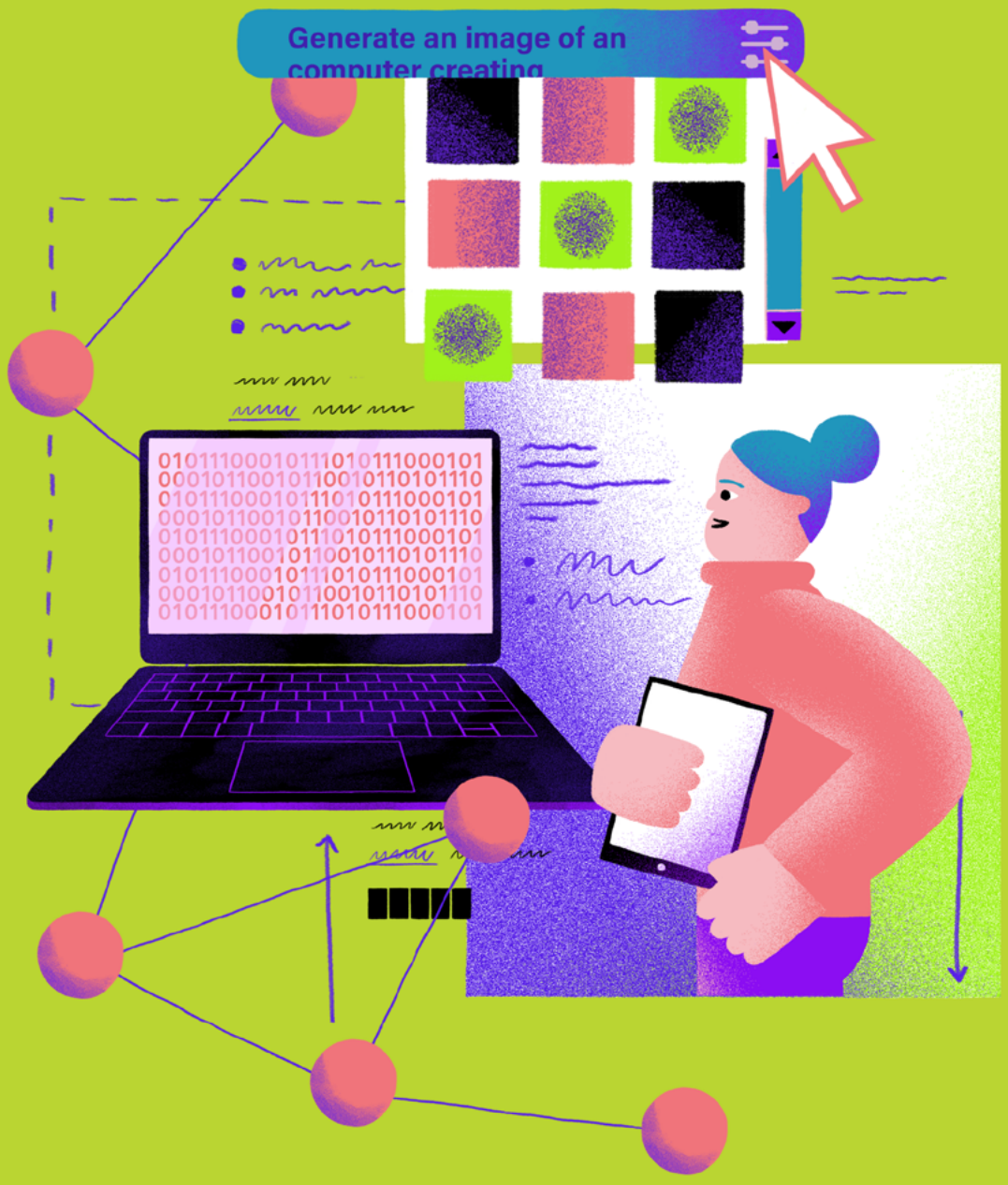
unesco

# COMPANION DOCUMENT:

# The Guidelines for the Governance of Digital Platforms and Generative Artificial Intelligence



JULY 2025

## DISCLAIMER

*Generative AI models, which are the subject of this document, have experienced accelerated growth in recent years, with further advancements observed in recent months. This rapid improvement, juxtaposed with significant knowledge gaps regarding the real-world impact of this technology, underscores the critical governance challenges at hand. This document endeavors to analyze the potential benefits and risks of AI advancements on human rights and freedom of expression, while acknowledging the limited availability of robust scientific evidence on AI trajectory and effective safeguard mechanisms.*

Published in 2025 by the United Nations Educational, Scientific and Cultural Organization, 7, place de Fontenoy, 75352 Paris 07 SP, France ©UNESCO.



*The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.*

*The recommendations herein build on those outlined in the Guidelines for the Governance of Digital Platforms, are based on multistakeholder inputs, existing frameworks as well as the latest research and existing evidence. In recognition of the dynamic nature of this field, this document constitutes the initial release of a living document, subject to periodic reviews and updates to reflect subsequent technological changes and impacts. All stakeholders are invited to contribute to future iterations by providing comments and suggestions to* internetconference@unesco.org.

# 1. Introduction

The emergence of Generative Artificial Intelligence (AI) has ignited widespread discussion regarding its potential to transform and elevate human prosperity. However, the field remains marked by substantial uncertainty, and robust empirical evidence on the real-world impact on human rights is critically needed to inform evidence-based policy. Without effective governance, generative AI's opportunities risk remaining unrealized or being distributed inequitably, exacerbating existing disparities. As highlighted by the United Nations Secretary-General's High-level Advisory Body dedicated to this technology topic: **'AI governance is crucial – not merely to address the challenges and risks, but also to ensure that we harness AI's potential in ways that leave no one behind'**.[1]

Although innovation in AI systems has occurred for many decades, the last two years have been marked by an exponential rise in development and attention to the technology, particularly in its generative capabilities.[2] In a short period of time, generative AI, which can create 'original' content in response to a simple text input, has reshaped the discourse on AI as well as its impacts on society.[3] Unlike innovation requiring infrastructure and hardware, scaling AI to millions of users doesn't involve meaningful investment per extra user due to its zero marginal or distribution costs. For this reason, the AI adoption rates outpaced that of personal computers and the internet and generative AI has swiftly transformed our information ecosystem.[4] This profound shift is especially evident in how people, particularly younger demographics, access information. **Increasingly, AI-powered platforms are perceived as more trustworthy than other outlets**. This trend has led to a bypass of established news websites, with younger users predominantly favouring recommended feeds and chatbots as their primary information sources. This differs sharply from conventional search engines, which historically served as intermediaries, directing users to original news content. In contrast, generative AI tools now directly parse, synthesize, and repackage information, fundamentally shifting information access away from original journalistic sources.

More broadly, despite inherent uncertainties surrounding its ultimate impact, **experts largely concur on generative AI's substantial potential to undermine information integrity.** This concern is not merely speculative; a significant majority of AI experts – 78% of those surveyed across diverse disciplines and 68 countries – expressed concern or strong concern regarding the damage generative AI could inflict on the integrity of information.[5] Deepfakes, voice clones or other forms of AI-generated content can severely undermine scientifically established facts and pose serious threats to democratic institutions and social trust more generally. In particular, this technology facilitates the creation of deceptive content that misrepresents individuals or events, including the fabrication of actions, statements, or altered event locations. Generative AI amplifies users' ability to rapidly and convincingly generate disinformation and hate speech which can then be disseminated to a broad audience on digital platforms. For instance, according to a UNESCO 2021 study, women journalists are disproportionately targeted with sexualized deepfakes, which not only demean those featured but also contribute to a chilling effect on journalism.[6] Moreover, the pervasive threat of

---

1 United Nations. 2024. *Governing AI for Humanity: The Final Report of the United Nations Secretary-General's High-level Advisory Body on Artificial Intelligence*. New York, United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

2 AI lacks a single definition or form but broadly refers to systems which process data and information in a manner resembling human intelligence, involving reasoning, learning, perception, prediction, planning or control. They comprise diverse technologies and techniques that allow the production and dissemination of information and content, with self-learning and adaptive characteristics. See UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

3 At the time of drafting, the originality and novelty of AI-generated content remain subject to debate and ongoing copyright litigation. This document does not aim to adjudicate these cases; therefore, any qualification of AI-generated content is bracketed.

4 Current adoption rate studies predominantly focus on European and US markets, necessitating expanded research to accurately gauge global uptake.

5 The final report of United Nations Secretary-General's High-level Advisory Body on AI provides an overview of expert risk perceptions on AI-related trends and risks from 348 AI experts across disciplines and 68 countries in all regions. Damage on information integrity ranks the highest with 78 percent of experts being concerned or very concerned about this risk. See *Governing AI for Humanity, op. cit.*

6 Posetti, J. et al. 2021. *The Chilling: global trends in online violence against women journalists; research discussion paper.* Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377223.

such malicious content, and the resulting risk of reputational damage and psychological abuse, can deter individuals from public engagement, even without direct targeting.

Beyond disinformation-related harms, this technology also presents significant concerns for the overall protection and promotion of human rights, artistic, and cultural production as well as for the safety of women, girls, and communities in situations of vulnerability and marginalization. While there is ample evidence of AI susceptibility to biases (notably due to incomplete or unrepresentative datasets), gender-based discrimination and harms resulting from the use and misuse of generative AI is substantial.

Furthermore, and perhaps most critically, generative AI and AI-powered platforms pose a significant threat to the pluralism of voices and the diversity of content. This danger stems partly from the unequal distribution of AI resources, encompassing data, hardware infrastructure, and specialized expertise, coupled with increasing market concentration among AI providers. This lack of diversity and imbalance manifests also in the outputs of generative AI models, which exhibit a tendency to overrepresent dominant cultural groups, potentially leading to the misrepresentation or underrepresentation of other communities on a substantial scale. The consistent underrepresentation of low-resource languages in training data results in suboptimal performance for their speakers, and these communities face the risk of being portrayed exclusively through an 'outsider' perspective, rather than their own cultural expressions. Such an outcome actively reinforces the values and predispositions of dominant cultural and political groups, while simultaneously marginalizing the linguistic, historical, and cultural diversity of other communities, thereby systematically constricting the digital landscape of perspectives and limiting genuine pluralism.

**Without governance to ensure accountability and cooperation to build capacity, facilitate access and increase diversity and pluralism, the existing 'AI divide' could further widen and become entrenched**, accentuating AI risks, particularly to freedom of expression, and limiting its meaningful contribution to scientific advances, economic growth and progress on the Sustainable Development Goals (SDGs).

While there is a growing consensus on the critical need to govern this rapidly evolving technology, a **global governance deficit** persists with respect to AI, and its most advanced generative capabilities in particular. Notwithstanding recent progress and ongoing efforts, the patchwork of norms and institutions is still nascent, disjointed and full of gaps.[7]

**This document aims to fill some of these gaps by drawing parallels and lessons from other fields of governance.** Using *the UNESCO Guidelines for the Governance of Digital Platforms* (hereinafter referred to as 'the Guidelines') as its foundation, it considers how critical principles for the governance of digital platforms apply to generative AI in a way that safeguard freedom of expression, access to information and other human rights, while ensuring robust safety and risk mitigation measures.[8] Furthermore, it highlights some emerging good practices to enable agility and adaptation to rapid advancements in technology. In this way, and consistent with *the Guidelines'* overarching philosophy, the approach prioritizes human rights at the core of technology governance. This is crucial for fostering consistent global responses and mitigating the risk of fragmented regulations or disparate approaches that could ultimately compromise universal norms. By embedding human rights deeply within the framework of AI governance, it seeks to ensure a unified and equitable global trajectory for this rapidly advancing technology.

7 See, for example, Gutiérrez, J. D. 2024. *Consultation Paper on AI Regulation: Emerging Approaches Across the World*. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000390979. The paper discusses the emerging approaches to regulating the value chains of artificial intelligence systems worldwide.

8 UNESCO. 2023. *The Guidelines for the Governance of Digital Platforms*. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000387339.

In particular, this document may serve as a resource for a range of actors, including:

- Policymakers and governance bodies in identifying principles and inclusive processes that could be considered across technologies;
- digital AI developers and providers in their design, policies and practices;
- journalists and media in their use of generative AI, when scrutinizing technological developments and companies, and in support of media and information literacy; and
- other stakeholders, such as civil society and researchers, in their advocacy and accountability efforts.

## 1.1 Objectives and Structure

**This document systematically reviews each of the universal principles presented in *the Guidelines* and considers their application in the governance of generative AI.** It also delineates specific areas necessitating the expansion of current governance frameworks to adequately address the implications of generative AI for freedom of expression and access to information, as well as for fostering the promotion and dissemination of diverse cultural content.

This undertaking is also in line with the ethos of *the Guidelines.* Developed as a living document, *the Guidelines* are subject to periodic review and updates, with several companion documents supporting implementation efforts. This document and its recommendations are intended to equip stakeholders with essential tools to address online harms and safeguard human rights in the digital age. In the face of rapid AI advancements and the emergence of diverse regulatory approaches, this paper offers universal principles for upholding human rights and freedom of expression and provides practical guidance for media and civil society confronting online hate speech, discrimination, and disinformation. This includes multistakeholder coalitions, such as those established by **UNESCO's Social Media 4 Peace (SM4P) project**, as well as broader networks of fact-checkers and organizations focused on these issues.[9] It can support these stakeholders in formulating and implementing advocacy strategies with digital platforms and relevant authorities to address the systemic transformations driven by this technology and, crucially, to **safeguard freedom of expression and the right to access information.**

The chapters of this document largely mirror *the Guidelines'* organizational structure. This document first defines key notions related to generative AI and provides necessary context on the existing AI governance landscape and reported gaps. It then considers *the Guidelines'* multistakeholder, system-based, and human rights-respecting approach and its relevance for the governance of AI. It also details roles and responsibilities of States and AI actors.[10] Finally, it offers recommendations by identifying areas where further research and additional mechanisms may be required to better address the risks and opportunities of generative AI as it pertains to freedom of expression and access to information.

Considering that *the Guidelines* are this document's foundations, many concepts as well as the language and framing are borrowed from the original text.[11]

## 1.2 Methodology

**This document consists of a meticulous review of each of the Guideline's principles and their application to the design, training, development and deployment of generative AI**. It aims to offer practical guidance to safeguard freedom of expression, access to digital information, and diverse

9 UNESCO. Social Media 4 Peace. https://www.unesco.org/en/social-media4peace.

10 AI actors are those who play an active role in the AI system lifecycle, including organizations and individuals that develop, deploy, or operate AI.

11 Sentences directly extracted from *the Guidelines* are marked in italic with reference to the exact article.

cultural content, while countering the dissemination and (artificial) creation of content that could be permissibly restricted under international human rights standards.[12] The content of this paper was also informed by a **series of qualitative interviews** with key experts and stakeholders involved in the governance of AI as well as Social Media 4 Peace Coalition members.

This document draws on insights from *the Guidelines'* extensive consultations and the evidence provided by UNESCO commissioned studies which shaped the development of *the Guidelines.*[13] Furthermore, this document references *the UNESCO Recommendation on the Ethics of Artificial Intelligence,* highlighting several specific articles that directly address the profound impact of AI on information ecosystems and freedom of expression.[14] It also considers **complementary global frameworks** which address the risks of AI and its impact on human rights, including the United Nations Secretary-General's AI Advisory Body reports, as well as *the International Scientific Report on the Safety of Advanced AI, the OECD Trustworthy AI Principles* and *the G7 Hiroshima Process International Guiding Principles of the Organizations of Developing Advanced AI system***.** It also takes into account international instruments on related issues such as information integrity including *the Windhoek +30 Declaration, the United Nations Pact for the Future, Global Digital Compact and Declaration on Future Generations and the Global Principles for Information Integrity.*

Finally, close attention has been paid to the **burgeoning field of research and scientific inquiry** into the risks and opportunities of generative AI. A comprehensive yet non-exhaustive bibliography is provided to allow readers to further explore the literature that has informed the document's findings and recommendations.

12 United Nations. 1966. *International Covenant on Civil and Political Rights*. New York, United Nations. Article 19.3.

13 UNESCO. Consultation Process. Internet for Trust. https://www.unesco.org/en/internet-trust/guidelines-consultation-process?hub=71542.

14 UNESCO. 2021. Recommendation on the Ethics of the Artificial Intelligence. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

# 2. Terminology

Generative AI is an emerging field which is characterized by a wide array of contentious terminology that lacks precise definitions or well delineated boundaries. This section aims to define key notions related to generative AI. It should not be conceived as an attempt to create a broad consensus on this matter, only to clarify how specific terms are used in this document. Similarly, it does not constitute an exhaustive list of definitions, it only focuses on technology or processes necessary to explain generative AI which are relevant for its governance.

## 2.1 Foundation models or General-Purpose AI (GPAI)

Most generative AI applications today are built upon foundation models which are typically trained on broad data and can perform, or can be adapted to perform, a wide variety of tasks.[15] The concept of '**general-purpose AI' or GPAI,** is sometimes preferred to foundation models and can be found in recent policy processes and texts such as the European Union's Artificial Intelligence Act.[16]

Another common term is **large language models (LLMs)**, which refers to a type of models defined by their extensive parameter counts and pretraining on vast amounts of text data. They serve as the core technology for a significant portion of today's foundation models (though not all, as some are being trained on vision, robotics, or reasoning and search). These models excel at various text-based tasks, such as question-answering, translation, and summarization.

One defining characteristic of foundation models, as suggested by the name, is the fact that they can be used as the **foundation for many other applications**. While they underpin most generative AI systems today, the majority of generative AI users do not interact directly with them. Rather, they engage through applications built on them. For example, the popular application ChatGPT is built on the GPT-3.5 and GPT-4 families of foundation models.[17]

A second important aspect of foundation models is the **scale of data** involved in training them. Training generative AI models necessitates vast datasets, frequently sourced from publicly accessible webpages through automated web scraping. This process, which involves extracting content from the internet, including digital platforms, has been repeatedly shown to inadvertently collect significant quantities of copyrighted material and personal data.[18]

Finally, they are also characterized by a **high degree of opacity**. This is in part due to the fact that their complex computational processes are hard to explain. It also stems from the proprietary nature of these models, where commercially sensitive training data, methodologies, and decision-making processes are usually not open to public scrutiny.

**Because of their broad use and purpose, foundation models have become a critical consideration in the governance of AI.** Their versatility brings significant uncertainties concerning their societal and systemic implications. Furthermore, as the base for a wide range of AI tools, any errors or issues at the foundation model level (and related dataset) may impact end users and non-users in diverse yet significant ways.

---

15 The term 'foundation model' was popularized in 2021 by researchers at the Stanford Institute for Human-Centered Artificial Intelligence, in collaboration with the Stanford Center for Research on Foundation Models, an interdisciplinary initiative set up by the Stanford Institute for Human-Centered AI.

16 The European Commission. *The AI Act Explorer: EU Artificial Intelligence Act. Official Journal version of 13 June 2024*. Brussels, European Union. https://artificialintelligenceact.eu/ai-act-explorer.

17 Jones, E. 2023. *Explainer: What is a Foundation Model?* Ada Lovelace Institute. https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/.

18 Data protection authorities have cautioned against the use of web scraping techniques through which individuals may lose control of their personal information when these are collected without their knowledge, against their expectations, and for purposes potentially different from those of the original collection. They also highlight that web scraping may not comply with relevant data protection principles, including data minimization and accuracy, insofar as there is no assessment on the reliability of the sources. See European Data Protection Supervisor. 2024. *First EDPS Orientations for ensuring data protection compliance when using Generative AI systems. Brussels,* European Data Protection Supervisor. https://www.edps.europa.eu/system/files/2024-06/24-06-03_genai_orientations_en.pdf; 2024. *Opinion 28/2024 on Certain Data Protection Aspects Related to the Processing of Personal Data in the Context of AI Models*. Brussels, European Data Protection Board. https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en.

## 2.2 Generative AI

Generative AI is a category of AI that can generate seemingly new content in a variety of formats – also known as **modalities** which can include text, images, sounds and videos – in response to user inputs also referred to as **prompts**. Generative AI can be unimodal (operating and generating content in a single mode or type of data) or multimodal (handling and generating content in multiple modes or types of data simultaneously). This generative capacity is predicated on the analysis and learning from extensive datasets, often incorporating significant portions of internet-derived data.

Generative AI is often regarded not only as a **critical turning point** in the development, but also adoption of artificial intelligence. Despite prior broad utilization of the technology in academic circles, this class of AI and related applications mark a revolution in accessibility and were rapidly endorsed by internet users.[19]

Creating generative AI (foundation) models and applications is a complex process and the sequence of required tasks varies between providers. However, it is worth noting that this technology is developed and deployed following a series of distinct stages including but not limited to pre-training, fine-tuning, system integration, deployment, and post deployment updates or model alignment which each requires different stakeholders, methods, and resources. Breaking down generative AI into phases can help reveal intervention points and fine tune governance systems.[20]

The intricate and multifaceted nature of generative AI has led to the development of diverse lifecycle models, each highlighting different phases of the process and the interdependent relationships among various providers. Although a comprehensive overview of the generative AI life cycle is not the objective of this document, the subsequent simplified diagram Figure 1, focusing specifically on content generation, will provide a useful framework.[21] Each stage presents opportunities for governance intervention (the below list is purely indicative and non-exhaustive).

Similarly, identifying **AI actors** across the generative AI supply chain may provide useful distinctions for future accountability mechanisms.[22] Unlike conventional physical products, characterized by a linear relationship among manufacturers, supply chains, retailers, and consumers, the AI ecosystem exhibits a complex, non-linear interplay between developers, vendors, and end-users. Consequently, the allocation of responsibility among diverse AI actors, including those involved in development and deployment, remains a critical challenge.

This document adopts the broad OECD definition of **AI actors,** encompassing ***those who play an active role in the AI system lifecycle, including organizations and individuals that deploy or operate AI.***[23] Where possible however it tries to distinguish particular roles, especially between 'upstream providers' who develop foundation models used for their generative capabilities and less technical 'downstream organizations' who deployed user-facing applications.

19 ChatGPT, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users just two months after its launch in November 2022, making it the fastest-growing consumer application in history.

20 Lee, K. et al. 2024. *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*. Journal of the Copyright Society of the U.S.A. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551.

21 Ibid.

22 Küspert, S. et al. 2023. *The Value Chain of General-Purpose AI: A Closer Look at the Implications of API and Open-Source Accessible GPAI for the EU AI Act*. London, Ada Lovelace Institute. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.

23 OECD. 2019. *Artificial Intelligence and Responsible Business Conduct*. Paris, OECD. https://web-archive.oecd.org/site/mneguidelines/RBC-and-artificial-intelligence.pdf.

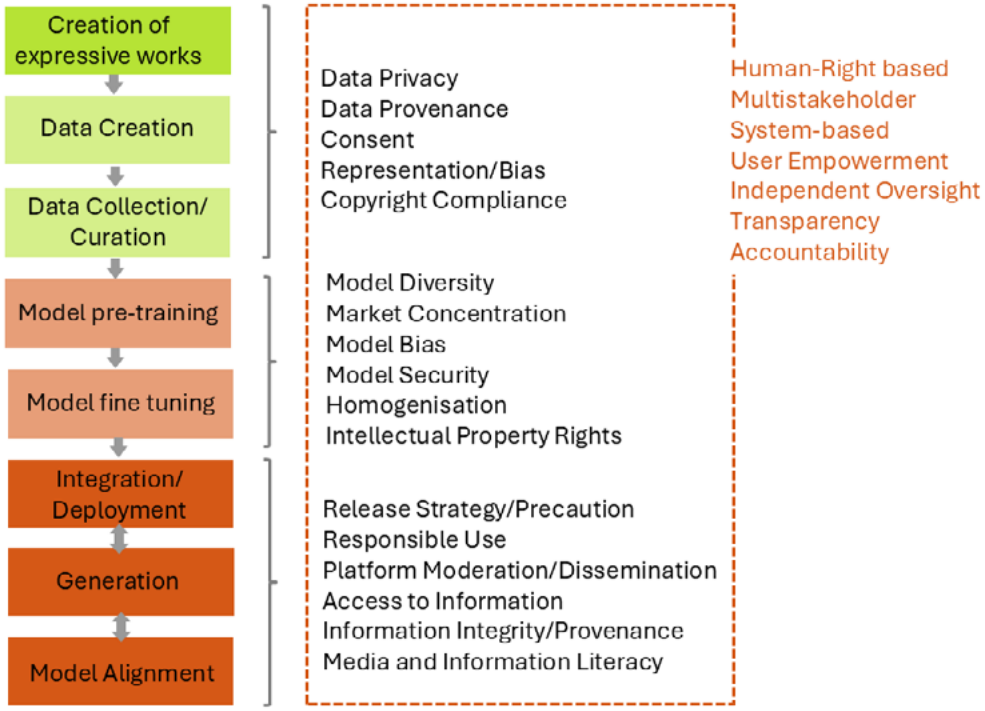



**Figure 1** - Generative AI Life Cycle and Governance Interventions

# 3. The Governance System

## 3.1 AI Governance

> *There is no shortage of documents and dialogues focused on AI governance. Hundreds of guides, frameworks, and principles have been adopted by governments, companies, and consortiums, and regional and international organizations. […] Yet, none of them can be truly global in reach and comprehensive in coverage. This leads to problems of representation, coordination and implementation (Governing AI for Humanity: Final Report, Articles xii and xiii).*

As exposed by the United Nations Secretary-General's High-level Advisory Body on AI, the governance of AI is characterized by a patchwork of norms and principles as well as significant gaps across many metrics including inclusivity, accountability, effective compliance, and transparency to name only a few.[24] Moreover, existing efforts are often disjointed, if not contradictory and have largely failed to sufficiently take into considerations long-standing, shared norms including commitments to the Universal Declaration of Human Rights.[25] To date, many sets of AI principles produced by companies, governments, civil society, and international organizations scarcely refer to human rights. Of those that do, only a small proportion engage with human rights in depth or adopt human rights as a fundamental framework.[26]

The latest advances in AI and its generative capabilities introduce additional governance challenges, including understanding the nature and potential impact of the seemingly original content it generates. While it is beyond the scope of this document to provide a detailed taxonomy of the actual and potential risks of generative AI,[27] experts generally agree on its potential to undermine information integrity and significantly affect (positively and negatively) freedom of expression and access to information.[28] Furthermore, these advances are poised to exacerbate the already vast digital divide and threaten diverse cultural content. Currently, only a limited number of AI providers, typically global technology leaders from developed countries, have the requisite skills, financial capital, and computing power to create the most sophisticated generative AI models and applications. A pervasive concern is that AI multinationals, in their rapid pursuit of innovation, may neglect the critical necessity of ensuring their technology is secure, representative, diverse, and fully aligned with international human rights standards. This apprehension materialized prominently in March 2023, when over 3,000 organizations and experts from academia, industry, and civil society collectively signed an open letter. This letter emphatically called for 'all AI labs to immediately pause for at least 6 months the training of powerful AI systems' citing profound societal risks and the need for a more deliberate and responsible approach to AI development.[29]

---

24 Compiling works that span disciplinary and geographical boundaries, the joint UNESCO and Mila publication, provides further insights for identifying and understanding current gaps in AI governance. See UNESCO/Mila. 2023. *Missing Links in AI Governance*. Paris/Montréal, UNESCO/Mila. https://unesdoc.unesco.org/ark:/48223/pf0000384787.

25 Jones, K. 2023. *AI Governance and Human Rights: Resetting the Relationship*. London, Chatham House. https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights/03-governing-ai-why-human-rights.

26 A paper reviewing 36 prominent sets of AI principles from around the world found that only 23 referred to international human rights. Only one-half of the government documents reviewed include any reference to human rights. See Fjeld, J. et al. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. Cambridge, Berkman Klein Center, Harvard University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482.

27 For a more detailed overview see Office of the United Nations High Commissioner for Human Rights. 2023. *Taxonomy of Human Rights Risks Connected to Generative AI: Supplement to B-Tech's Foundational Paper on the Responsible Development and Deployment of Generative AI*. Geneva, OHCHR. https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf; for a deep dive on experts' levels of concern about AI risks across multiple domains, see 2024. *Governing AI for Humanity, op. cit.* For major risks associated with the malicious use of generative AI and new risks that may emerge in the future, see Privitera, D. et al. 2025. *International AI Safety Report 2025.* Mila - Quebec AI Institute; Bengio, Y., and McDermid, J. A. 2024. *International Scientific Report on the Safety of Advanced AI: Interim Report*. Mila - Quebec AI Institute.

28 Concerns are diverse and range from enhanced abilities of technology to generate large-scale dis- and mis-information campaigns, to disseminate of confidently stated but erroneous or false content by which users may be misled or deceived, to facilitate the production of and access to violent, inciting, radicalizing, or threatening content.

29 In March 2023, 33,000 signatories from academia, industry and civil society signed an open letter calling on 'all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4' (GPT-4 is a foundation model upon which ChatGPT, a generative AI application, is built on). Future of Life Institute. 2023. *Pause Giant AI Experiments: An Open Letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Although many relevant governance regimes and principles predate the advent of generative AI, they are not obsolete. Rather, they require careful interpretation and application to address the specific characteristics of this new technology.

Complementing this adaptation of existing frameworks, new governance systems are also being formulated to directly regulate foundation models and their generative functionalities. This is the case for instance with the EU AI Act where a number of provisions have been added to regulate general-purpose (i.e., foundation) models, shifting the focus from specific use cases to the technology itself.[30] Similarly, in December 2024, the Brazilian Senate approved a significant new *AI bill* (Bill No. 2338/2023), which, at the time of the publication of this document, was advancing to the Chamber of Deputies. This proposed legislation adopts a comprehensive risk-based approach, aiming to impose specific obligations on developers, distributors, and applicators of high-risk AI systems. *The Bletchley Declaration,* signed by 28 countries in November 2023, also expresses a collective commitment to proactively manage the potential risks associated with highly capable general-purpose AI models.

The challenges of governing generative AI are vast and multi-faceted. The generative revolution demands agile yet inclusive governance systems which promote universal norms while taking into account different national contexts. It requires establishing common principles in a rapidly evolving landscape where societal implications of AI may only become apparent in several years. **The difficulty also lies in striking the right balance between the significant risks posed by this class of AI and its potential benefits and opportunities.** Finally, it is worth noting that humankind still lacks the distance of time and concrete evidence to identify with certainty what are the best mechanisms to effectively govern this technology. In the face of prevailing uncertainty, this document concedes that it cannot offer a definitive solution to the complex governance challenges posed by generative AI. However, it maintains that the principles articulated herein, informed by established governance frameworks, must form the bedrock of both present and future discourse on this critical issue. Fundamentally, any effective governance framework for generative AI must be rooted in a human rights-based, multistakeholder approach to effectively navigate its inherent opportunities and risks. It is unequivocally evident that governance demands a multistakeholder model, encompassing not only governmental and corporate entities but also the broader societal sphere. The current concentration of power within AI markets imperils pluralism, thereby restricting the breadth of voices and perspectives represented in digital spaces and hindering robust public debate. Consequently, online environments risk becoming dominated by content, norms, and ethical standards dictated by a limited cohort, rather than reflecting the diverse composition of the global population. Governance must aim to safeguard human rights, protect users from potential harms and maximize the benefits afforded by (generative) AI for all, while also ensuring accountability for both intended and unintended consequences of its development.

## 3.2 Digital Platform Governance

Despite its relative novelty, the governance of digital platforms is a fast-advancing field which draws lessons from decades of work in the domain of broadcast regulation, including governmental or corporate interventions that deals with content issues – regardless of the source of the content. In recent years, significant progress has been made in the development of a multistakeholder approach to shape the evolution and use of the internet and its main gateways – digital platforms.

In particular, *the UNESCO Guidelines for the Governance of Digital Platforms* are a global standard setting document to safeguard freedom of expression while dealing with content that could be permissibly restricted under international human rights standards. They propose fair, clear, and shared measures including greater transparency of technology companies and their financing; the

30 The European Commission. *The AI Act Explorer: EU Artificial Intelligence Act. Official Journal version of 13 June 2024*. Brussels, European Union. https://artificialintelligenceact.eu/ai-act-explorer.

establishment of independent regulators; the promotion and development of tools and policies that enhance critical thinking and empower users; promote measures to ensure the protection of women and girls as well as for other groups that are marginalized or in situations of vulnerability; and, above all, the safeguarding and strengthening of freedom of expression, access to information and diverse cultural content, as well as other human rights. *They can be applied, as relevant, to diverse processes that touch on the governance of content on digital platforms, regardless of form or field* (**Article 43**).

### 3.2.1 Similarities and Relevance

**While the governance dedicated to digital platforms may vary from the emerging frameworks focusing on (generative) AI, there is significant overlap**. This is due in part to digital technologies' deep interconnection: AI is increasingly deployed in the operations and systems of digital platforms, while the data generated on digital platforms (and the internet more generally) is often used to train foundation models. Similarly, both technologies are global in nature and require consistent policy responses around the world to avoid the fragmentation of regulations or approaches that compromise human rights**.** Synergies also stem from the fact that both domains exhibit similar challenges, including extreme market concentration (most of the leading AI providers are in fact digital platforms companies), a high degree of uncertainty, scale and velocity of technology, a lack of a common level of skills and standards and the necessity to balance significant benefits with societal risks and harms. Finally, the governance of digital platforms is particularly relevant when considering mechanisms to manage AI generative capabilities.

Consequently, *the Guidelines,* which outline a human rights-respecting and system-based governance system, provide a relevant and applicable set of principles to positively shape the applications of generative AI and their potential impact on freedom of expression and access to information.

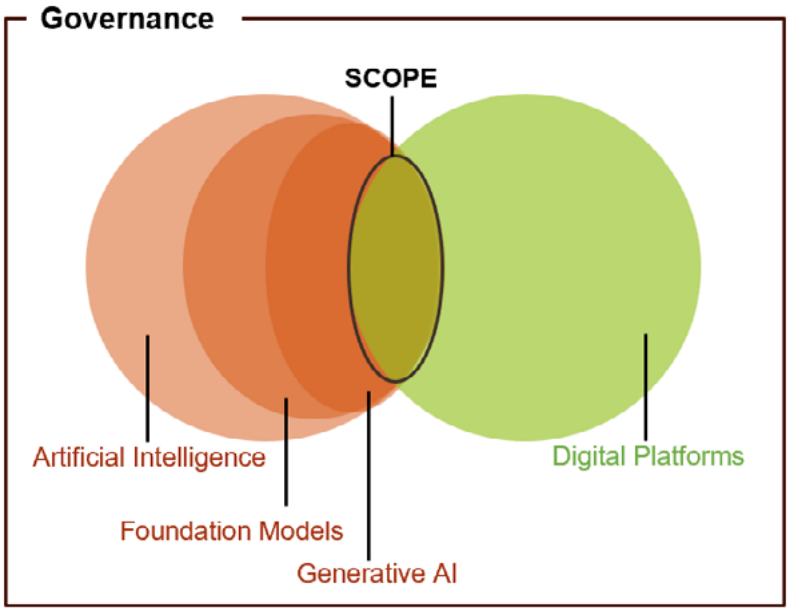



**Figure 2 –** Scope of the companion document

While significant parallels exist between the governance of digital platforms and generative AI, key distinctions warrant our attention. **Firstly, the underlying technologies** of generative AI and foundation models differ substantially from those supporting internet infrastructures and digital platforms. Consequently, seemingly similar governance provisions, such as technical solutions and engineering processes, will likely diverge considerably in their practical implementation.

Secondly, the **type of content** managed by these companies as well as related users' rights need to be differentiated. With the exceptions of initial prompts and developer work, there is limited involvement

of human users in the creation of seemingly original AI-generated content.[31] As of now, machines do not possess the right to freedom of expression or any other human rights. Furthermore, creating content through popular generative AI applications does not automatically imply the dissemination of the prompted content to a large public. By contrast, most of the content on digital platforms is expected to be user-generated, therefore engaging individuals' rights to freedom of expression. It also involves users' ability to share, widely or with their communities, their opinions as well as their literary, scholastic, or creative works.

Despite the unique characteristics of AI-generated content, its implications for freedom of expression, particularly regarding access to information, are significant. A primary application of generative AI is user-specific information retrieval. As adoption increases**, this technology is poised to become a crucial tool for accessing public interest information** – defined as publicly relevant information necessary for informed decision-making across key aspects of life, including the exercise of human rights. In the contemporary information landscape, characterized by pervasive information overload, effectively organizing, sorting, and filtering information has become critical. Consequently, generative AI's potential to provide plural, accessible, high-quality, user-centric information may become essential for the effective exercise of this right.

As with other human rights, the right to access information requires reinterpretation and adaptation for the digital age. Experts have emphasized the need for a more comprehensive understanding of this right, extending beyond the traditional focus on access to publicly held information.[32] This broader conception encompasses new dimensions such as availability, quality, stability, cultural appropriateness, agency, and usability, which may provide fruitful considerations for the governance of generative AI. Sections **4.1.1** and **4.2**, which address the responsibilities of States and AI actors, further analyze the impact of generative AI on access to information.

## 3.3 The Guidelines Approaches

*The UNESCO Guidelines for the governance of digital* platforms offer a series approaches for the governance system to effectively balance human rights and respect *the UN Charter* while preventing harms. The following section outlines some of the key characteristics presented in the document which provide a pertinent framework for governing generative AI.

### 3.3.1 Human rights-based Approach

The Guidelines recognize that *the application of rules and regulations in every governance system must adhere to human rights* which are presented as **the compass for all decision-making, at every stage and by every stakeholder.** As such, the text includes multiple and frequent references to international human rights law and standards, with a particular focus on freedom of expression and Articles 19 (3) and 20 of the ICCPR.

Too frequently overlooked, the relevance of human rights for the governance of generative AI cannot be overstated. Human rights international standards provide a robust, recognized framework. This long-standing, trialed, and tested legal system offers a shared language and universal mechanisms to identify and mitigate potential harms arising from generative AI broad adoption. They also present a legitimate basis to consider equitable distribution of AI's potential benefits.

In this regard, human rights **ought to be the starting point for normative constraints on generative AI, the baseline that new principles, obligations, and rights may appropriately complement.** The Guidelines are one such complement and provide concrete mechanisms to safeguard human rights including freedom of expression and access to information in the digital era.

31 It is pertinent to acknowledge that the datasets used to train generative AI models are, in part, comprised of human-generated content.

32 Martins, P. 2024. *Applying a Tech Lens to the Right to Information: Part 1*. Centre for International Governance Innovation (CIGI). https://www.cigionline.org/static/documents/DPH-paper-Martins_2z90GoX.pdf.

### 3.3.2 Multistakeholder Approach

*The Guidelines* call for a multistakeholder approach to the governance of technology.[33] The imperatives of multistakeholder inputs and meaningful participation are not unique to *the Guidelines* and are a recurring requirement in most international frameworks dedicated to AI. For example, *the UNESCO Recommendation on the Ethics of Artificial Intelligence,*[34] as well as the High-level Advisory Body on Artificial Intelligence, both identify adaptive multistakeholder collaboration as the root of AI governance and recommend strengthening multistakeholder policy dialogues covering this subject.[35]

The Guidelines however provide more specific guidance with regards to efforts needed to establish meaningful representation not only in terms of participation but also in c*ontributing to oversight and achieving the necessary checks and balances through institutionalized involvement and scrutiny* (**Article 69**). Similarly, they offer further considerations regarding groups which may require particular attention including **communities in situations of vulnerability and marginalization, as well as women and girls, journalists, artists, human rights defenders, and environmental defenders**. Dedicated provisions for these groups can be found throughout the text including with regards to risks assessments and due diligence (**Article 88**), mechanisms to gather testimony and specific user experiences (**Article 90**), non-discrimination (**Article 93**), and special protections (**Article 103**).

These targeted recommendations are particularly relevant for the governance of AI as most communities included in *the Guidelines* are susceptible to being disproportionately affected by AI related harms. For example, the 2023 UNESCO Issue Brief *Your Opinion does not matter any way* exposes technology-facilitated gender-based violence and significant harms resulting from the misuse of generative AI. In particular, the authors noted that 'generative AI has amplified existing methods and increased the potential avenues for technology-facilitated gender-based violence faced by many communities online.' A subsequent UNESCO study *Challenging systematic prejudices: an investigation into bias against women and girls in large language models* echoed these findings. Despite ongoing efforts, the Issue Brief reveals worrying tendencies in generative AI to produce persistent gender bias, as well as homophobia and racial stereotyping.

While multistakeholder participation is crucial for effective AI governance, achieving it presents significant challenges. The rapid pace of technological advancement and the complexity of these technologies often hinder the development of well-structured and time-intensive participatory processes. Moreover, translating diverse stakeholder inputs into concise and actionable interventions remains a substantial hurdle. Nevertheless, such participation is invaluable, as it broadens perspectives and introduces vital new indicators and sectoral nuances.

Despite these challenges, there are notable examples of initiatives fostering diverse perspectives in AI governance. *The European Union's Code of Practice on GPAI,* designed as a central resource for GPAI model providers, stands out for its highly inclusive and iterative approach. Organized by the EU AI Office, the Code is shaped by four distinct working groups, each led by global experts from law, civil society, the technical community, and academia. These groups have engaged over 1,000 stakeholders through open questions, seeking input to 'highlight areas for further progress.'

Another significant example is the African Observatory on Responsible Artificial Intelligence, established in 2022. This Observatory promotes the ethical development and use of AI technologies

33 It also worth noting that *the Guidelines* themselves were produced through an extensive multistakeholder consultation process in which UNESCO received more than 10,000 comments from 134 countries engaged in this process. This represents one of the largest and most open consultation exercises ever conducted by the Organization.

34 See for examples Article 8d (on objectives), Article 47 (on participation throughout the AI system life cycle) and Article 54 (on governance mechanisms) in UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. Paris, UNESCO.

35 UNESCO champions collaboration, uniting diverse stakeholders to enhance the quality of decision-making. This approach was integrated into the Organization's fields of competence years ago, particularly in its approach to digital technologies. This framework is built upon the principles of human rights, openness, inclusive access, multistakeholderism, and gender equality. Numerous documents within UNESCO's digital library offer detailed recommendations on multistakeholder processes. For the inclusive development of AI policies see UNESCO and Innovation for Policy Foundation (i4Policy). 2024. *Multistakeholder AI Development: 10 Building Blocks for Inclusive Policy Design*. Paris, UNESCO / Innovation for Policy Foundation. https://unesdoc.unesco.org/ark:/48223/pf0000382570.

across the African continent. Its network, comprising leaders from academia, civil society, industry, and policymakers, effectively amplifies 'African Voices' in the global AI governance discourse.

### 3.3.3 System-based Approach

The system-based approach outlined in *the Guidelines* defined stakeholders' responsibilities in accordance to processes that they need establish to prevent and mitigate technology negative externalities while promoting freedom of expression and access to information. Put simply, this approach focuses on **systems and processes rather than the nature** of each individual piece of online content (including AI-generated content). It also recognizes the **inherent interconnectedness of stakeholders**, their relations, complementarity and feedback loops.[36] Finally, it takes a holistic lens, looking at the functioning of the internet as a whole (**Article 21**).

The emphasis on systems and processes allows not only to avoid undue interference in freedom of expression but also to proactively address uncertainty. In the face of the significant unknowns surrounding the existing and potential impacts of generative AI models, it is essential to ensure that all stakeholders take proactive steps to mitigate negative externalities through the implementation of sufficient precautionary measures and safety procedures. This approach also seeks to ensure the long-term efficacy of governance frameworks. Recognizing the dynamic nature of human discourse, risks, and technologies, it focuses on the adaptability and resilience of governance mechanisms in the face of ongoing disruptions.

### 3.3.4 User empowerment and media and information literacy

The Guidelines also highlight the importance of media and information literacy (**Articles 74-84**) to ensure that all stakeholders are effectively playing their part in the governance system.

> » *Media and information literacy programmes should put an emphasis on the empowerment of users and ensure that they have the skills and knowledge that will enable them to interact with content critically and effectively in all forms of diverse media and with all information providers* ***[…]*** *(***Article 76***).*

Empowering users to critically interact with machine-generated and curated information is essential for both digital platforms and AI providers.

Research has demonstrated that digital technologies may monopolize users' attention and may risk manipulating their thoughts and opinions. Generative AI in particular increases the likelihood for humans to be influenced as well as over-rely or develop dependencies on the tools they employ. Notwithstanding the absence of malicious intent, generative AI applications have the potential to exert influence on human behavior. Furthermore, the human-like characteristics of certain classes of AI, such as chatbots, can cultivate trust and may result in the uncritical acceptance of (dis/mis) information or disclosure of personal information.

Effectively governing digital technologies is a sociotechnical issue. While technical and regulatory actions may offer solutions to mitigate risks and distribute benefits, their impact is ultimately contingent on social realities. Therefore, engaging all citizens through a whole-of-society approach, improving citizens' understanding of digital data, technology and their relationship to it as well as increasing their awareness of its pitfalls and limitations are essential to enable a safe and inclusive digital environment. To that end, multiple articles and sections of *the Guidelines* are dedicated to this issue including the responsibilities of both governments and digital platforms to pilot user-facing tools that gives them a clear understanding about the origin and context of content, support user empowerment and allocate adequate resources for global and targeted media and information literacy programmes (**Articles 78-84**). In particular, Media and Information Literacy (MIL) in the context of AI should include a strong focus on data privacy, helping users understand how their

36 See also next chapter on shared responsibilities.

personal data may be collected and used to train algorithms, and empowering them to make informed decisions in digital spaces.

Complementing *the Guidelines,* UNESCO also published *a policy brief on Media and Information Literacy* responses to the evolution of generative artificial intelligence as well as a toolkit designed to support youth organizations in integrating MIL into their strategies, policies, and actions.[37] These documents emphasize the importance of a holistic approach encompassing technical tools and off-platform investments in information literacy and capacity-building programs to foster viable and sustainable knowledge societies. This also echoes several key Recommendations on the Ethics of Artificial Intelligence. For example, **Articles 101 and 102** specifically urges Member States to provide adequate AI literacy education to the public on all levels and promote the acquisition of 'prerequisite skills' for AI education. This includes foundational abilities like basic literacy, numeracy, coding, and digital skills, alongside media and information literacy.

Finally, this reflects the consensus among practitioners and experts that online content should be evaluated within a broader contextual framework, rather than through a simplistic binary of AI-generated versus human-created content. The ability to distinguish between genuine and AI-generated content, while valuable, does not automatically validate the trustworthiness of the encompassing narrative or claim. Conversely, synthetic content can serve as a powerful medium for creative expression, facilitating artistic innovation, political satire, and humorous commentary.[38] Establishing content trustworthiness does not hinge exclusively on its perceived authenticity but also on its contextual embedding and the claims it purports. While technical solutions such as watermarking may mitigate some risks (see also section 4.2.3 on transparency of provenance), they should be complemented by media and information literacy toolkit enabling everyday users to trace and corroborate claims in a more neutral, process-oriented way. This approach emphasizes critical thinking and verification through lateral reading and corroborating evidence including context clue and historical data.[39]

## 3.4 Regulatory arrangements

» *Depending on the context, accountability and compliance mechanisms for the governance of digital platforms may include complementarity and convergency within different regulatory arrangements, such as: self-regulatory […], co-regulatory structures and mechanisms […] as well as statuary regulatory frameworks […]* ***(Article 45).***

**Universal by design, *the Guidelines* aim to apply to a wide range of regimes and acknowledge that the effective implementation of its principles can take many forms.** They do not favour any particular type of interventions, rather they recognize the diversity of possible arrangements including self-regulation mechanisms, co-regulation structures and statutory regulations. This structural versatility also reflects existing divergences among countries in their strategies to regulate platforms, with no clearly established global models or best practices.[40]

Similarly, generative AI governance today is a kaleidoscope of existing norms and emerging regulations, enforced in various manners by states, regional and intergovernmental organizations. While some governments have chosen to implement robust legal frameworks to regulate AI, others have

37 Frau-Meigs, D. 2024. *User Empowerment through Media and Information Literacy Responses to the Evolution of Generative Artificial Intelligence (GAI)*. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388547; Acero Pulgarín, S. et al. 2024. *Journey through the MILtiverse: Media and Information Literacy Toolkit for Youth Organizations*. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000392035.

38 WITNESS. 2021. *Just Joking! Deepfakes, Satire and the Politics of Synthetic Media*. https://www.witness.org/deepfakes-and-satire-report-released.

39 Google. 2024. *Determining Trustworthiness Through Context and Provenance*. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/determining_trustworthiness_en.pdf.

40 Afina, Y. et al. 2024. *Towards a global approach to digital platform regulation.* Chatham House. https://www.chathamhouse.org/sites/default/files/2024-01/2024-01-17-towards-global-approach-digital-platform-regulation-afina-et-al.pdf.

preferred industry led initiatives or opted to strengthen co-regulatory structures instead[41]. In the absence of global consensus and sufficient evidence to determine with certainty what regulatory arrangements may prove to be most effective, *the Guidelines'* flexible approach and emphasis on complementary measures and shared duties is particularly suitable to inform the governance of generative AI. While AI actors bear the responsibility for ensuring that the design, use, and deployment of this technology are consistent with human rights principles, states should focus their efforts on the enforcement of measures promoting systemic risk mitigation, transparency, accountability, and the alignment of relevant systems and processes with international human rights standards.

» *Regulatory arrangements should be effective and sustainable, taking into account the available local resources and the main priorities needing attention […] Independent oversight is needed for all forms of regulation. The process for developing regulation should be open, transparent, and evidence-based.* ***(Article 56).***

For each model, *the Guidelines* also offer specific recommendations, all of which are applicable for the governance of generative AI. For example, they recommend enhancing accountability such as independent periodic mandatory audits to assess companies' compliance with self-regulatory codes (**Articles 57 and 58**) or state-enforced penalties and funding to ensure co-regulatory structures achieve their stated objectives (**Article 59**).

### 3.4.1 Additional provisions on statutory regulations

Given the increased complexity for lawmakers to legislate on rapidly evolving technology, *the Guidelines* further elaborate on **statutory regulations (Articles 60-73**) and the process required to establish such frameworks. **It should be emphasized that *the Guidelines* do not prescribe statutory regulation universally but instead provide guidance to States on establishing such frameworks should they deem it necessary**. Two provisions in particular bear much significance for the governance of generative AI.

#### a. Independent regulatory authorities

There is considerable uncertainty about the societal implications of generative AI and experts disagree not only on its overall trajectory – whether positive or negative – but also the pace of its disruption.[42] In such unpredictable context, policy makers can only partially delineate *ex ante* rules that will effectively oversee generative AI, even in the imminent future. **Consequently, when establishing a statutory framework, the effective enforcement of the law by regulatory bodies is of equal or greater importance than the legislative process itself.** Moreover, regulating a fast-evolving, ubiquitous technology requires the continuous interpretation of legal frameworks to consider their application to new and emerging contexts. It necessitates active coordination between regulatory authorities and key stakeholders including technology companies, researchers and civil society but also between regulators themselves to provide consistent and coherent guidance across domains.

» *Statutory regulation […] should be considered only when there is independence in decision-making of the regulatory authorities involved in its implementation […].* ***(Article 60).***

**Independent regulatory oversight** is a cornerstone of the governance model presented *in the Guidelines* and therefore the subject of many provisions. In particular, readers are invited to consult **Articles 68-73** which outlines the essential characteristics of independent regulatory bodies as well as the conditions necessary to perform effectively their mandates including *sufficient technical expertise and funding not subjected to governmental discretion.*

41 G'sell, F. 2024. *Regulating Under Uncertainty: Governance Options for Generative AI*. Stanford Cyber Policy Center. https://cyber.fsi.stanford.edu/content/regulating-under-uncertainty-governance-options-generative-ai.

42 *International AI Safety Report 2025.* Mila - Quebec AI Institute. https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf; Bengio, Y., and McDermid, J. A. 2024. *International Scientific Report on the Safety of Advanced AI: Interim Report*. Mila - Quebec AI Institute. https://assets.publishing.service.gov.uk/media/6716673b96def6d27a4c9b24/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

The establishment and reinforcement of independent regulatory bodies is also one of the major efforts around the implementation of *the Guidelines*. To that end, UNESCO has facilitated the establishment of the **Global Forum of Networks** which brings together major international networks of regulatory authorities. This new body provides a space for collaboration, enabling regulators to share information globally and consider innovative mechanisms for the governance of digital platforms with a human rights-based approach. Unprecedented in its scope and objectives, the Global Forum of Networks is still a new endeavour. It may pave ways for similar initiatives around AI and may provide a valuable blueprint to foster regulators' cooperation in related domains.

## b. Differentiated and Asymmetrical Obligations

» *When defining the digital platforms that should be in the scope of statutory regulation, the regulatory authorities should identify those platforms that have relevant presence, size, and market share in a specific jurisdiction* ***(Articles 66).***

The technology sector is characterized by the lack of pluralism due to the high degree of market concentration, which is prevalent for both generative AI providers and digital platforms. All technology companies should abide by the principles contained in *the Guidelines* and respect human rights. Yet, dominant technology companies – who offer services to most internet users today – bear particular responsibilities with regards to safeguarding freedom of expression and information integrity, while dealing with content that could be permissibly restricted under international human rights standards. **Therefore, statutory obligations may require market leaders to demonstrate higher standards of transparency and accountability.** This layered approach differentiates duties for small (often national) companies with multinational platforms which are the main gateways to the internet and digital 'information points' for billions of users.

**Article 67** provides a set of criteria to define companies in scope of regulations. These characteristics such as size, market share and functionality are equally relevant when considering generative AI providers.

Current research may offer an additional and complementary criteria worth considering: **computing power**. Substantial computing power is required to train foundation models which underpin generative AI applications. At present, the majority of AI computational resources are concentrated within the private sector. **Compute governance** aims to manage access to and allocation of computing resources.[43] Given the potential impact of generative AI on markets and society, it may be beneficial to expand access to AI computing infrastructure to other sectors, such as academia, civil society, and government entities. Balancing the distribution of 'compute' is becoming a key point of intervention to address the growing AI divide.[44]

43 It is beyond the scope of this document to provide a detailed analysis of the benefits and risks of governing computing power. For further information on this subject, readers are invited to consult: Sastry, G., 2024. *Computing Power and the Governance of Artificial Intelligence*. https://arxiv.org/abs/2402.08797.

44 For example, see the OECD AI Compute and Climate group is to promote compute access. 2024. OECD. https://oecd.ai/en/compute.

# 4. An Enabling Environment Requires Shared Responsibility

» *Creating […] an enabling environment is not simply an engineering question. It is also an endeavour that calls for the engagement of societies as a whole and therefore requires whole-of-society solutions.* ***(Article 22)***

## 4.1 Joining efforts towards safeguarding human rights and mitigating technology risks

Generative AI reduces the cost and complexity of analyzing data. It impacts freedom of expression and access to information as it is widely used to seek, retrieve and adapt information to users' needs and requirements. The promise is that generative AI, particularly through chatbot interfaces, could enable new ways to access digital content and improve effectiveness of existing information retrieval systems. With its enhanced contextual understanding, generative AI could better capture the meaning of queries, providing more accurate and relevant results to users. With sound governance, generative AI tools have the capacity to transform complex or technical information into accessible formats across languages. They can, for example, render scholarly articles more understandable for wider audiences. Furthermore, they can bolster democratic and civic engagement by providing clear and concise information on critical policy matters, voting processes, and parliamentary functions. This class of AI promises to benefit education and individuals' learning by generating personalized materials and adapting its content to student learning styles and paces.

In addition, generative AI offers novel avenues for freedom of expression, extending beyond its frequently cited risks. While synthetic images or videos are often portrayed negatively, they also possess the potential to facilitate political critique, satire, parody, humour, and artistic exploration. Artists are leveraging generative AI to create new forms of art, including visual art, music, and literature, opening up creative possibilities previously unattainable with traditional tools. Furthermore, generative AI can be employed to denounce systemic brutality and violence while safeguarding vulnerable individuals. For example, Amnesty International utilized AI-generated imagery, (the images included text stating that they were produced by AI), to obscure the identities of protesters, thereby protecting them from potential arbitrary arrest and prosecution.

Despite these expected benefits, this technology can also carry significant implications for the right to privacy as it facilitates the analysis of both textual and visual content created or posted by users in the digital space, thereby potentially improving the efficiency of real-time social media monitoring and censorship. As identified in OHCHR Taxonomy of Human Rights Risks, generative AI tools supercharge existing forms of surveillance and privacy violations on a large scale.

In addition, generative AI's ubiquitous and opaque nature hinders transparency and complicates the identification of responsible parties within AI companies but also between AI providers and deployers (which can be both public or private actors). Users may find it hard if not impossible to establish which actor may be responsible for a particular use design, functionality or restriction. (See section on **accountability 4.2.4).**

Taking a holistic lens, the governance system outlined in *the Guidelines* is underpinned by the notion of shared responsibility. At a fundamental level, it recognizes that the efforts required to safeguard human rights and mitigate technology risks are not one sided. **While some duties may fall on AI actors, others such auditing and assessing, promoting cooperation, improving media and information literacy are imparted predominantly to different actors, users and non-users.** In other words, if corporate and technology driven solutions are a necessary condition, technical improvements and private sector led initiatives alone are insufficient to ensure freedom of speech,

access to information and a safe and secure digital environment. Thus, *the Guidelines* also focuses on the roles of the state, international organizations (**Articles 33-34**), media, civil society, academia and other stakeholders (**Articles 35-41**) involved in the governance of digital platforms.

Notwithstanding this document primarily focuses on state and corporate interventions which remain controversial both at the international and national levels and highlights the key role of civil society, independent researchers and other stakeholders to advance a human rights and evidence-based approach of the governance of generative AI. The next section is dedicated to reviewing *the Guidelines'* principles and their application to different stakeholders but particularly to governments and companies efforts to govern generative AI.

## 4.2 The role of civil society, independent researchers, and other stakeholders

The active engagement of every stakeholder in the governance of digital platforms and the safeguarding of freedom of expression, access to information, and other human rights constitutes a fundamental pillar of *the Guidelines.* A series of **Articles (35-41)** specifically addresses and underscores their vital importance, further complementing the recommendations on multistakeholder processes outlined in section 3.2.3.

Just as with digital platforms, the contributions of these groups are paramount for the effective governance of generative AI. In an industry defined by homogeneity, centralization, and inherent imbalances, civil society, advocacy groups, journalists, researchers, and other independent bodies are indispensable. They play a vital role in comprehending generative AI's impact – especially on vulnerable and marginalized populations, women and girls, journalists, artists, and human rights defenders – while offering much-needed alternative perspectives, and diligently monitoring and reporting on corporate policies and government actions that affect human rights.

Moreover, in light of the considerable uncertainty and the critical gaps in understanding the societal and local impacts of these technologies, these groups play a crucial role in establishing robust evidence on generative AI's capabilities, opportunities and risks. Specifically, **independent researchers** ought to be a cornerstone of generative AI governance. Their unbiased contributions are vital for delivering timely, impartial, and scientific knowledge, which is necessary to counteract the information asymmetries that currently exist between well-resourced AI companies and the broader global community. Ultimately, global instruments like this one rely on burgeoning research and scientific inquiry to provide sound and balanced recommendations.

***The Guidelines* underscore the importance of researchers** through articles dedicated to supporting their role, including provisions for data access to non-personal and pseudonymous data held by digital platforms for researchers, journalists, and advocacy groups **(Articles 116-118).** Further reinforcing this, *the UNESCO Recommendation on the Ethics of Artificial Intelligence* also includes recommendations on access to and interdisciplinarity of AI research **(Articles 109-111).**

Finally, the UN High-Level Advisory Body on AI has proposed establishing an **independent international scientific panel on AI**. This panel would be composed of diverse, multidisciplinary experts serving voluntarily in their personal capacity. With support from the proposed United Nations AI office, relevant UN agencies, and other international organizations, its key mandates would encompass issuing an annual report surveying AI-related capabilities, opportunities, risks, and uncertainties. It would also be responsible for producing thematic research digests on how AI can contribute to achieving the SDGs, and for identifying areas of scientific consensus on technology trends and areas requiring additional research.

## 4.3 States duties to respect, protect, and fulfil human rights

States duties presented in *the Guidelines* reflect the governance approach outlined in the previous chapter, including respect and promotion of human rights (**Articles 26**), the necessity to consider multistakeholder perspectives and establish special protections for communities (**Article 28 b-g**) as well as the establishment of sufficient mechanisms to ensure checks and balances including independent regulatory functions (**Article 28m**).

### 4.3.1 Protection of Freedom of Expression and Access to Information

The Guidelines provide relevant considerations regarding state constraints on access to information including **Articles 29a,** which highlights that *States should refrain from imposing measures that prevent or disrupt general access to the dissemination of information*. However, pressures on governments to suspend popular generative AI applications have grown. To date, several countries have instituted an explicit prohibition against these tools. While it is not the aim of this document to evaluate the compatibility of restrictive measures with international standards for legality, legitimacy or proportionality, it is worth noting that. with the increasingly broad adoption and integration of this technology in education and business operations, restricted access to generative AI tools or blanket ban are likely to cause substantial disruption to the dissemination of information as well as significant collateral damages.

Moreover, public bodies' transparency on how they use and deploy generative AI, their interactions with technology actors but also the requirements they impose on AI providers is an essential component for the effective governance of this technology. In this regard, **Article 27** stipulates that states *have an obligation to be fully transparent and accountable about the requirements they place upon digital platforms,* ensuring legal certainty and legal predictability, which are essential preconditions for the rule of law, as well as **Articles 28i** – *States should ensure that any restrictions imposed upon platforms consistently follow the high threshold set for restrictions on freedom of expression, based on the application of Articles 19 (3) and 20 of the ICCPR, respecting the conditions of legality, legitimate aim, necessity, and proportionality* offer relevant recommendations for government interactions with all technology providers.

### 4.3.2 Protection of artistic and journalistic content

**Articles 28f** on state responsibilities to recognize the **importance of artistic works** for the renewal of cultural production as well as Article **28g** on state duties to guarantee **users' rights to privacy and data protection** deserve particular attention for the governance of generative AI.

In recent months, significant concerns have arisen from the extensive use of large datasets, often sourced through web scraping, to train generative AI (foundation) models. In many cases, this has been shown to incorporate and regurgitate large amounts of copyrighted content which includes artistic and journalistic works or personal information. With regards to personal data, generative AI applications may accidentally memorize, aggregate or leak sensitive information.[45] Furthermore, the underlying patterns and structures within datasets could potentially be exploited by malicious actors to purposely extract individuals' details. Regarding copyrighted content, numerous lawsuits worldwide allege infringement by generative AI models partially trained on proprietary text and images. In certain instances, these models have been reported to generate outputs strikingly

45 Data protection authorities have cautioned against the use of web scraping techniques through which individuals may lose control of their personal information when these are collected without their knowledge, against their expectations, and for purposes potentially different from those of the original collection. They also highlight that web scraping may not comply with relevant data protection principles, including data minimization and accuracy, insofar as there is no assessment on the reliability of the sources. See European Data Protection Supervisor, op. cit.

similar to existing creative works.[46] These cases have highlighted the urgent need for governments to update intellectual properties regimes to clarify ownership of AI-generated content and ensure that creators' rights are protected. As many experts noted, current intellectual property laws are increasingly ill-suited to account for content generated through the synthesis of vast quantities of existing data and content. This fundamental mismatch highlights the urgent need for a new conceptual framework to define the economic relationship between source content and the novel content generated by AI systems.

This document does not aim to provide definitive answers to these questions or to speculate on the outcomes of ongoing litigation. Its aim however is to ensure that the governance systems, which will ultimately legislate on these issues, exhibit a human rights-based approach which emphasizes both the need for protection of freedom of expression, intellectual property and the promotion of cultural expressions while enjoying the benefits of scientific progress and its applications.[47] To that end, several multistakeholder initiatives are worth mentioning such as *the 2023 Principles for Fair Compensation.*[48] These principles are intended to help in the design, implementation and evaluation of public policy mechanisms that encourage technology companies and news publishers to engage with each other to develop fair economic terms. Readers are also invited to consult the UNESCO recent *Issue Brief on AI and the future of journalism* which offers valuable recommendations on this subject including establishing transparent frameworks and standards for collaboration with publishers and creators with a focus on diversity and inclusivity (including cooperation beyond dominant, English-speaking outlets).[49]

### 4.3.3 Plurality and Diversity

States must guarantee and actively defend freedom of expression in all its diversity, and encourage greater plurality of datasets and outputs as well as diverse ecosystems of AI developers and deployers. In particular, states should promote and invest in (across the AI supply chain from data collection to applications) efforts to support languages and cultures that are under-digitalized and poorly represented on the web. This notion is also strongly underscored in *the UNESCO Recommendation on the Ethics of Artificial Intelligence.* For example, Article 67 explicitly states that Member States should implement mechanisms to promote and increase diversity and inclusiveness that reflect their populations within AI development teams and training datasets.

This state responsibility arises from the fact that most generative AI systems are trained on extensive datasets predominantly composed of English-language content and data originating from countries in the Global North (see also section 4.2.2 on bias and discrimination). This practice perpetuates the values, interests, and biases of dominant cultural and political groups, effectively marginalizing the linguistic, historical, and cultural diversity of other communities. The issue is exacerbated by the lack of diversity due to the current market concentration, particularly concerning advanced AI models, where a limited number of private AI actors wield control over the essential data, computational resources, and infrastructure for training such models. This centralization erodes pluralism, constricting the range of voices and perspectives represented in digital spaces and impeding robust public discourse. Consequently, online environments increasingly reflect the norms and ethics of a select few, rather than the diverse composition of the global population.

---

46 Current generative AI models, due to their capacity to memorize fragments of their training data, may inadvertently reproduce these fragments. In certain instances, these models can even go beyond mere reproduction of fragments. Researchers at Stanford University demonstrated that a chatbot could be induced to regurgitate substantial portions of popular literary pieces. See Henderson, P. et al. 2023. *Foundation Models and Fair Use*. Preprint, Stanford University. https://arxiv.org/pdf/2303.15715.

47 United Nations General Assembly. 2015. *Report of the Special Rapporteur in the field of cultural rights: Copyright policy and the right to science and culture* (Doc. A/HRC/28/5.) New York, United Nations.

48 These Principles were adopted by participants at a conference held at the Gordon Institute of Business Science (GIBS) in Johannesburg, South Africa, on 14 July 2023. See Global Forum for Media Development. 2023. *Big Tech and Journalism: Principles for Fair Compensation*. https://gfmd.info/engagements/big-tech-and-journalism-principles-for-fair-compensation.

49 Schiffrin, A. 2024. *AI and the Future of Journalism: An Issue Brief for Stakeholders*. Paris, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000391214.

Furthermore, some of the principles articulated earlier, such as the concept of **asymmetrical requirements** (section 3.4.1.b), which imposes greater obligations on market leaders than on smaller (often local) entities, mechanisms for managing access to and allocation of **computing resources** or more generally facilitating **access to AI assets** (expertise, compute and data) for researchers and experts working with low-resources and minority languages, offer valuable frameworks for state consideration.

**Promoting open solutions** and expanding access to both data sets and foundation models may offer additional mechanisms for promoting diversity and plurality.[50] **At the data level,** experts have highlighted that equitable access to data is essential for nurturing a just, informed, and inclusive society in the digital era. Regarding the UNESCO Recommendation on the Ethics of Artificial Intelligence, for instance, Article 75 is specifically dedicated to the promotion of open data. It urges Member States to consider mechanisms such as open repositories for publicly funded or publicly held data and source code, as well as data trusts, to support the safe, fair, legal, and ethical sharing of data. Another example includes the African Commission on Human and People's Rights which advocates for states to ensure that 'data held by public institutions, as well as that held by private actors where there is an overriding public interest in access, should be made publicly available by default, in alignment with the principle of maximum disclosure, except where justified by regional and international *human rights standards*.[51] **At the model level**, states could consider open-source models to enable users to download, modify, and share models or their parts.[52] In particular, open sourcing enables broader participation in AI development, facilitating large-scale collaboration. It leverages diverse expertise, perspectives, and collective effort to enhance AI safety research, and expand the frontiers of AI capabilities for more regional and diverse uses.

While open access fuels research, innovation, and AI safety, fostering transparency, collaborative flaw detection, it also presents risks. Increased participation and diverse perspectives can drive beneficial innovation for more communities, but open models can be exploited for malicious purposes, with limited developer oversight and irreversible distribution. Recognizing this, a consensus is emerging to assess the 'marginal' risk of open-source models, comparing their potential danger to existing closed models and alternative technologies.[53] This document does not delve into the specific applications of open-source models; however, it's vital to note that numerous research, scientific and international organizations have formulated important guidelines and principles for responsible open foundation models, particularly regarding state approaches to these technologies.[54] For example the UN Open Source Principles are comprised of eight guidelines and provide a framework to guide the use, development and sharing of Open Source software.

### 4.3.4 Civil servant code of conduct and procurement

» *States should strongly discourage – including through measures such as professional codes of conduct – public officials from disseminating disinformation, including gendered disinformation; misinformation; and intimidating or threatening the media.* ***(Article 28j)***

50 Open solutions also aligned with other initiatives championed by UNESCO, such as Open Educational Resources (OER), Open Access (OA) to scientific information, Free and Open Source Software (FOSS), and Open Data, which have long facilitated the free flow of information and knowledge.

51 African Commission on Human and Peoples' Rights. 2024. *Resolution on Promoting and Harnessing Data Access as a Tool for Advancing Human Rights and Sustainable Development in the Digital Age* (Doc. ACHPR/Res. 620 (LXXXI) 2024.) https://achpr.au.int/en/adopted-resolutions/620-data-access-tool-advancing-human-rights-and-sustainable-development.

52 Some state-of-the-art general-purpose AI Capabilities of general-purpose AI 1.1 How general-purpose AI is developed models, such as GPT-4o, are on the closed end of the spectrum, while others sit more towards the open end of the spectrum. For example, Llama-3.1 has 'open' weights that are available for public download

53 Privitera, D. et al. 2025. *International AI Safety Report 2025.* Mila - Quebec AI Institute; Bengio, Y. et al. 2024. *International Scientific Report on the Safety of Advanced AI: Interim Report*. Mila - Quebec AI Institute.

54 See Seger, E. et al. 2023. *Open-Sourcing Highly Capable Foundation Models*. Centre for the Governance of AI. https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf; Klyman, K. 2023. *How to Promote Responsible Open Foundation Models*. Stanford HAI. https://hai.stanford.edu/news/how-promote-responsible-open-foundation-models; Srikumar, M. et al. 2024. *Risk Mitigation Strategies for the Open Foundation Model Value Chain*. Partnership on AI. https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/.

Generative AI tools could profoundly reshape the nature of civil servant work, civic engagement and public service delivery. Several countries are already leveraging such technologies to make services more efficient, seamless, and integrated. These benefits, however, need to be carefully balanced with generative AI's significant risks for information integrity and discrimination as well as uncertainty regarding copyright and privacy compliance. In line with **Article 28j**, governments may consider developing further guidance to clarify how civil servants should approach the use of generative AI in their official functions.[55]

This could include for example the necessity to systematically verify the accuracy of generative AI outputs for bias and misinformation or avoid inputting any prompt that could include classified or sensitive information or that reveals the intent of government.[56]

It is also worth noting that a number of actors including civil society and human rights experts have called for governments to develop **AI-specific public procurement guidelines** to ensure that AI models and applications used by public bodies respect human rights and due processes.

## 4.4 The responsibilities of technology companies to respect human rights

The vast majority of *the Guidelines'* principles focused on digital platforms are directly applicable to generative AI companies and require little to no explanation about how they may be translated to AI governance. Readers are therefore encouraged to consult the original text for a more granular understanding of its content (**Articles 85-144**). They are also summarized in Article 30:

*Digital platforms should comply with five key principles:*

- ***Platforms conduct human rights due diligence**, assessing their human rights impact, including the gender and cultural dimensions, evaluating the risks, and defining the mitigation measures.*
- ***Platforms adhere to international human rights standards** […]. Design should ensure non-discrimination and equal treatment and that harm is prevented; **content moderation and curation policies and practices** should be consistent with human rights standards, whether these practices are implemented through automated or human means, with knowledge of local languages and linguistic context as well as respect for cultural diversity, and adequate protection and support for human moderators.*
- ***Platforms are transparent and open about how they operate** […]. This includes transparency about the tools, systems, and processes used to moderate and curate content on their platforms, including in regard to algorithmic decisions and the results they produce.*
- ***Platforms make information accessible** for users to understand the different products, services, and tools provided, and to make informed decisions about the content they share and consume […].*
- ***Platforms are accountable to relevant stakeholders** […] they give users the ability to seek **appropriate and timely redress** against content-related decisions, including both users whose content was taken down or moderated and users who have made complaints about content.*

Rather than paraphrasing *the Guidelines'* articles, the following section highlights areas where additional considerations and further investigations may be required as well as where generative AI could offer potential benefits for the governance system itself.[57] It also directs readers' attention

55 To that end, UNESCO launched the Schools of Public Administration and Actors for Research and Knowledge on AI (SPAARK-AI Alliance) in 2025. This global network collaborates on training civil servants in artificial intelligence (AI) and digital transformation. https://www.unesco.org/en/articles/strengthening-public-sector-ai-capabilities-worldwide-unesco-and-partner-schools-public.

56 The UK Government Digital Service. 2024. *Guidance to Civil Servants on Use of Generative AI*. London, UK Government. https://www.gov.uk/government/publications/guidance-to-civil-servants-on-use-of-generative-ai/guidance-to-civil-servants-on-use-of-generative-ai.

57 These areas were identified and suggested during expert interviews and consultations with stakeholders.

to other AI governance instruments such as *the UNESCO Recommendation on the Ethics of Artificial Intelligence* further outlining complementarity between both fields of governance.[58]

### 4.4.1 Generative AI and Online Content

As mentioned in the previous chapter, AI-generated content does not require the significant contribution of a human author. This does not mean however that AI companies' content policies do not have profound implications on freedom of expression and access to information in particular.

The following section focuses on (generative) AI developers' and providers' content decisions and processes. Specifically, it considers corporate policies on two categories:

A. Synthetic content that AI systems generate in response to a human prompt and measures to restrict specific types of AI-generated outputs.

B. Content moderation systems based on generative AI (foundation) models used to moderate content published by users hosted on digital platforms.

It is important to note that these categories do not constitute an exhaustive list of generative AI impacts on online content. Generative AI is still a nascent field, and much uncertainty remains with regards to the extent of its societal implications. This document focuses on two specific outcomes and relations with regards to content moderation as summarized in the graphic below.

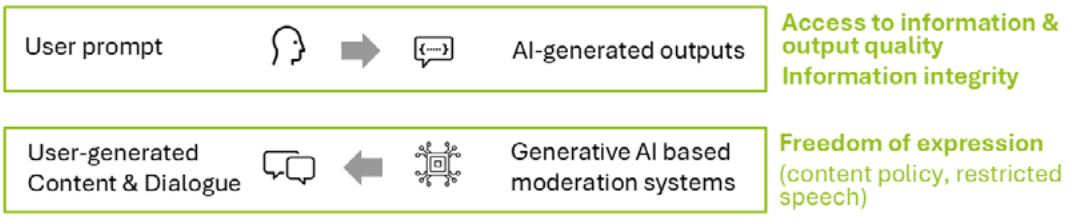



**Figure 3 –** Generative AI-based moderation systems.

#### A. Synthetic content that AI systems generate in response to a human prompt

Considerations surrounding AI-generated content in relation to freedom of expression can be broadly analyzed according to two lenses: one pertaining the restrictions to access to information because of the outputs' quality, and then the broader impact on information integrity due to information manipulation and provenance.

» *When considering measures to restrict content, platforms should take into account the conditions on legitimate restrictions to freedom of expression as laid out in Article 19 (3) of the ICCPR, and the prohibition of advocacy to hatred that constitutes incitement against discrimination, hostility, or violence as laid out in Article 20 (2) of the ICCPR, including the six-point threshold test for defining such content outlined in the Rabat Plan of Action* **(Article 99).**

##### Access to information and outputs quality

Leading providers of generative AI applications already restrict a range of outputs, including those that they consider may increase their potential for harm, which in some cases can be content deemed sexually explicit, graphically violent, or hateful.[59] Although specific policies vary across the industry, companies may also limit content that is legally permissible but politically or culturally contentious. Early research suggests that **most popular chatbots lack precise definitions with**

58 UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence.
59 Google. 2024. *Generative AI Prohibited Use Policy*. Google. https://policies.google.com/terms/generative-ai/use-policy.

**regards to the categories they limit and use broad or open-ended restrictive clauses.**[60] According to experts, generative AI providers have generally adopted more restrictive usage policies than most social media platforms, even though – unlike social media posts – the output of chatbots is not automatically disseminated to the public, but requires users to introduce specific requests.[61]

There is a number of caveats to these findings. While generative AI users have the right to access information, **generative AI models have, thus far, displayed a tendency to sometimes produce outputs that are inaccurate or nonsensical**. This phenomenon in which generative AI will, spontaneously, produce 'confidently stated but erroneous or false content', is sometimes referred to as 'confabulation', a term some prefer to the terms 'hallucination' or 'fabrication'.[62] Another well documented issue with chatbots is their inventing fake sources or providing inaccurate references. These inaccuracies and 'hallucinations' can have damaging effects, particularly when it implicates real individuals.[63] Equally, the questions of bias and discrimination are persistent across chatbots which are addressed in the following section. **Currently, none of the existing moderation techniques used to curate AI generated outputs are foolproof or comprehensive**.[64] The generative AI industry thus faces a difficult trade-off between offering unrestricted access to AI-generated information and exposing users to potentially inaccurate or harmful content.[65]

Notwithstanding these inherent limitations, freedom of expression experts have encouraged AI **providers to clarify their terms and provide reasons justifying restrictions and the potential harms that prohibited content may cause.** In line with *the Guidelines'* transparency principle (**Article 115 a-d**), these efforts could also contribute to enhancing users' media and information literacy. For example, instead of generic replies – such as 'I am unable to help you with that' – to problematic prompts, **AI providers could further invest in 'push back' outputs and offer users with explanations on why their query was not processed.** This would enable users to better understand why certain prompts are not permitted as well as the inherent limitations of the AI systems they interact with. This also mirrors good practices in the governance of digital platforms *which should notify users when their content is removed and the reason behind it. This would allow users to understand why that action on their content was taken, the method used (through automated means or after human review), and under which platform rules action was taken* **Article 110.**

### Information manipulation and provenance

The convergence of generative AI, which has augmented users' ability to produce harmful synthetic or manipulated content, and digital platforms, which facilitate its widespread dissemination, has led to the unprecedented proliferation of misinformation and disinformation at a scale and speed which jeopardizes the integrity of information ecosystems.

An ever-growing percentage of online text, images, audio, and video reportedly originates or is altered by generative AI systems. As a result, there is increasingly a lack of clarity about the authorship, source, and authenticity of digital content (see also section 4.1.3.). **Transparency on the provenance** of digital content is becoming a significant concern for both AI providers which are responsible for the technology that generates synthetic content and digital platforms which bear the duties to monitor content that they distribute**.**

60 Calvet-Bademunt, J. and Mchangama, J. *Freedom of Expression in Generative AI – A Snapshot of Content Policies.* The Future of Free Speech. https://futurefreespeech.org/report-freedom-of-expression-in-generative-ai-a-snapshot-of-content-policies/.

61 Further analysis also considers that the Rabat Plan of Action, is a key global standard that introduced a six-part test to determine hate speech and provides guidance on how to balance freedom of expression and incitement to hatred. One of the six elements in the test – the extent of the dissemination of content – is likely less worrying in generative AI than in social media.

62 Autio, C. et al. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

63 One specific example commonly used to illustrate this challenge includes ChatGPT references to a nonexistent sexual harassment case.

64 This includes, for examples, filtering training data, fine-tuning models with human feedback, limiting input queries, and filtering output content.

65 Note that this excludes content that is permissibly restricted under international human rights law and standards.

- **AI providers** should deploy technical measures to label and identify content created by machines. The use of watermarking,[66] fingerprinting,[67] markup and metadata[68] to signal content provenance can help detect content that may be partly or wholly synthetic. Provenance information should also be integrated into product design, transparency requirements (see also section 4.2.3. on transparency) and industry standards including how content was generated and the tools that were employed. In recent years, content provenance has been a promising area of collaboration across the content creation and distribution ecosystem such as the Coalition for Content Provenance and Authenticity (C2PA). Furthermore, these measures may also offer significant benefits for media and information literacy efforts. Clear, user-facing labels providing provenance information can empower individuals to make informed choices about the content they consume. Finally, content provenance techniques offer authors and journalists a mechanism for tracing and identifying instances of unauthorized content use. By embedding key information about the content creator – including the author's name, publication date, and copyright details – directly within the digital material, watermarks provide evidence of ownership. However, no technical solution exists that can reliably detect and appropriately watermark or label all synthetic content, especially when it comes to plain text. Provenance methods remain limited, susceptible to removal or modification by skilled users, and prone to false positives.[69] Consequently, numerous stakeholders have called for further research within the AI community to develop more robust methods for distinguishing AI-generated content.
- **On the side of digital platforms,** experts have recommended to adopt proactive solutions for identifying falsified content including automated checks for watermarks or labeling deepfakes and AI-generated posts (a provision which is also included in *the Guidelines* under **Article 115o**). They have also highlighted the importance of developing more diverse, robust and accessible mechanisms to encourage users, non-users or third parties to flag and identify (harmful) synthetic content. The critical function of reporting is identified in *the Guidelines* under **Articles 123 to 125**.

As mentioned in the previous chapter, technical solutions, while essential, cannot sufficiently address ubiquitous threats to information integrity. Users' engagement and enhancing media and information literacy offer a relevant and complementary mitigation strategy. For example, some AI providers are encouraging their users to disclose AI involvement in the content they share. Similarly, technology companies should proactively inform audiences when they are interacting with a chatbot. Beyond these corporate initiatives, the United Nations Global Principles for Information Integrity offers further recommendations for multistakeholder action.[70]

Finally, it is paramount to recognize that information integrity is particularly critical within the **context of electoral processes.** Freedom of expression and access to information are foundational to democratic governance. While challenges to information integrity have always been present during political contests like elections, they have intensified in speed, sophistication, and potentially scale, largely due to the capacities recently introduced generative AI. This reality is thoroughly explored in *the UNESCO Issue Brief: Freedom of Expression, Artificial Intelligence and Elections*, which provides a comprehensive overview of how AI influences freedom of expression in electoral contexts.[71] The brief offers practical approaches for practitioners and partners to navigate this fast-evolving

66 Watermarking refers to the process of embedding unique and detectable signals in AI- generated content. Watermarking could also be used to indicate genuine content and provide authors with a method to trace and identify unauthorized use of their content. For example, watermarks can embed crucial information about the content creator, such as the author's name, publication date, and copyright details, directly into the digital material.

67 Fingerprinting refers to the process of converting a piece of content and relevant information about it into a compressed sequence of numbers (the 'fingerprint') that is stored in a database.

68 Markup allows to include information about a piece of content in its metadata, such as its date of capture, generation or editing.

69 For more information of the limitations of existing methods see Privitera, D. et al. 2025. *International AI Safety Report 2025.* Mila - Quebec AI Institute; Bengio, Y. et al. 2024. *International Scientific Report on the Safety of Advanced AI: Interim Report*. Mila - Quebec AI Institute.

70 United Nations. 2024. United Nations Global Principles for Information Integrity. https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf.

71 Patel, A. 2025. *Freedom of Expression, Artificial Intelligence and Elections*. Paris, UNESCO / UNDP. https://unesdoc.unesco.org/ark:/48223/pf0000393473.

environment, highlighting both the opportunities and the serious risks, such as the proliferation of disinformation, deepfakes, and the amplification of hate speech, all of which can significantly undermine democratic processes and public trust.

### B. Content moderation systems based on generative AI (foundation) models

Existing automated content moderation systems have drawn significant criticism for failing to effectively identify harmful content or appreciate the nuance, irony or context of human speech. According to some experts, (LLMs-based) generative AI has the potential to significantly improve automated content moderation on digital platforms. The promise is that it could effectively handle not only content moderation decisions but also content policy development.[72] Allegedly, generative AI could enable more consistent labeling, faster feedback loops for policy refinement and reduce the need for human moderators.[73]

Nonetheless substantial concerns persist about the models' capacity to effectively account for cultural diversity, including linguistic variation and contextual nuances, as well as to adopt their policies and processes to regional human rights standards (see sections 4.1.4 on plurality and diversity and 4.2.2 on bias and discrimination.)

These risks mean that external scrutiny and third-party audits of content moderation activities are vital. Auditing is an essential mechanism to ensure accountability for content choices as highlighted in **Article 103**. *Digital platforms should commission regular external audits, with binding follow-up steps, of the automated and human tools used for content moderation, curation, and recommender mechanisms for their precision, accuracy, and for possible bias or discrimination across different content types, languages, cultures, and contexts […].* Furthermore, as generative AI models could be prompted to provide explanations for every moderation decision, they can themselves further facilitate the auditing process.

Similar to content moderation on digital platforms, generative AI companies also rely on human annotators to label large amounts of deleterious content in the original dataset. While this practice occurs at an earlier stage in the process compared to social media and may not necessitate human involvement as frequently and systematically, it still requires sufficient protection and training for the individuals responsible for classifying content. Here again *the Guidelines* provide relevant directions which can be found in **Article 101**: *Human content moderators, whether employed by platforms directly or hired as outside contractors through outsourced roles, should be adequately trained, fluent in the language(s) used on the platforms and familiar with local linguistic and cultural contexts, evaluated, vetted, and psychologically supported. Platforms should further put in place* ***well-funded and well-staffed support programmes for content moderators to minimize harm caused to them through their reoccurring exposure to violent or disturbing content while at work***.

## 4.4.2 Bias and discrimination

» *Digital platforms should ensure non-discrimination and equal treatment in their design processes, as well as in their content moderation and curation policies, practices, and systems. This encompasses addressing biases, stereotypes, and discriminatory algorithms or content moderation practices that affect women and girls, as well as groups in situations of vulnerability and marginalization, including indigenous communities.* ***(Article 93)***

72 In 2023, OpenAI proposed the use of GPT-4 for content moderation, highlighting its potential for efficient classification of harmful content and reducing reliance on human moderation. See Open Ai. 2023. Using GPT-4 for content moderation. https://openai.com/index/using-gpt-4-for-content-moderation/#LilianWeng.

73 Barrett, P. M. and Hendrix, J. Is Generative AI the Answer for the Failures of Content Moderation? *Tech Policy Press.* https://www.techpolicy.press/is-generative-ai-the-answer-for-the-failures-of-content-moderation/.

Whether prompted or spontaneously, AI-generated content has been shown to exhibit bias, leading to skewed, discriminatory or distorted outputs. This issue is consistent across all modalities including text, images, audio, and video. These flaws often mirror the limitations of the underlying datasets and often appear to be persistent in time and across applications.

AI-enabled discrimination and bias can take many forms which often intersect,[74] affecting communities and cultures across the world. As mentioned in the previous chapter, AI discriminatory tendencies severely impact women and girls and may reinforce gender bias or increase avenues for technology-facilitated gender-based violence. Additionally, the widespread use of foundational models may lead to 'outcome homogenization,'[75] where prejudices present in the training data are perpetuated across a broad spectrum of AI applications. For example, generative AI outputs often overrepresent dominant cultural groups, potentially leading to the misrepresentation or underrepresentation of other groups at scale.[76] Typically, 'low-resource' or minority languages are poorly represented, if at all, in datasets, resulting in suboptimal performance of generative AI applications for speakers of these languages. This disparity in performance may exacerbate existing digital divides between linguistically, geographically, and culturally diverse populations within and between nations.[77] It can further solidify detrimental biases and potentially institutionalize systemic exclusion, limiting the ability of communities or individuals in situation of vulnerability or marginalization to exercise control over the representation of their identities in media and across the internet.

In responses to these challenges, AI providers and researchers have deployed a variety of methods to mitigate or remove bias and improve fairness in AI systems. It is beyond the scope of this document to detail these technical solutions, however it is essential to note that complete and global fairness in AI systems may not be attainable.[78] According to independent experts and as highlighted in the International Scientific Report on the Safety of Advanced AI: '*establishing general-purpose AI systems that are fair across all measures and across different cultural, social and scientific contexts remains challenging.* ***No existing measure can entirely eliminate all potential risks of bias and unfairness which are inherent in the development of highly capable AI systems'***.[79]

This somewhat somber conclusion is nevertheless counterbalanced by scientists' and experts' call to further improve current models and strive for the development of fairer, if not fair, generative AI applications. Bias in its various forms should be addressed comprehensively and systematically, requiring specialized detection and risk assessments. To that end, *the Guidelines* provide useful guidance with regards to instances which may require **additional human rights risk assessments,** including prior to any significant design changes or in responses to emergencies or crises (**Article 88**). Furthermore, in line with *the Guidelines'* **multistakeholder approach**, scientific evidence has demonstrated that meaningful representation and participation in the AI life cycle can help reduce risks of discrimination. It also stresses the importance of '*creating spaces to listen, engage, and involve users, including those who have experienced harassment or abuse, their representatives, and users from groups in situations of vulnerability and marginalization* […]' (**Article 89 and 90**) and '*effective and accessible complaints mechanisms for members of groups in situations of vulnerability and marginalization'* (**Article 123**).

---

74 As indicated in the 2023 International Scientific Report on the Safety of Advanced AI, intersectional bias remains difficult to address. For instance, a generative AI system might be fair to Asians and women separately, yet be biased toward Asian women.

75 Bommasani, R. et al. 2022. *Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?* Cornell University. https://arxiv.org/abs/2211.13972.

76 Aldahoul, N. et al. 2024. AI-generated Faces Influence Gender Stereotypes and Racial Homogenization. Consensus. https://consensus.app/papers/aigenerated-faces-influence-gender-stereotypes-and-aldahoul-rahwan/2545a38cd56c5faeb0778ea572f317da/.

77 United Nations Human Rights. 2024. Taxonomy of Human Rights Risks Connected to Generative AI. B-Tech. United Nations Human Rights Office of the High Commissioner. https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf.

78 Mathematical results suggest that it may not be possible to satisfy all aspects of fairness simultaneously under reasonable assumptions. This impossibility theorem of fairness is supported by results indicating the complexity of training unbiased general-purpose AI models.

79 Considering that no existing techniques can fully ensure non-discrimination and equal treatment in generative AI models, *UNESCO Recommendation on the Ethics of Artificial Intelligence* may offer useful alternatives to *the Guidelines*, including the necessity for AI actors to *'make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system.'* (Article 29). UNESCO. 2021. Recommendation on the Ethics of Artificial Intelligence. UNESCO.

In addition to *the Guidelines,* ***the UNESCO Recommendation on the Ethics of Artificial Intelligence*** also extensively addresses the crucial issue of gender and AI systems. This includes a dedicated chapter outlining policy interventions, specifically **Articles 87-93. Additional recommendations can also be found in UNESCO reports dedicated to gender-related harms online**. For example, the 2024 report on systematic prejudices suggests that stakeholders identify characteristics, contexts and output properties for which AI models must ensure equitable performance (recognizing that models cannot achieve fairness in all contexts at all times). Moreover, improving transparency and encouraging third-party monitoring could enable faster and more comprehensive improvement. Collaborating with independent observers and groups that research gender bias and harms *by allowing safe access to data and trends will allow all parties to better understand such harms, and work towards finding innovative solutions.*[80] Information access for research purposes is also identified in *the Guidelines* as a potential risk mitigation measure in **Articles 116 to 118**.

Finally, it is worth noting that the governance mechanisms outlined in *the Guidelines* aim to promote **cultural diversity, cultural expression, and diverse cultural content in the digital space.** For example, **Article 115** urges technology companies to *provide clear information on the safeguards applied in relation to content moderation and curation that are put in place to protect freedom of expression and access to information and diverse cultural content […].* **Articles 119** on **languages and accessibility** also stipulates that *platforms should have their full terms of service available in the official and primary languages of every country where they operate, ensure that they are able to respond to users in their own language and process their complaints equally, and have the capacity to moderate and curate content in the user's language.*

### 4.4.3 The need to ensure meaningful transparency

Due in part to their novelty and extreme complexity,[81] the inner workings of AI models remain opaque. Notwithstanding significant challenges, the provision of detailed information to regulatory bodies, third parties, users, and the general public regarding potential harms associated with the use or misuse of generative AI is indispensable for its governance including the effective assessment and mitigation of human rights risks.

Transparency standards for the foundation models underpinning generative AI are gradually emerging and include documenting and communicating the engineering decisions that characterize those models. For example, the development of '***model cards',*** which are technical datasheets offering insights into the model's capabilities, limitations, and potential biases, have become established practices amongst industry leaders.

> » *Transparency should be meaningful – the information provided should be as clear and concise as possible, and as detailed and complex as necessary. Transparency is not simply the provision of legal texts or a data dump, but about providing stakeholders with the information they need to make informed decisions (**Article 112**).*

While they may provide a starting point to improve the transparency of generative AI models, these cards remain noncompulsory and non-standardized. As such, they may not offer *meaningful transparency* – sufficient information or sufficiently accessible information for external stakeholders. Furthermore, foundation models exhibit significant cross-lingual differences in safety and capabilities, making it challenging to provide accurate information across languages and cultures. Yet, opacity in generative AI models is not merely a consequence of technological complexity that hinders users' comprehension of the underlying mechanisms. It is further compounded by the practices of market leaders driving innovation in the field. Many private companies choose to conceal critical details citing both short-term competitive pressures and security concerns as justifications for

80 Chowdhury, R. and Dhanya, L. 2023. *Your Opinion Doesn't Matter, Anyway: Exposing Technology-Facilitated Gender-Based Violence in an Era of Generative AI.* Paris, UNESCO. https://www.unesco.org/en/articles/your-opinion-doesnt-matter-anyway.

81 Generative AI models, often based on complex neural networks with billions of connections, have become so intricate that their internal decision-making processes are no longer traceable, even to experts.

withholding information from the public. Although some of these concerns may be legitimate, many experts, researchers as well as employees from AI companies have identified serious gaps over a number of transparency indicators.[82] For example, the Standford Foundation Model Transparency Index scores 10 major foundation developers on 100 transparency metrics.[83] Critically, the Index shows that AI **developers disclose little to no information about the real-world impact of their foundation models**. This includes indicators such as affected market sectors, affected individuals, usage reports, and geographic statistics.

In addition to impact, data access and trustworthiness were systematically, across all AI developers, the least transparent domains. As highlighted by the authors of the study: '*this means that the* ***public has little to no information about who uses foundation models, where foundation models are used, and for what purpose.*** *The lack of transparency regarding these matters inhibits effective governance of foundation models, as it is difficult for governments or civil society organizations to pressure companies to responsibly deploy their models if there is no information about the impact of deployment.*' Drawing parallels with the governance of digital platforms, they also highlight that shortcomings in downstream transparency, in this case by foundation model developers, mirror issues faced by social media companies in the last decade. They therefore recommend adopting certain improvements implemented by digital platforms, such as including detailed usage policy violations and government requests for user data in transparency reports.[84]

Transparency is one of the core principles of *the Guidelines* and several articles are dedicated to defining meaningful transparency in relations to the companies' terms of services (**Article 115a-d**), content moderation and curation policies (**Article 115e-j**) complaints mechanisms (**Article 115k)** and advertising practices (**Article 115l-o**). Many are directly pertinent for the governance of generative AI and could offer a relevant starting point to direct efforts and research.

Another sensitive issue around **transparency involves personal data**. Section 4.2.3 already partially covers this question, yet it is worth highlighting **Article 115j** which stipulates that technology companies should be transparent on *how personal data is collected, used, disclosed, stored, and shared, and what treatment is made of users' personal data, including which personal and sensitive data is used to make algorithmic decisions for the purpose of content moderation and curation.*

Due partly to web scraping techniques, many foundation models are trained on publicly available data containing personal information without the knowledge or consent of the individuals to which it pertains. As a result, generative AI applications have been known to 'leak' information about individuals whose data was used in training. Moreover, these applications could further facilitate malicious use by providing new tools to access sensitive personal data. Experts agree that current privacy-enhancing methods do not effectively apply to foundation models because of their scale. Therefore, many have called for developing better mechanisms for individuals to control and trace their data. This also echoes calls to improves data transparency which remains a key area of opacity across the industry.[85] Reflecting *the Guidelines'* principles*, the 2025 International Safety Report* suggests a series of potential interventions including user-friendly interfaces for managing data permissions and establishing clear processes for individuals to access, view, correct, and delete their data.[86]

82 In June 2024, several former employees of leading generative AI companies published an open letter in which they advocate for robust whistleblower protections for employees and urge AI companies to cultivate a 'culture of open criticism' that encourages, rather than penalizes, those who voice their concerns.
83 Center for Research on Foundation Mondels. 2024. *The Foundation Model Transparency Index.* Standford. https://crfm.stanford.edu/fmti/May-2024/index.html.
84 Bommasani, R. et al. 2024. *The Foundation Model Transparency Index v1.1*. Stanford. https://crfm.stanford.edu/fmti/paper.pdf.
85 Ibid.
86 Privitera, D., op. cit.

### 4.4.4 Building comprehensive accountability mechanisms

Accountability and legal liability for generative AI systems presents a complex governance challenge, particularly when it involves proprietary foundation models. The opaque nature of these systems hinders transparency and complicates the identification of responsible parties. Affected communities may find it hard if not impossible to establish that a particular malfunction within the AI system directly caused their harm, and whether or not this harm was intended or foreseeable by developers or deployers. Experts have also highlighted that individuals harmed by (generative) AI may face substantial initial costs and endure considerably lengthier legal processes compared to cases that do not involve AI.

As per other key issues presented in this document, existing mechanisms to report generative AI harms and redress such harms are embryonic and spotty. Here again *the Guidelines* could offer fruitful considerations to further develop **context specific accountability** in generative AI applications and models. For example, they underscore the importance of establishing **accessible reporting mechanisms** for not only direct users but also for non-users and third parties representing their interests. The document also emphasize the necessity for technology actors to commission **regular external audits and independent assessments** of the impacts of their systems on human rights, cultural diversity, and gender equality. These audits, which must include binding follow-up steps, should cover diverse content types, languages, cultures, and contexts. The results of these reviews should be made public (**Article 103**). Similarly, technology companies must acquire the necessary means to understand local contextual conditions when responding to user complaints, ensuring a culturally sensitive system design **(Article 123).**

In line with the principles of user empowerment (see section 3.2.5), equally crucial in the governance of digital platforms and generative AI, *the Guidelines* recommend that technology providers offer users **options to adjust content curation and moderation systems**. This empowers users to control the content they see, including access to diverse sources and viewpoints on trending topics **(Article 109).** Finally, **redress processes** should also include clear, easily accessible channels for complaints in the user's language and notifications about the results of their appeal **(Article 126).**

The question of appropriate reporting mechanisms is also pertinent to the concluding section of this chapter, concerning AI product release strategies. Current practice often sees the establishment of reporting mechanisms and redress processes only after significant deployment, or not at all. It is essential that effective user communication, coupled with robust reporting and appeals systems, be considered during the design phase, rather than as a reactive measure.

### 4.4.5 Release strategy

In light of ongoing uncertainty, significant risks and severe governance gaps, experts have questioned whether sufficient precautionary assessments were conducted before deploying the most advanced generative foundation models. Suspicion of insufficient precautions led numerous researchers and industry leaders to call leading AI developers to pause the training of advanced AI systems. In the 2023 open letter 'Pause Giant AI Experiment', over 33,000 diverse signatories argued that the level of 'planning and management' of AI labs was insufficient. Their action was also motivated by the fear that corporate providers were engaged in an out-of-control race to develop and deploy systems whose outputs could not be can understood, predicted, or reliably controlled.

> *Their action was also motivated by the fear that corporate providers were engaged in an out-of-control race to develop and deploy systems whose outputs could not be understood, predicted, or reliably controlled.*

To date, there has been no substantial effort to halt the development and release of advanced AI models. Instead of pausing progress, calls for a moratorium have primarily served to emphasize the need for establishing stronger governance mechanisms to distribute the potential benefits and to mitigate the risks associated with developing increasingly powerful AI systems.

This document advocates for the application of the **precautionary principle** to guide the development, deployment, and governance of generative AI. In situations where potential harms associated with generative AI are identified, preventive measures should take precedence over reactive responses. Furthermore, emphasis should be placed on anticipating and mitigating potential risks before they materialize. To that end, it seems imperative that AI developers demonstrate the presence of robust safeguards for human rights and safety mechanisms prior to the release of new products, thereby precluding the deployment and testing of models directly on users. Finally, decisions regarding the development and deployment of generative AI should involve meaningful engagement with a wide range of stakeholders, including local experts, policymakers, and communities.

# 5. Conclusion

The trajectory of generative AI remains uncertain. It is crucial to recognize that its future is not predetermined but rather can be shaped by inclusive and well-informed governance decisions that uphold freedom of expression, diversity, and plurality.

*The Guidelines for the Governance of Digital Platforms* outline a set of responsibilities, duties, and recommendations relevant to the governance of generative AI. These include fostering an enabling environment for freedom of expression and access to information; establishing independent governance systems; and adhering to principles such as human rights due diligence, alignment with international human rights standards, transparency, accountability, and user empowerment.

Such principles are fundamental to a human rights approach to digital platform governance that safeguards freedom of expression and access to information. However, given the specific characteristics and challenges posed by generative AI technologies, *the Guidelines* also highlight additional responsibilities and recommendations that are important to consider when developing appropriate governance frameworks.

## For all stakeholders

- **Examine existing global instruments for their applicability to new fields of technology governance**. While these instruments may not address all actual and potential harms of generative AI and may lack specific guidance for direct implementation in corporate and engineering processes, their adaptable and universal nature allows for application to new circumstances and technological disruptions. They offer critical principles and a baseline upon which new practices and standards can and should be built. In particular, instruments governing digital platforms, such as *the Guidelines,* are highly relevant when considering mechanisms to positively shape the applications of generative AI and its potential impact on freedom of expression and access to information.
- **Support the reinterpretation and integration of human rights standards into national and regional regulatory frameworks related to the digital sphere**. Develop more comprehensive understanding of these rights including access to information, extending beyond their traditional analogue focus and consider new dimensions such as availability, quality, stability, cultural relevance, agency, and usability. These updated dimensions may offer valuable insights for the development of effective governance frameworks for generative AI especially when it comes to ensuring access to information and mitigating risks such as disinformation and privacy violations.
- **Enable the creation of an independent international scientific panel** to collate and catalyze cutting-edge research (which this document heavily relied on) to inform scientists, policymakers, States and other stakeholders seeking scientific perspectives on AI technology or its applications from an impartial, credible source. This panel should prioritize the generation of robust evidence concerning the impact and potential risks of AI on the full spectrum of human rights, as well as the identification and dissemination of best-in-class mitigation mechanisms.

## For States

- In cooperation with all relevant stakeholders, **shape the direction of national AI research and development**, support and invest in efforts to create generative AI models and applications for languages and cultures that are insufficiently digitalised and poorly represented on the web; facilitate access to AI assets (data, computing power and expertise) to researchers in low resources and minority languages. Promoting open solutions and expanding access to both data sets and foundation models may offer additional mechanisms for promoting diversity in freedom of expression.

- **Consider governance mechanisms to ensure fair distribution of computing power,** including managing access to and allocation of computing resources. This could include expanding access to AI computing infrastructure to non-commercial entities, such as academia, civil society, and government bodies. This could also foster greater domestic AI innovation systems that better serve national and local cultures.
- **Clarify how the existing national legal frameworks apply to generative AI**, placing particular attention on updating and interpreting copyright legislations for AI generated content.
- **Maintain full transparency and accountability regarding requirements placed on AI actors.** Carefully consider the benefits of generative AI tools across various domains and avoid broad restrictions or blanket bans, which, similar to internet shutdowns, may cause substantial disruption to information dissemination and significant collateral damage.
- **Require third-party monitoring and testing.** This includes the establishment of dedicated and independent regulatory bodies in the context of statutory arrangements but also mechanisms such as legal safe harbors or government-mediated access regimes, thereby enabling independent researchers and auditors to conduct comprehensive assessments within secure environments without necessitating the public release of source code and model weights.[87] The creation of regulatory sandboxes, allowing for the controlled testing of AI products with selected communities under regulatory oversight, represents a further potential avenue for evaluation.[88]
- **Develop AI-specific public procurement guidelines** (including for example minimum thresholds on explainability and transparency) to ensure that AI models and applications used by public bodies respect human rights. This also includes disclosing all use of AI in the public sector and clarifying how civil servants should approach the use of generative AI in their official functions. **As purchasers** of AI systems, government entities can exert negotiating power in procurement to increase transparency. Notably, some foundation model developers stated that their participation was driven by requests from customers to understand the transparency of their products. **As influential procurers** of technology, they can also play a standard-setting role.
- Mandate the development and updating of **media and information literacy** educational curricula, including **AI literacy** through formal and non-formal education.

## For technology companies and platform providers

- **Urgently improve transparency especially disclosure of information regarding the real-world impact** of foundation models and generative AI applications. This includes detailed information about who uses foundation models, where foundation models are used, and for what purpose**, with particular attention to geographical, cultural and linguistic indicators.** AI developers should offer greater clarity on cross-lingual differences in safety and capabilities and provide greater information on AI systems impact across languages and cultures.
- **Invest in and deploy best-in-class mitigation measures, demonstrating effective implementation of all feasible measures for all relevant communities in each relevant language.** Transparency requires that information about these mitigation measures be specific and detail the affected communities and context.
- Recognizing that AI models cannot achieve fairness in all contexts at all times, **jointly establish with all relevant stakeholders characteristics and output properties for which AI models must ensure equitable performance within a given context.** Disclose the biases for which models and applications are tested, the specific communities targeted in testing, the mitigation measures deployed for each group, and effective mechanisms for affected communities to provide feedback.

87 Raji, I. D. et al. 2022. Outsider Oversight: Designing a Third-Party Audit Ecosystem for AI Governance. *Association for Computing Machinery.* https://dl.acm.org/doi/10.1145/3514094.3534181.
88 Ferrandis, C.M. and Perset, K. 2023. *Regulatory Sandboxes in Artificial Intelligence*. Paris, OECD. https://oecd.ai/en/wonk/sandboxes.

- **Jointly establish with all relevant stakeholders acceptable thresholds and sufficient precautionary assessments** to be conducted before releasing advanced generative AI products. As a minimum, tech companies and platform providers need to ensure their products are safe and conduct thorough human rights risk assessments before introducing them to the public.
- **Develop best in class industry standards across the generative AI supply chain including specific requirements for investors, developers, providers, and deployers.** Notwithstanding the establishment of new industry groups, current efforts are largely characterized by individual initiatives and voluntary commitments without independent oversight. Adopting joint standards and ensuring accountability and compliance towards those standards, including independent periodic mandatory audits are crucial steps in the establishment of an effective governance system for generative AI.
- **Invest in research and development for more robust methods to detect and label AI-generated content.** Consider content provenance solutions which engage actors across the content creation and distribution ecosystem including media companies, news publishers, journalists and content creators.
- **Provide precise definitions of restricted categories of AI-generated content and offer users clear explanations and justifications for these restrictions**, including the potential harms such content may cause. Avoid broad or open-ended restrictive clauses that disproportionately impact user access. This includes developing clear terms of service detailing guardrails and safeguards, and monitoring use for inappropriate content.
- **Prioritize solutions that empower users and support media and information literacy efforts.** This includes the provision of relevant information and readily accessible tools designed to facilitate user comprehension of the various AI products, services, and tools available. With respect to chatbots, this further entails the provision of clear advisories regarding content, context, and provenance, as well as raising user awareness on potential harms.
- **Promptly implement accessible and user-friendly reporting systems**, alongside comprehensive redress procedures, when faced with significant deficiencies in existing mechanisms. Tech companies and platforms providers should also actively promote and fund interdisciplinary research on robust redress frameworks, including best practices for dispute resolution, digital forensics tools for source attribution, and appropriate measures to deal with harmful AI content.

## ABOUT THIS COMPANION DOCUMENT

This companion document is part of UNESCO's efforts to help stakeholders effectively implement *the Guidelines for the Governance of Digital Platforms* and uphold principles of transparency, accountability, due diligence, user empowerment, and alignment with international human rights standards.

The document will serve as a guide for the multistakeholder coalitions being set up by UNESCO through the Social Media 4 Peace project, to formulate and implement advocacy strategies with digital platforms and relevant authorities to curb the spread of harmful content through generative AI, while protecting Freedom of Expression.

It has been prepared by **Marjorie Buchser** Senior Consultant with UNESCO and Chatham House Associate Fellow.

This document has greatly benefited from the significant inputs of experts from various regions of the world including Africa, South Asia, Europe, North and Latin America.

This document has also been enriched by the valuable feedback and comments of numerous UNESCO colleagues who dedicated their time to review it and offer substantial and relevant suggestions. In particular, **Tawfik Jelassi, Sylvie Coudray, Guilherme Canela de Souza for their strategic guidance.** Invaluable inputs and amendments were provided by **Ana Cristina Ruelas, Samrita Menon, Adeline Hullin, Mikel Aguirre Idiaquez, Tim Francis, Prateek Sibal, and Jaco du Toit.**

It is also appropriate to express gratitude to **Ophélie Kukansami Léger, Dissarintr Tovikkai, Lucas Novaes Ferreira, Daria Kovaleva and Daniel Joshua Brini** for their assistance throughout the drafting of this document.

**Document code**: CI/FMD/FEJ/2025/70

Funded by
the European Union

The production of this document was produced with the financial support of the European Union as part of the Social Media 4 Peace project. Its contents are the sole responsibility of its author(s) and do not necessarily reflect the views of the European Union.