Human Rights Across the GenAI Value Chain

May 2025

Copyright © 2025 by Business for Social Responsibility (BSR)

Agenda

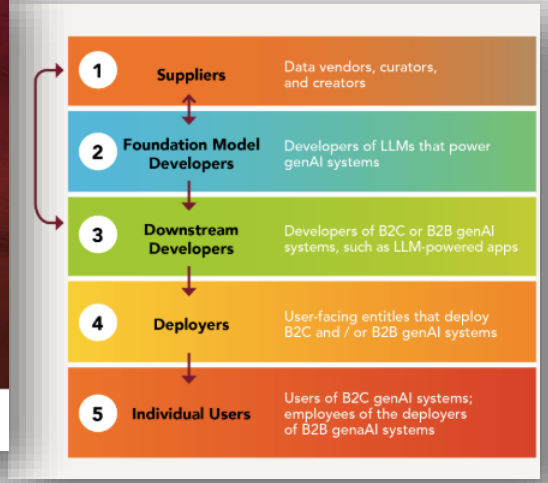
1. Intro to Generative AI and Human Rights
2. The Generative AI Value Chain
3. The Human Rights Impacts of Generative AI
4. Recommendations
5. Q&A

BSR
© 2020 Business for Social Responsibility

Why Human Rights and GenAI?

Human Rights and the GenAI Value Chain

A Human Rights Assessment of the GenAI Value Chain



Responsible AI Practitioner Guides

- ★ Overview of Responsible AI Practitioner Guides
- 1 Fundamentals of a Human Rights-Based Approach to Generative AI
- 2 Governance and Management
- 3 Impact Assessment
- 4 Risk Mitigation
- 5 Conducting Stakeholder Engagement
- 6 Policies and Enforcement
- 7 Aligning Transparency & Disclosure Practices with Human Rights
- 8 Remedy for Generative AI-Related Harms

BSR Speakers

Lindsey Andersen
Associate Director
San Francisco

Hannah Darnton
Director
San Francisco

J.Y. Hoh
Associate Director
Singapore

Samone Nigam
Manager
San Francisco

01 Introduction

What is Generative AI?

Generative AI (genAI): A type of AI that is capable of producing content in various modalities, including natural language text and speech, image, audio, video, and computer code.

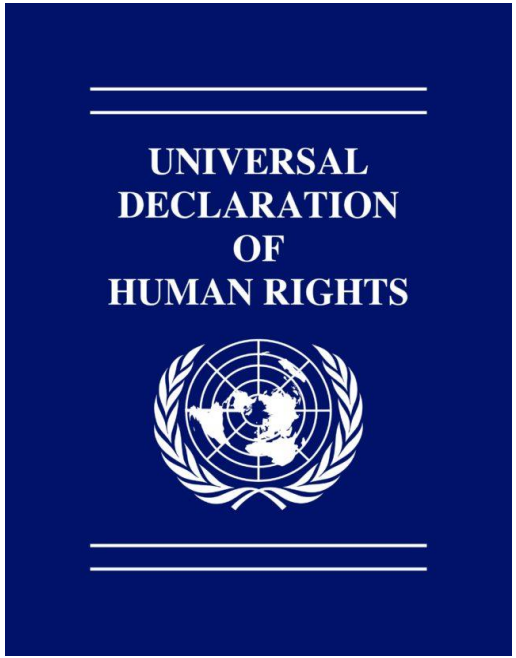
Key Capabilities	How it Works	Examples
• Text generation • Image creation • Audio and video synthesis • Code generation	• Trained on large datasets • Learns patterns, structures, and relationships • Generates outputs in response to prompts	• ChatGPT • DALL·E • Zoom AI Companion • Microsoft Copilot … & many more

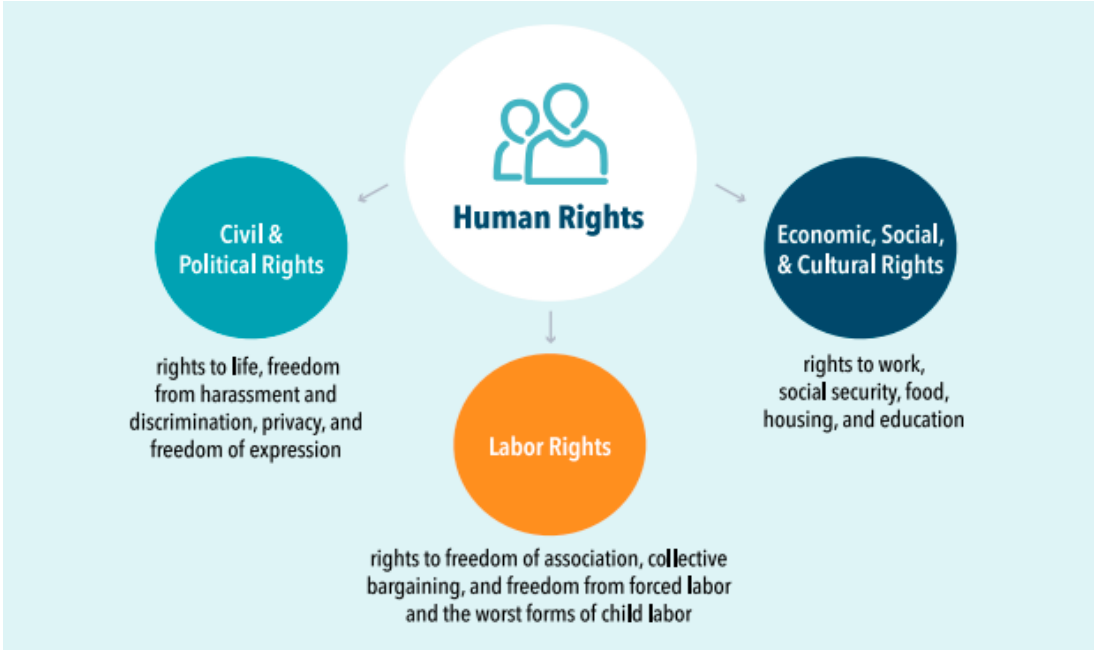
Known Risks and Challenges Associated with GenAI and LLMs

1	Inaccuracy / Hallucinations	A prominent airline pays damages after chatbot outputs false information about discounts to a customer
2	Privacy Breaches	Staff at a top tech company input top secret, proprietary data into ChatGPT
3	Under-representative Outputs	Image generation models reinforce stereotypes by only generating men when asked to produce images of doctors and CEOs
4	Harmful Content	The National Eating Disorder Association chatbot generates harmful instructions about weight loss and diet culture

What are Human Rights?

Human rights are **inherent** to us all, regardless of nationality, sex, national or ethnic origin, color, religion, language, or any other status. They are globally agreed upon standards of achievement for all, covering a wide range of independent yet interconnected civil, political, economic, social, cultural, and environmental rights.





What are Human Rights?

- They are codified in United Nations **international human rights instruments**
- They have been incorporated into international treaties, regional human rights instruments, and national constitutions and legal codes
- They are **overseen by the UN Human Rights System** — tasked with monitoring and protecting human rights and further developing them / releasing guidance about what they mean in practice

International Human Rights Instruments

- Universal Declaration of Human Rights
- International Covenant on Civil & Political Rights
- International Covenant on Economic, Social & Cultural Rights
- Thematic Conventions — e.g. labor rights, child rights, rights of persons with disabilities

UN Guiding Principles on Business and Human Rights (UNGPs)

UN Guiding Principles on Business and Human Rights

PROTECT

State Duty to Protect

Safeguard human rights through appropriate policies, regulation, and adjudication.

RESPECT

Corporate Responsibility to Respect

Company must “know and show” human rights impacts and steps to mitigate and eliminate.

REMEDY

Provide Access to Remedy

States and companies must remedy abuses through judicial and non-judicial means.

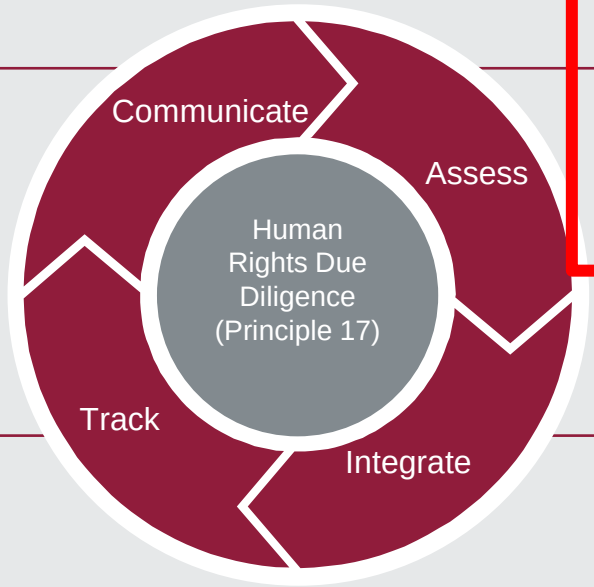
Human Rights Due Diligence and Human Rights Assessments

Communications (Principle 21)

"In order **to account for how they address their human rights impacts**, business enterprises should be prepared to **communicate this externally**"

Assessing Impacts (Principle 18)

"In order to gauge human rights risks, business enterprises should **identify and assess any actual or potential adverse human rights impacts**"



Tracking (Principle 20)

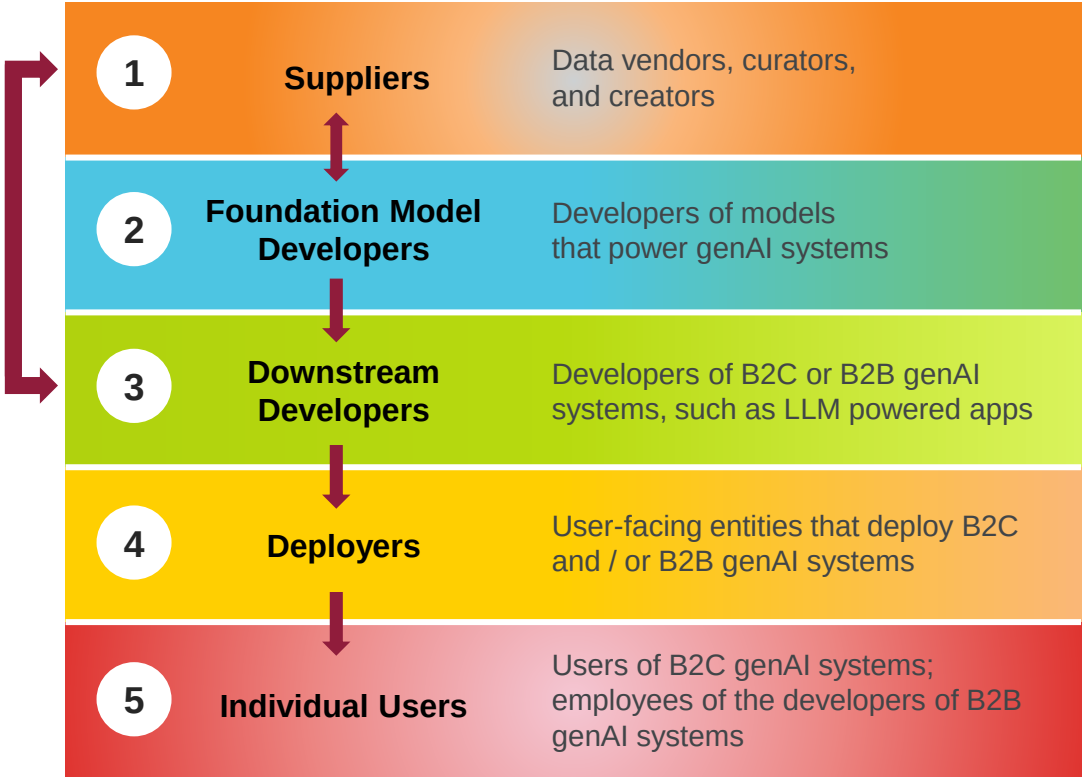
"In order to verify whether adverse human rights impacts are being addressed, business enterprises should **track the effectiveness of their response**"

Integration and Action (Principle 19)

"In order to prevent and mitigate adverse human rights impacts, business enterprises **should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action**"

02 The Generative AI Value Chain

The Generative AI Value Chain

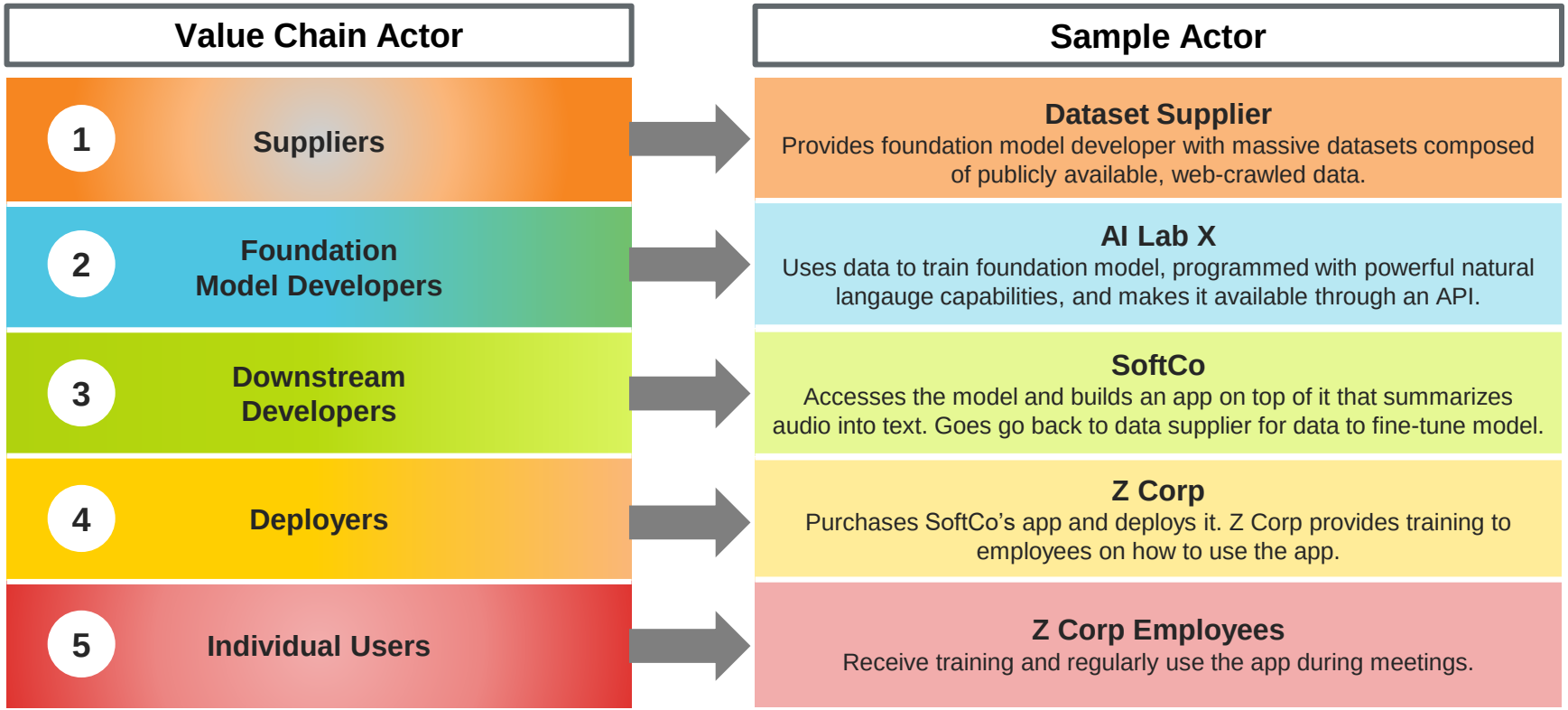


Some entities straddle multiple categories. For instance, OpenAI is a supplier, a foundation model developer, and a deployer.

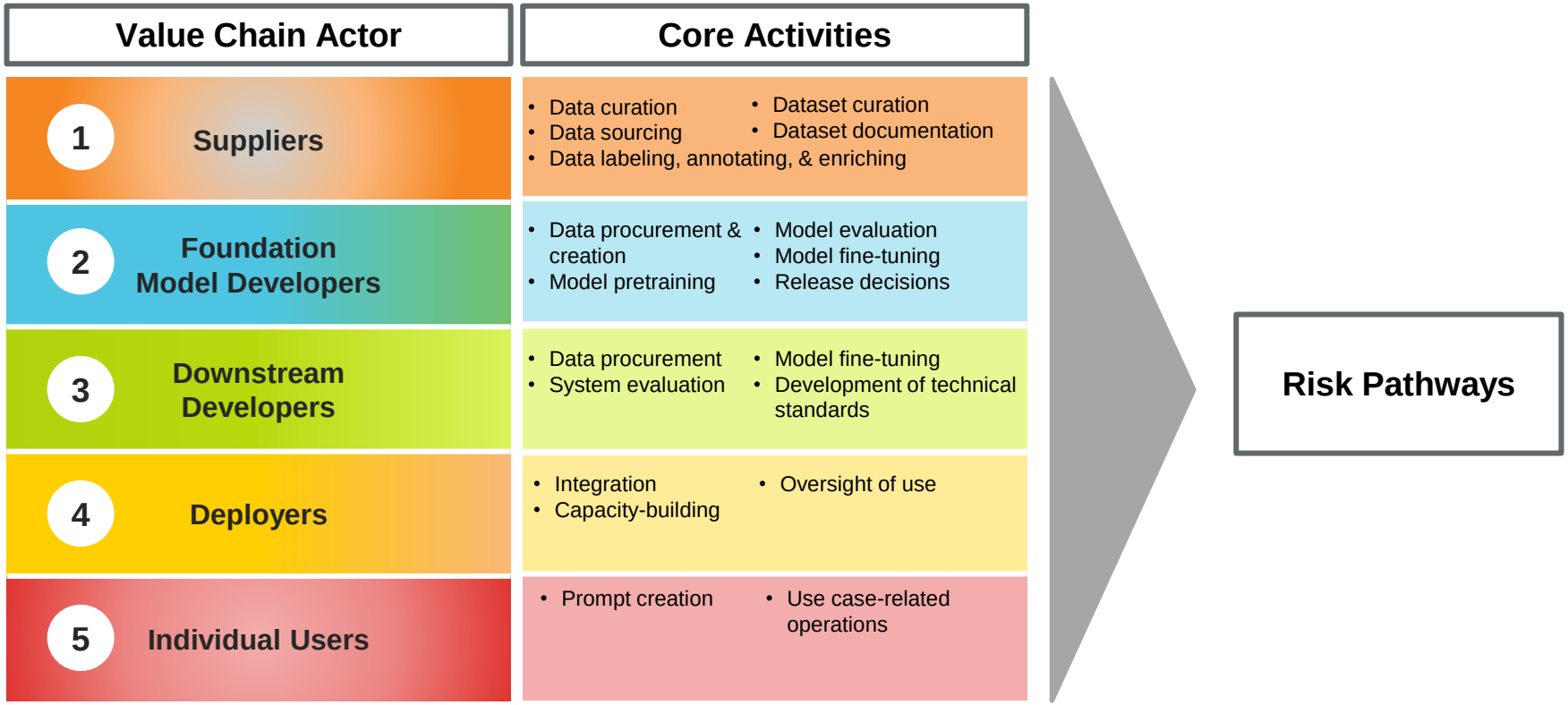
Actors across the value chain work closely together, particularly during the training and evaluation process. Developers may go back to suppliers multiple times for additional data.

Other actors, such as regulators or investors, may be “value-chain adjacent” – meaning that their activities have impacts across the value chain.

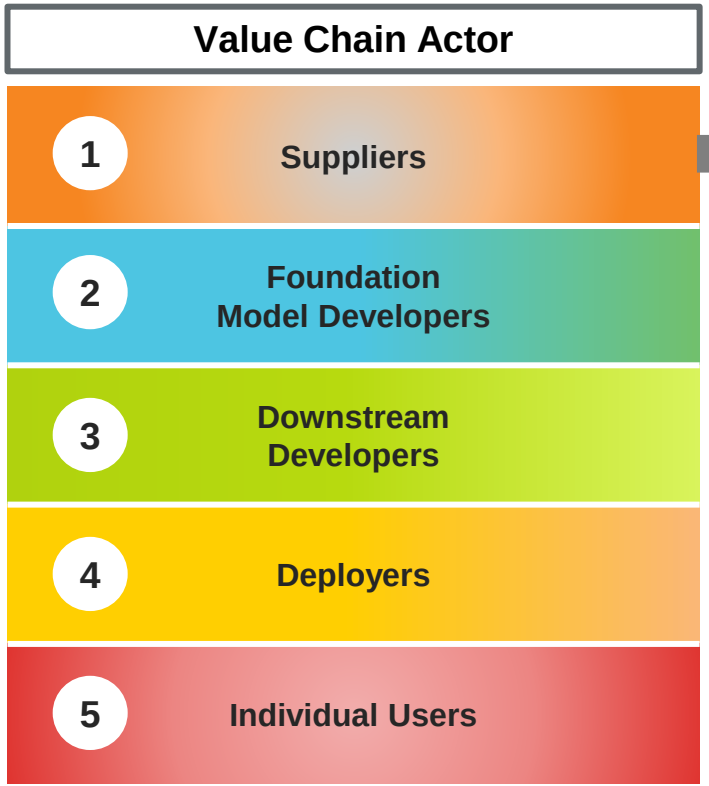
The GenAI Value Chain | Example: Call Summarization Tool



The GenAI Value Chain | Core Activities and Risk Pathways



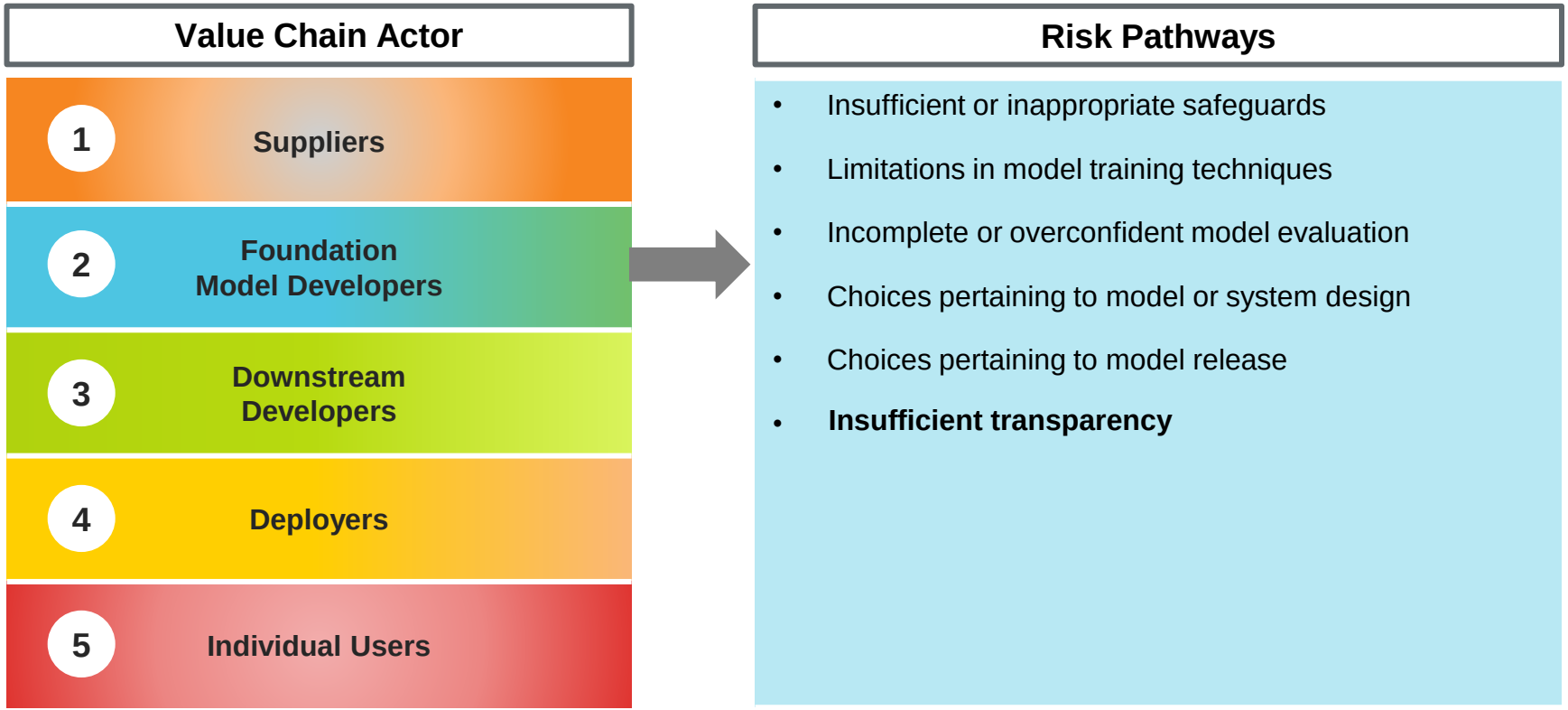
The GenAI Value Chain | Risk Pathways for Suppliers



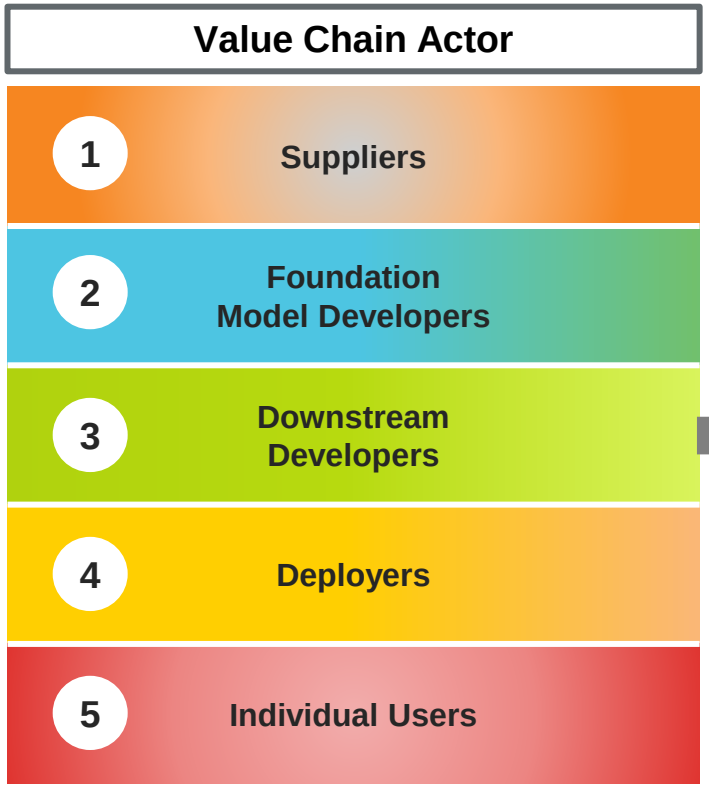
Risk Pathways

- Insufficient or misleading dataset documentation
- Failure to engage with downstream applications
- Misalignment between dataset and downstream use
- Failure to clean biased data
- **Unrepresentative dataset curation**
- Poor data labeling guidance
- Ethical remuneration and working conditions for data labelers, annotators, and creators

The GenAI Value Chain | Risk Pathways for Foundation Model Developers



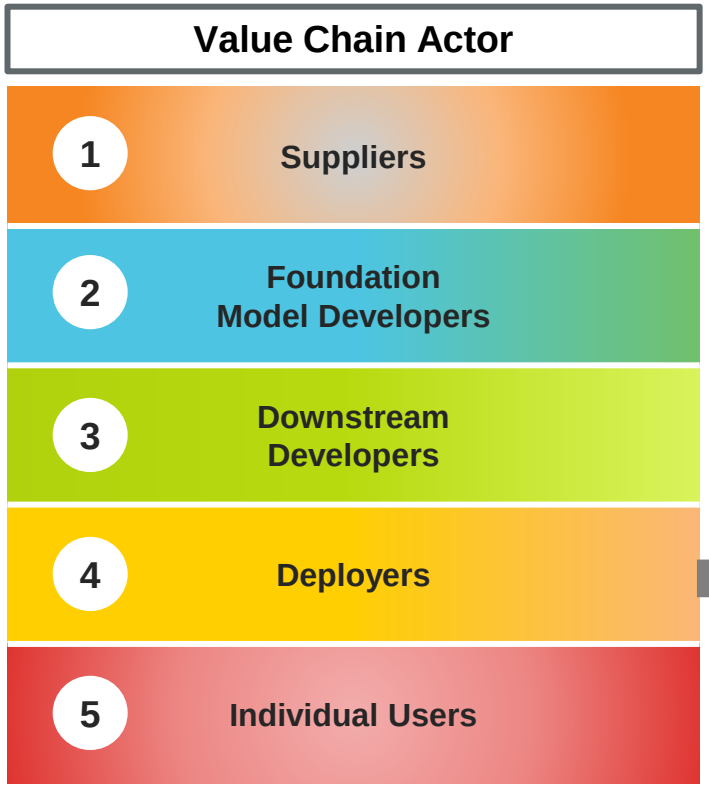
The GenAI Value Chain | Risk Pathways for Downstream Developers



Risk Pathways

- Ineffective technical mitigations
- Choices pertaining to model fine-tuning
- **System design or intended deployment context**
- Lack of feedback channels for deployers and individual users
- Invalid or insufficient evaluation

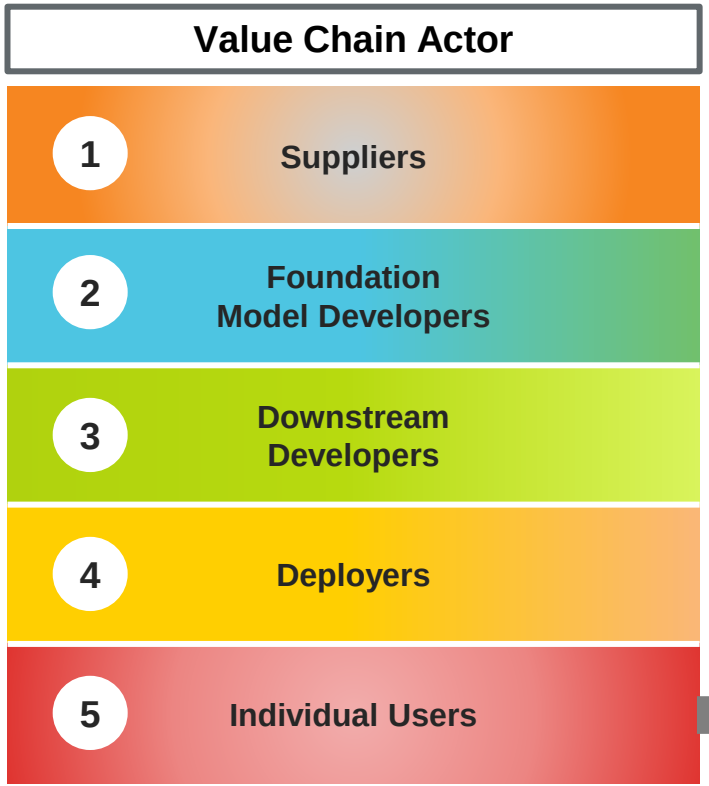
The GenAI Value Chain | Risk Pathways for Deployers



Risk Pathways

- Limited genAI literacy and/or an absence of safety culture
- **Choices pertaining to deployment and integration**
- Ineffective human oversight
- Ineffective scalable oversight
- Limited awareness pertaining to developer safeguards

The GenAI Value Chain | Risk Pathways for Individual Users



Risk Pathways

- Intentional manipulation, misuse, or abuse
- Unintentional misuse or harm
- Limited genAI literacy
- Overreliance
- **Automation bias**

03 Human Rights Impacts

List of Internationally Recognized Human Rights

- Right to equality and non-discrimination
- Right to life, liberty, and personal security
- Freedom from slavery
- Freedom from torture and degrading treatment
- Due process and fair trial rights
- Freedom from arbitrary arrest and exile
- Right to privacy
- Freedom of movement
- Right to asylum
- Right to a nationality and the freedom to change nationality
- Right to marriage and family
- Right to own property
- Freedom of thought
- Freedom of religion and belief
- Right to remedy
- Freedom of opinion, expression, and access to information
- Right of peaceful assembly and association
- Right to political participation
- Right to social security
- Labor Rights (e.g. safe working conditions, adequate remuneration, right to join unions)
- Right to rest and leisure
- Right to adequate living standards
- Right to health
- Right to education
- Right to participate in the cultural life of the community
- Right to benefit from scientific advancement
- Right to internet access
- Right to a healthy environment
- Disability rights (e.g. right to accessibility)
- Child Rights

Salient Human Rights Impacts

1

Equality and Nondiscrimination

- Outputs reinforce stereotypes / encourage discrimination
- Over / under / mis-representation
- Process discrimination
- Inaccessibility

2

Access to Information

- Hallucination
- Disinformation
- Flooding of synthetic content

3

Privacy

- Personal data leaked in model outputs
- Personal data revealed through invasive analytics
- Generation of non-consensual explicit imagery

4

Economic, Social, and Cultural Rights

- Decrease in employment opportunities
- Infringement on intellectual property rights
- Facilitation of cybersecurity attacks

5

Bodily Integrity

- Incitement of societal violence
- Facilitation of violence or harassment
- Connection to critical system failures

6

Freedom of Thought and Opinion

- Decline in critical analysis skills
- Hyper-personalized behavioral nudging

Salient Human Rights Impacts

1

Equality and Nondiscrimination

- Outputs reinforce stereotypes / encourage discrimination
- Over / under / mis-representation
- Process discrimination
- Inaccessibility

2

Access to Information

- Hallucination
- Disinformation
- Flooding of synthetic content

3

Privacy

- Personal data leaked in model outputs
- Personal data revealed through invasive analytics
- Generation of non-consensual explicit imagery

4

Economic, Social, and Cultural Rights

- Decrease in employment opportunities
- Infringement on intellectual property rights
- Facilitation of cybersecurity attacks

5

Bodily Integrity

- Incitement of societal violence
- Facilitation of violence or harassment
- Connection to critical system failures

6

Freedom of Thought and Opinion

- Decline in critical analysis skills
- Hyper-personalized behavioral nudging

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	**Poor dataset curation**	**Training dataset does not contain key info about taxpayer dates and procedures**

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Training dataset does not contain key info about taxpayer dates and procedures
Foundation Model Developers	**Model training limitations**	**Model training choices result in lower levels of accuracy**

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Training dataset does not contain key info about taxpayer dates and procedures
Foundation Model Developers	Model training limitations	Model training choices result in lower levels of accuracy
Downstream Developers	**Model fine-tuning choices, HuRi-impacting use cases / app domains**	**Adequate fine-tuning for accuracy not conducted due to commercial pressures to release the app**

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Training dataset does not contain key info about taxpayer dates and procedures
Foundation Model Developers	Model training limitations	Model training choices result in lower levels of accuracy
Downstream Developers	Model fine-tuning choices, HuRi-impacting use cases / app domains	Adequate fine-tuning for accuracy not conducted due to commercial pressures to release the app
Deployers	**Choices pertaining to integration, limited genAI literacy / safety culture**	**Chatbot integrated into tax filing tools without telling users to fact-check outputs**

Access to Information | Connections to Value Chain Actors

Risk Category: **Hallucination** - GenAI models may “hallucinate” by producing false or misleading outputs.

Risk Example

A genAI chatbot designed to assist taxpayers in filing their taxes provides them with incorrect information about filing dates and procedures.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Training dataset does not contain key info about taxpayer dates and procedures
Foundation Model Developers	Model training limitations	Model training choices result in lower levels of accuracy
Downstream Developers	Model fine-tuning choices, HuRi-impacting use cases / app domains	Adequate fine-tuning for accuracy not conducted due to commercial pressures to release the app
Deployers	Choices pertaining to integration, limited genAI literacy / safety culture	Chatbot integrated into tax filing tools without telling users to fact-check outputs
Individual User	**N/A**	**Chatbot's guidance relied upon without fact-checking**

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	**Poor dataset curation**	**Supplied dataset contains personal information.**

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Supplied dataset contains personal information.
Foundation Model Developers	**Insufficient safeguards, model training limitations**	**Privacy techniques, such as differential privacy, are not applied during model training.**

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Supplied dataset contains personal information.
Foundation Model Developers	Insufficient safeguards, model training limitations	Privacy techniques, such as differential privacy, are not applied during model training.
Downstream Developers	**Ineffective technical mitigations**	**No product policy against queries that might result in personal data leaks.**

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Supplied dataset contains personal information.
Foundation Model Developers	Insufficient safeguards, model training limitations	Privacy techniques, such as differential privacy, are not applied during model training.
Downstream Developers	Ineffective technical mitigations	No product policy against queries that might result in personal data leaks.
Deployers	**N/A**	**(Downstream developer is also the deployer in B2C products)**

Privacy | Connections to Value Chain Actors

Risk Category: **Data leakage** – Model outputs may contain sensitive personal data.

Risk Example

A genAI chatbot is trained on a dataset containing sensitive personal information, such as addresses, bank account details, and passwords. When the chatbot is queried about specific individuals, some of that data is leaked in the outputs.

Entity	Risk Pathways	Example Connection to the Entity
Suppliers	Poor dataset curation	Supplied dataset contains personal information.
Foundation Model Developers	Insufficient safeguards, model training limitations	Privacy techniques, such as differential privacy, are not applied during model training.
Downstream Developers	Ineffective technical mitigations	No product policy against queries that might result in personal data leaks.
Deployers	N/A	(Downstream developer is also the deployer in B2C products)
Individual User	**N/A**	**Users try to get the chatbot to leak information about people they know.**

Vulnerable Groups

The UNGPs state that companies should pay "particular attention to the rights and needs of, as well as the challenges faced by, **individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized."**

Vulnerable Group Analysis for Access to Information

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

04 Recommendations

Summary of Recommendations

Cross-Ecosystem	• Include all value chain in risk assessments • Transparency and communications between value chain actors	• Collaborate on safety standards • More collaboration with affected stakeholders
Suppliers	• Workflows for responsible data collection and curation • Work with developers on data documentation standards	• Work with developers on standards for data annotation
Foundation Model Developers	• Create data procurement requirements and evaluate datasets • Create public documentation for model release	• Create responsible use guidance for downstream developers • Work with downstream developers to address risks
Downstream Developers	• Consider risks of B2C or B2B models • Engage deployers and users throughout genAI product lifecycle	• Incorporate reporting mechanisms for deployers, users, and other stakeholders
Deployers	• Report issues to downstream developer • Engage downstream developers in system design	• Invest in genAI literacy and safety culture • Consider labor rights impacts before adopting genAI

Cross-Ecosystem Recommendations

- Consider the role of the value chain in risk assessment / HRDD efforts
- Transparency, communication, and collaboration across the value chain to mitigate risk

Supplier Recommendations

- Create workflows for responsible data collection, curation, and provision
- Work with developers on standards for data annotation and data documentation

Foundation Model Developer Recommendations

- Publish model documentation and responsible use guidance for downstream developers
- Collaborate closely with downstream developers to address risks

Downstream Developer Recommendations

- Consider the deployment context; engage deployers and users throughout product lifecycle
- Develop reporting mechanisms for product use issues

Deployer Recommendations

- Invest in capacity-building for developing genAI literacy and promoting a culture of safety
- Communicate proactively with developers about product use

Join us for Webinar 2!

Practitioner Guidance for Human Rights-Based AI Governance

May 21

8-9 am PDT and SGT

Register at www.bsr.org/events

04 Q&A