcement, the military, intelligence, and the provision of social services. Certain industries can also be high risk, such as healthcare, defense, law enforcement, telecommunications.

The Geography: In countries with a history of human rights violations, weak rule of law, and a high rate of corruption, misuse of products and services may be more likely. Government entitites in these geographies are also more likely to be high risk. However, it is important to note that misuse of products and services can occur in "low-risk" geographies as well.

4. Integrating Human Rights Into Impact Assessment Across The Generative AI Value Chain

This section provides guidance for actors across the genAI value chain seeking to integrate a human rights-based approach into their impact assessment processes and approaches. (See Section 4 of BSR's Human Rights Assessment of the Generative AI Value Chain for a detailed description.) It is important for actors to consider their role in the value chain and the impacts that might flow to the subsequent actor. For example, foundation model developers should assess potential impacts associated with their models being used by downstream developers.

The guidance below serves as a starting point for identifying human rights impacts at each point of the genAI value chain and should not be considered comprehensive. Salient impacts and relevant questions to identify impacts will vary depending on the context, product, use case, etc.

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.

1. Suppliers

Risk Pathways:

Impacts at the genAI supplier level most often stem from processes related to dataset curation. The following are common risk pathways for suppliers that can lead to downstream adverse human rights impacts. Suppliers should familiarize themselves with these risk pathways and how they may emerge as part of their processes and operations. (Note: for

more information on each of these risk pathways, refer to Section 4 of BSR's Human Rights Assessment of the Generative AI Value Chain.)

- **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. Data suppliers may also omit data provenance in their dataset documentation.
- **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.
- **Misalignment between dataset and downstream use:** Some datasets may be excessively broad or contain material irrelevant to a specialized domain.
- **Failure to clean biased data:** Data suppliers may fail to recognize and rectify entrenched biases in their foundational datasets.
- **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 identifying 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 pay, or not pay, data creators.

Example questions to inform impact assessment:

The following questions are examples of human rights considerations that should be integrated into supplier impact assessment processes.

- **Equality and nondiscrimination:** Have datasets been curated with consideration for equitable representation and balance for the intended purpose/use of the product? Have representational needs been defined to be contextually specific to the intended use case or application of the training dataset?
- **Equality and nondiscrimination:** Is there data annotation guidance and have annotators been adequately trained on it? Has data annotation guidance been reviewed to ensure equality and nondiscrimination? For example, in what ways could data labeling guidance contribute to discriminatory representation of particular individuals or communities?
- **Privacy and personal property rights:** Did data subjects provide informed consent for their data to be included in datasets? Are there guidelines for responsible data collection? Are there effective mechanisms for verifying that datasets have been assembled responsibly?
- **Labor rights:** Are data enrichment workers provided fair compensation and working conditions aligned with International Labour Organization standards?

- **Remedy:** Have measures been taken in the dataset curation process to support developers' ability to identify and address the source of model performance issues? Specifically, has data provenance information been included in the dataset and has the dataset been appropriately documented?

2. Foundation Model Developers

Risk Pathways:

Impacts at the foundation model level stem from data procurement, model training and evaluation processes, and model release and disclosure decisions. The following are common risk pathways for foundation model developers that can lead to downstream adverse human rights impacts. Foundation model developers should familiarize themselves with these risk pathways and how they may emerge as part of their processes and operations. (Note: for more information on each of these risks, refer to Section 4 of BSR's Human Rights Assessment of the Generative AI Value Chain.)

- **Insufficient or inappropriate safeguards:** Safeguards implemented at the foundation model level may not be sufficient or appropriate for mitigating against downstream risks. In other cases, safeguards at the foundation model level may limit downstream model performance.
- **Limitations in model training techniques:** Some training techniques that foundation model developers utilize could enable downstream risks.
- **Incomplete or overconfident 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. Evaluations often cannot capture the complex real-world impacts of LLMs.
- **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 architecture, the choice of the loss function, or hyper-parameters. These choices affect model behavior and accordingly may have downstream impacts.
- **Choices pertaining to model release:** Foundation model developers make decisions about the gradient of release for LLMs. Models may be released fully open, open-sourced 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.
- **Insufficient transparency:** Foundation model developers may not release sufficient information about the model that would enable downstream developers to identify and mitigate risks.

Example questions to inform impact assessment:

The following questions are examples of human rights considerations that should be integrated into foundation model impact assessment processes.

- **Life, liberty, and security; equality and nondiscrimination; etc.:** Have technical safeguards been enabled to ensure models do not output content that could be disturbing for the user, including violent, graphic, or obscene content; or content which would constitute hate speech? Have measures been taken to ensure that these safeguards are robust and not easily undone by downstream developers? Have safeguards been implemented with consideration for the type of planned model release (e.g., more conservative restrictions for open source models)?
- **Access to scientific advancements and its benefits:** Have technical safeguards been implemented with consideration for how they may limit downstream model performance (i.e., balancing model safety with model usefulness)? Have appropriate measures been taken to ensure that technical safeguards do not limit model applicability (e.g., in various contexts or domains) or useability (e.g., by particular individuals and communities, including linguistic communities) so that the technology may benefit as many people as possible?
- **Equality and nondiscrimination; freedom of expression and opinion; right to participate in cultural life:** Have model training techniques been assessed for their ability to enable outputs that represent diversity appropriately (e.g., in views, in representation of people and cultures)?
- **Equality and nondiscrimination:** Have models undergone robust evaluations to ensure that they do not directly perpetuate social equity issues and that they cannot be easily leveraged by downstream developers for discriminatory purposes?
- **Equality and nondiscrimination:** Have model architectures been evaluated for ways in which they may enable unintended bias or otherwise harmful outputs?
- **Remedy:** Has there been consideration for the ways that the model release approach (i.e., the gradient from closed to open source) may impact downstream human rights risk? Are decisions about model release made with consideration for how users may be able to access remedy for harms related to genAI systems that leverage foundation models?
- **Remedy:** Is there sufficient transparency about model characteristics to support downstream developers' ability to identify and address issues that arise when integrating, fine-tuning, or building on top of foundation models?

3. Downstream Developers

Risk Pathways:

Impacts at the downstream development level stem from individual use cases, training and evaluation processes, and user interfaces. The following are common risk pathways for downstream developers that can lead to downstream adverse human rights impacts. Downstream developers should familiarize themselves with these risk pathways and how they may emerge as part of their

processes and operations. (Note: for more information on each of these risks, refer to Section 4 of BSR's Human Rights Assessment of the Generative AI Value Chain.)

- **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:** 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. In other cases, technologies built on foundation models may be more likely to impact human rights due to the application domain.
- **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. 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.

Example questions to inform impact assessment:

The following questions are an example of the types and categories of human rights considerations that should be integrated into a HRBA to impact assessment processes related to downstream development.

- **Life, liberty, and security; equality and nondiscrimination; etc.:** Have technical safeguards been enabled to ensure the system does not output content that could be disturbing for the user, including violent, graphic, or obscene content; or content which would constitute hate speech? Have measures been taken to ensure that these safeguards are robust and not easily circumvented by bad actors?
- **Equality and nondiscrimination:** Have systems undergone robust evaluations to ensure that they do not directly perpetuate social equity issues and that they cannot be easily leveraged by deployers or individual bad actors for discriminatory purposes?
- **All rights:** Have proposed use cases undergone any kind of impact prior to development and deployment to understand the ways that the individual use case and/or its application in specific domains may have adverse impacts on people?

- **Remedy:** Have mechanisms that meet the UNGP's effectiveness criteria for operational grievance mechanisms (i.e., legitimate, accessible, predictable, equitable, transparent, rights-compatible, source of learning) been designed to enable deployers and individual users to report issues and/or harms resulting from use of a genAI system? Have internal processes been established for addressing complaints, providing remedy, and mitigating risks?

4. Deployers

Risk Pathways:

Impacts at the system deployment level stem from the deployment context. The following are common risk pathways for genAI deployers that can lead to downstream adverse human rights impacts. Deployers of genAI systems should familiarize themselves with these risk pathways and how they may emerge as part of their processes and operations. (Note: for more information on each of these risks, refer to Section 4 of BSR's Human Rights Assessment of the Generative AI Value Chain.)

- **Limited genAI literacy and/or an absence of safety culture:** Deployers may fail to establish a deployment safety culture as they are unaware of the risks of 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 implementing the technology in ways or into work streams that are inappropriate.
- **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.
- **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. 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.
- **Ineffective scalable oversight:** Deploying genAI tools at scale will often necessitate scalable oversight mechanisms, such as 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.

Example questions to inform impact assessment:

The following questions are an example of the types and categories of human rights considerations that should be integrated into risk assessment processes related to deployment.

- **All rights:** Which functions across the organization are deploying genAI solutions? For what use cases? Do any of these use cases pose significant risks to people? Have proposed use cases undergone any kind of impact prior to development and deployment to understand the ways that the individual use case and/or its application in specific domains may have adverse impacts on people?
- **Equality and nondiscrimination; public services; health; education; etc.:** Has there been thorough assessment of the types of harms that could result from the integration of genAI systems into workplaces, workflows, processes, and operations? Has there been adequate investment into capacity building and knowledge sharing for individuals operating genAI systems to ensure that systems are used appropriately and safely and do not result in harm specific to the deployment context?
- **Remedy:** Have processes and procedures been put in place to prevent harms related to the use of genAI systems (e.g., human oversight) and to ensure those impacted have access to remedy if or when harms do result from the use of genAI systems?
- **Labor rights:** Has there been thorough assessment of how the integration of genAI systems into workplaces, workflows, processes, and operations may impact workers' well-being, including how impacts on availability of quality work opportunities?

5. Key Resources

Human Rights Assessment Resources: The following resources contain more detailed information about human rights assessments and how to conduct them:

- **UNGPs Interpretive Guide:** Provides guidance on how to interpret the UNGPs, alongside examples.
- **OECD Due Diligence Guidance:** A subcomponent of the OECD guidelines that describes how companies should carry out due diligence. These guidelines are in line with the UNGPs.
- **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 impact assessment.
- **BSR's Approach to Human Rights Assessments:** Describes how BSR conducts HRAs for companies across all industries, including various types and levels of depth. This approach can be used in any context.
- **BSR FAQ on Human Rights Assessment:** An FAQ that seeks to answer common questions about HRAs.
- **The Danish Institute for Human Rights HRIA Guidance and Toolkit:** An in-depth explainer on how to conduct standalone HRIAs that is meant to be applicable to all industries. It includes a variety of tools and illustrative case studies.
- **Assessing AI: Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems:** The Center for Democracy and Technology's report mapping the spectrum of AI assessment approaches, from narrowest to broadest and from least to most independent, to identify which approaches best serve which goals.

Appendix 1: List Of Human Rights

This annex contains detailed descriptions of the following rights that are most relevant to AI.

Summary List of Internationally Recognized Human Rights

The following list of human rights is a summarized version of the rights listed in all the international human rights instruments listed above. It includes those most relevant for AI/technology more broadly. Some rights have been combined for simplicity.

Civil and Political Rights

- Right to life, liberty, and security (including both physical and psychological security)
- Freedom from torture, inhuman or degrading treatment
- Freedom from slavery
- Freedom from arbitrary arrest or detention
- Right to a fair trial and the presumption of innocence
- Right to equality and nondiscrimination
- Freedom of movement
- Right to seek asylum
- Right to privacy
- Freedom of expression and opinion
- Freedom of thought
- Freedom of religion
- Access to information
- Freedom of assembly and association
- Right to marry and have a family
- Right to personal property
- Right to political participation (e.g., to vote)

Economic, Social, and Cultural Rights

- Labor rights (e.g., just working conditions, fair wages, right to form unions)
- Right to social security
- Right to an adequate standard of living (e.g., adequate food, housing)
- Right to health
- Right to education
- Right to participate in cultural life
- Right to benefit from scientific advancement
- Right to a healthy environment
- Right to internet access

Thematic Rights

- Children's rights (e.g., to free expression, access to age-appropriate information, freedom from exploitation)
- Disability rights (e.g., accessibility)
- Indigenous people's rights

The list of human rights has been segmented into three categories:

1. Civil and political rights
2. Social, economic, and cultural rights
3. Thematic rights

Civil and Political Rights

RIGHT	DESCRIPTION
Right to life, liberty, and security (including both physical and psychological security)	Everyone has the right to life, liberty, and security of person. This includes protection from national and individual security, including physical integrity of citizens to external threats, abuse by government officials or other citizens, and state or private surveillance. **International Human Rights Treaties and Conventions** • Article 3 UDHR • Article 6 and 9 of the ICCPR • ILO Occupational Safety and Health Convention (No. 155) • Article 14, CRPD • Article 16, ICRMW

Freedom from torture, inhuman or degrading treatment	No one shall be subjected to torture or to cruel, inhuman, or degrading treatment or punishment. **International Human Rights Treaties and Conventions** • Article 5 UDHR • Article 7 ICCPR • Article 37, CRC • Article 15, CRPD • Article 10, 17, ICRMW • UN Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CTIDTP)
Freedom from slavery	No one shall be held in slavery or servitude; slavery and the slave trade shall be prohibited in all their forms. **International Human Rights Treaties and Conventions** • Article 4 UDHR • Article 8 of the ICCPR • ILO Forced Labour Convention (No. 29) • ILO Abolition of Forced Labour Convention (No. 105) • ILO Worst Forms of Child Labour Convention (No. 182) • ILO Standards on Forced Labour
Freedom from arbitrary arrest or detention	No one is to be deprived of their liberty or exiled from their country without having first been found of criminal offense by a legal statute; the government cannot deprive an individual of their liberty without proper due process of law. Anyone who has been the victim of unlawful arrest or detention shall have an enforceable right to compensation. **International Human Rights Treaties and Conventions** • Article 9 UDHR • Article 9 ICCPR
Right to a fair trial and the presumption of innocence	Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him. Courts must be competent to interpret the law, independent from the government, and able to make an independent judgement. A presumption of innocence should be maintained in a fair and public trial, where a proper defense can be mounted, leaving the obligation to prove each element of offense beyond reasonable doubt to the responsibility of the prosecution. A trial cannot be brought for something that was not a criminal offense in law at the time it was done. **International Human Rights Treaties and Conventions** • Article 10 & 11 UDHR • Article 14, ICCPR • Article 13, CRC

Right to equality and nondiscrimination	All human beings are born free and equal in dignity and rights and everyone is entitled to all rights and freedoms, without any distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or status. All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination and against any incitement to such discrimination. Discrimination occurs when an individual or group is disproportionately impacted compared to others. Any action or treatment must not violate one's freedom from discrimination, regardless of its positive or negative outcomes. **International Human Rights Treaties and Conventions** • Article 1, 2, & 7 UDHR • Article 2, 3, & 26 ICCPR • Article 2 & 3 ICESCR • Article 2 CEDAW • Article 2, 4, & 5 CERD • Article 5 CRPD • Article 17 DRIP • ILO C111
Freedom of movement	Everyone has the right to freedom of movement and residence within the borders of each state, and everyone has the right to leave any country, including his own, and to return to his country. **International Human Rights Treaties and Conventions** • Article 13 UDHR • Article 12 ICCPR • Article 18 CRPD • Article 8, 39 ICRMW
Right to seek asylum	Everyone has the right to seek and to enjoy in other countries asylum from persecution. Persons persecuted by one's own country may be protected by another sovereign authority. **International Human Rights Treaties and Conventions** • Article 14 UDHR • UN Convention Relating to the Status of Refugees (all articles)

Right to privacy	No one shall be subject to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has a right to the protection of the law against such interference or attacks. The right to privacy includes any information shared with authorized parties that cannot be divulged to third parties without the express consent of the individual. A reasonable expectation of privacy and confidentiality must be maintained. Every individual has the right to ascertain whether their personal data is stored in automatic data files and for what purpose. Moreover, every individual should be able to ascertain which bodies control or may control their files and be able to request rectification or elimination in cases where files contain incorrect data or have been collected/processed contrary to provisions of the law. **International Human Rights Treaties and Conventions** • Article 12 UDHR • Article 17 ICCPR • General Comment No. 16 on Article 17 ICCPR • Article 16 CRC • Article 22 CRPD • Article 14 ICRMW
Freedom of expression and opinion	Everyone has the right to freedom of opinion. This right includes the freedom to hold opinions without interference or fear of punishment. Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive, and impart information and ideas through any media and regardless of frontiers. Everyone has the right to communicate their views, however unpopular, without interference or fear of punishment. **International Human Rights Treaties and Conventions** • Article 19 UDHR • Article 19 ICCPR • Article 13 ICRMW
Freedom of thought, religion, or belief	Everyone shall have the right to freedom of thought, conscience, and religion. This right shall include freedom to have or to adopt a religion or belief of his choice, and freedom, either individually or in community with others and in public or private, to manifest his religion or belief in worship, observance, practice, and teaching. Everyone has the right to hold views on any issue without fear of punishment or censure. **International Human Rights Treaties and Conventions** • Article 18 UDHR • Article 18 ICCPR • Article 14 CRC

Access to information	Everyone has the right to seek, receive, and impart information and ideas through any media and regardless of frontiers. **International Human Rights Treaties and Conventions** • Article 19 UDHR • Article 19 ICCPR • Article 13 ICRMW
Freedom of assembly and association	Everyone has the right to gather with other people, in public or private, hold meetings, and organize peaceful demonstrations, including but not limited to societies, trade unions, political groups, and other associations. No one may force another to join any group if he or she does not wish to. **International Human Rights Treaties and Conventions** • Article 20 UDHR • Article 22 ICCPR • Article 15 CRC
Right to marry and have a family	Men and women of full age, without any limitation due to race, nationality, or religion, have the right to marry and to found a family. They are entitled to equal rights as to marriage, during marriage, and at its dissolution. Marriage shall be entered into only with the free and full consent of the intending spouses. The family is the natural and fundamental group unit of society and is entitled to protection by society and the state. **International Human Rights Treaties and Conventions** • Article 16 UDHR • Article 23 ICCPR • Article 23 CRPD

Right to personal property	The human right to intellectual property stems from the protection of moral and material interests resulting from any scientific, literary, or artistic production of which he is the author. This right safeguards the personal link between authors and their work even if the work becomes common property of mankind, and seeks to encourage active contributions to the arts and sciences. It also calls for adequate, material remuneration for authors. This human right does not necessarily coincide with intellectual property rights under national legislation or international agreements, such as protections present in copyright, patent, or IP regimes, and applies only to "moral and material interests" directly generated by the production. International rights recognize everyone's right to own property and the peaceful enjoyment of his possessions. This imposes an obligation not to interfere with, deprive a person of, or subject an individual's possession to control, while also calling for the action toward reasonable steps to protect property. Tangible property rights are understood as physical items—including but not limited to land, buildings, money, electronics, and equipment—that are already in possession or will be acquired through lawful means. **International Human Rights Treaties and Conventions** • Article 17 & 27 UDHR • Article 15 para. 1(c) ICCPR • Article 7 ICESCR • Article 12 CRPD • Article 15 ICRMW
Right to political participation (e.g., to vote)	Everyone has the right to take part in the government of his or her own country, directly or through freely chosen representatives, conduct public affairs, vote, and be elected at genuine periodic elections. Everyone has the right to equal access to public service in his country. **International Human Rights Treaties and Conventions** • Article 21 UDHR • Article 25 ICESCR • Article 8 ICESCR • Article 7 CEDAW • Article 41 ICRMW

Economic, Social, and Cultural Rights

RIGHT	DESCRIPTION
Labor rights (e.g., just working conditions, fair wages, right to form unions)	Everyone has the right to work, to free choice of employment, to just and favorable conditions of work, and to protection against unemployment. Everyone, without any discrimination, has the right to equal pay for equal work. Everyone who works has the right to just and favorable remuneration ensuring for himself and his family an existence worthy of human dignity, and supplemented, if necessary, by other means of social protection. Everyone has the right to form and to join trade unions for the protection of his interests. **International Human Rights Treaties and Conventions** • Article 23 UDHR • Article 27 CRPD • Article 7, 8 ICESCR • The Eight ILO Core Conventions
Right to social security	Everyone has the right to social security assistance when unable to work due to sickness, disability, maternity, employment injury, unemployment, or old age. **International Human Rights Treaties and Conventions** • Article 22 UDHR • Article 9 ICESCR • Article 26 CRC
Right to an adequate standard of living (e.g., adequate food, housing)	The right to an adequate standard of living for the health and well-being of himself and of his family recognizes survival and being able to live free of reasonably preventable suffering as a fundamental, basic right granted individuals and their families. This includes resources such as food, clothing, health care, and social services. **International Human Rights Treaties and Conventions** • Article 25 UDHR • Article 11 ICESCR • Article 27 CRC • Article 28 CRPD

Right to education	Education should be compulsory and free, at least in the elementary and fundamental stages. Forms of higher education should be made generally available and equally accessible. Education shall be directed to the full development of the human personality and to the strengthening of respect for human rights and fundamental freedoms. It shall promote understanding, tolerance, and friendship among all nations, racial or religious groups, and shall further the activities of the United Nations for the maintenance of peace. Parents have a prior right to choose the kind of education that shall be given to their children. **International Human Rights Treaties and Conventions** • Article 13 & 14 ICESCR • Article 26 UDHR • Article 28 CRC • Article 24 CRPD • Article 30 ICRMW
Right to participate in cultural life	Everyone has the right to freely participate, enjoy, and share in cultural life and scientific advancements and their benefits, while providing protection of the moral and material interests resulting from these productions. **International Human Rights Treaties and Conventions** • Article 27 UDHR • Article 15 ICCPR • Article 30 CRPD • Article 15 ICESCR
Right to benefit from scientific advancement	Everyone has the right to share scientific advancements and their benefits, and the right to the protection of the moral and material interests resulting from any scientific production of which he is the author. **International Human Rights Treaties and Conventions** • Article 27 UDHR • Article 15 ICCPR • Article 30 CRPD • Article 15 ICESCR
Right to health	Everyone has the right to the enjoyment of the highest attainable standard of physical and mental health. Everyone has the right to protection against threats to the state of well-being in which an individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively, and is able to make a contribution to his or her community. **International Human Rights Treaties and Conventions** • Article 12 ICESCR • Article 25 CRPD

Right to a healthy environment	Everyone on the planet has a right to a clean, healthy, and sustainable environment. A clean, healthy, and sustainable environment is necessary for the enjoyment of all human rights. States have the obligation to respect, protect, and promote human rights, including in all actions undertaken to address environmental challenges, and to take measures to protect the human rights of all, as recognized in different international instruments, and that additional measures should be taken for those who are particularly vulnerable to environmental degradation. All business enterprises have the responsibility to respect human rights, including the right to a clean, healthy, and sustainable environment. **International Human Rights Treaties and Conventions** • UN General Assembly Resolution 76/300 on the human right to a clean, healthy, and sustainable environment
Right to internet access	All sectors of society rely on the internet as a source of information, a means of participation in civil, political, economic, social, and cultural life, a means of gaining access to public services, including but not limited to education and health, a source of livelihood, and an arena for the exercise of human rights. Everyone has the right to the full protection of all human rights both online and offline. **International Human Rights Treaties and Conventions** • UN General Assembly Resolution A/HRC/RES/32/13 on the promotion, protection, and enjoyment of human rights on the internet

Thematic Rights

RIGHT	DESCRIPTION
Children's rights (e.g., to free expression, access to age appropriate information, freedom from exploitation)	**The UN Convention on the Rights of the Child** exists in recognition of the need to extend particular care to the child, which has been stated in the Geneva Declaration of the Rights of the Child of 1924 and in the Declaration of the Rights of the Child adopted by the General Assembly on 20 November 1959 and recognized in the Universal Declaration of Human Rights, in the International Covenant on Civil and Political Rights (in particular in articles 23 and 24), in the International Covenant on Economic, Social and Cultural Rights (in particular in article 10) and in the statutes and relevant instruments of specialized agencies and international organizations concerned with the welfare of children. It contains 54 articles enshrining the human rights of all children.

Disability rights (e.g., accessibility)	**The UN Convention on the Rights of Persons with Disabilities** recognizes the need to promote and protect the human rights of all persons with disabilities, including those who require more intensive support, and that discrimination against any person on the basis of disability is a violation of the inherent dignity and worth of the human person. It contains 47 articles enshrining the human rights of all persons with disabilities.
Indigenous peoples' rights	**The UN Declaration on the Rights of Indigenous Peoples** affirms that Indigenous peoples are equal to all other peoples, and that Indigenous peoples, in the exercise of their rights, should be free from discrimination of any kind. It contains 46 articles enshrining the human rights of all Indigenous peoples.

Appendix 2: Sample HRA Prioritization Criteria

The table below shows BSR's standard prioritization criteria for HRAs, and illustrates how scales can be defined for human rights assessment criteria to help assess the scope, scale, remediability, and likelihood of impacts. A five-point scale is used to align with common enterprise risk management (ERM) practices, which enables the impacts identified in HRAs to be incorporated into ERM systems. BSR then customizes the criteria to fit the scope of the assessment and align with other impact assessment processes. For example, scope and likelihood may be assigned numerical ranges that reflect a company's internal metrics and the reach of a given product or service.

Criteria	Levels				
Scope How many people are (or could be) affected by the adverse impact?	**Smallest** Smallest range of the relevant population impacted.	**Small** Limited/smaller range of the relevant population impacted.	**Medium** Majority of the relevant population impacted.	**Large** Larger majority of the relevant population impacted.	**Largest** Significant and/or all of the relevant population impacted.
Scale How serious are the impacts (or could they be) for affected individuals?	**Least Serious** Associated with indirect and/or minimal adverse impacts on physical, mental, civic, or material well-being.	**Moderately Serious** Associated with indirect and/or moderate adverse impacts on physical, mental, civic, or material well-being.	**Serious** Associated with direct and/or serious adverse impacts on physical, mental, civic, or material well-being.	**Very Serious** Associated with lasting adverse impacts on physical, mental, civic, or material well-being.	**Most Serious** Associated with irreversible adverse impacts on physical, mental, civic, or material well-being.

Remediability Can a remedy restore affected individuals to the same or equivalent position before the adverse impact?	**Remediable** Remedy would return those affected to the same or equivalent position before the adverse impact occurred.	**Likely Remediable** Remedy is likely to return those affected to the same or equivalent position before the adverse impact occurred.	**Possibly Remediable** Remedy may help return those affected to the same or equivalent position before the adverse impact occurred.	**Rarely Remediable** Remedy can rarely return those affected to the same or equivalent condition before the adverse impact occurred.	**Not Remediable** Remedy will not return those affected to the same or equivalent condition before the adverse impact occurred.
Likelihood What is the likelihood of the adverse impact occurring?	**Minor Probability** 0-10% chance of occurrence. Although a risk, it is highly unlikely that adverse impacts may occur.	**Some Probability** 10-40% chance of occurrence. There is some minor likelihood that adverse impacts may occur.	**Good Probability** 40-70% chance of occurrence. It's more likely than not that adverse impacts may occur.	**High Probability** 70-90% chance of occurrence. There is a high likelihood that adverse impacts may occur.	**Certain** 90-100% chance of occurrence. Averse impacts are currently occurring or certain to occur.

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