WORLD ECONOMIC FORUM

AI Governance Alliance
Briefing Paper Series

JANUARY 2024

Foreword

Paul Daugherty
Chief Technology and Innovation Officer (CTIO), Accenture

Jeremy Jurgens
Managing Director, World Economic Forum

John Granger
Senior Vice-President, IBM Consulting

Cathy Li
Head, AI, Data and Metaverse; Member of the Executive Committee, World Economic Forum

Our world is experiencing a phase of multi-faceted transformation in which technological innovation plays a leading role. Since its inception in the latter half of the 20th century, artificial intelligence (AI) has journeyed through significant milestones, culminating in the recent breakthrough of generative AI. Generative AI possesses a remarkable range of abilities to create, analyse and innovate, signalling a paradigm shift that is reshaping industries from healthcare to entertainment, and beyond.

As new capabilities of AI advance and drive further innovation, it is also revolutionizing economies and societies around the world at an exponential pace. With the economic promise and opportunity that AI brings, comes great social responsibility. Leaders across countries and sectors must collaborate to ensure it is ethically and responsibly developed, deployed and adopted.

The World Economic Forum's AI Governance Alliance (AIGA) stands as a pioneering collaborative effort, uniting industry leaders, governments, academic institutions and civil society organizations. The alliance represents a shared commitment to responsible AI development and innovation while upholding ethical considerations at every stage of the AI value chain, from development to application and governance. The alliance, led by the World Economic Forum in collaboration with IBM Consulting and Accenture as knowledge partners, is made up of three core workstreams – Safe Systems and Technologies, Responsible Applications and Transformation, and Resilient Governance and Regulation. These pillars underscore a comprehensive end-to-end approach to address key AI governance challenges and opportunities.

The alliance is a global effort that unites diverse perspectives and stakeholders, which allows for thoughtful debates, ideation and implementation strategies for meaningful long-term solutions. The alliance also advances key perspectives on access and inclusion, driving efforts to enhance access to critical resources such as learning, skills, data, models and compute. This work includes considering how such resources can be equitably distributed, especially to underserved regions and communities. Most critically, it is vital that stakeholders who are typically not engaged in AI governance dialogues are given a seat at the table, ensuring that all voices are included. In doing so, the AI Governance Alliance provides a forum for all.

As we navigate the dynamic and ever-evolving landscape of AI governance, the insights from the AI Governance Alliance are aimed at providing valuable guidance for the responsible development, adoption and overall governance of generative AI. We encourage decision-makers, industry leaders, policy-makers and thinkers from around the world to actively participate in our collective efforts to shape an AI-driven future that upholds shared human values and promotes inclusive societal progress for everyone.

Introduction to the briefing paper series

The AI Governance Alliance was launched in June 2023 with the objective of providing guidance on the responsible design, development and deployment of artificial intelligence systems. Since its inception, more than 250 members have joined the alliance from over 200 organizations across six continents. The alliance is comprised of a steering committee along with three working groups.

The Steering Committee comprises leaders from the public and private sectors along with academia and provides guidance on the overall direction of the alliance and its working groups.

The Safe Systems and Technologies working group, led in collaboration with IBM Consulting, is focused on establishing consensus on the necessary safeguards to be implemented during the development phase, examining technical dimensions of foundation models, including guardrails and responsible release of models and applications. Accountability is defined at each stage of the AI life cycle to ensure oversight and thoughtful expansion.

The Responsible Applications and Transformation working group, led in collaboration with IBM Consulting, is focused on evaluating business transformation for responsible generative AI adoption across industries and sectors. This includes assessing generative AI use cases enabling new or incremental value creation, and understanding their impact on value chains and business models while evaluating considerations for adoption and their downstream effects.

The Resilient Governance and Regulation working group, led in collaboration with Accenture, is focused on the analysis of the AI governance landscape, mechanisms to facilitate international cooperation to promote regulatory interoperability, as well as the promotion of equity, inclusion and global access to AI.

This briefing paper series is the first output from each of the three working groups and establishes the foundational focus areas of the AI Governance Alliance.

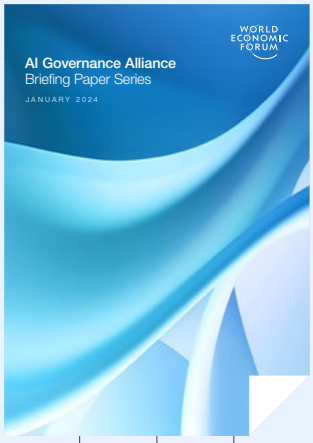
In a time of rapid change, the AI Governance Alliance seeks to build a multistakeholder community of trusted voices from across the public, private, civil society and academic spheres, united, to tackle some of the most challenging and potentially most rewarding issues in contemporary AI governance.

Reading guide

This paper series is composed of three briefing papers that have been grouped into thematic categories according to the three working groups of the alliance.

Each briefing paper of the report can also be read as a stand-alone piece. For example, developers, adopters and policy-makers who are more interested in the technical dimensions can easily jump to the Safe Systems and Technologies briefing paper to obtain a contemporary understanding of the AI landscape. For decision-makers engaged in corporate strategy and business implications of generative AI, the Responsible Applications and Transformation briefing paper offers specific context. For business leaders and policy-makers occupied with the laws, policies, principles and practices that govern the ethical development, deployment, use and regulation of AI technologies, the Resilient Governance and Regulation briefing paper offers guidance.

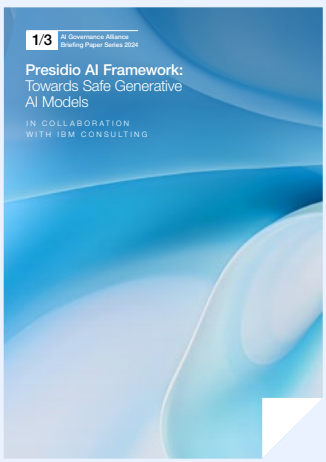
While each briefing paper has a unique focus area, many important lessons are learned at the intersection of these varying multistakeholder communities, along with the consensus and knowledge that emanate from each working group. Therefore, many of the takeaways from this briefing paper series should be viewed at the intersection of each working group, where findings become additive and are enhanced in context and interrelation with one another.

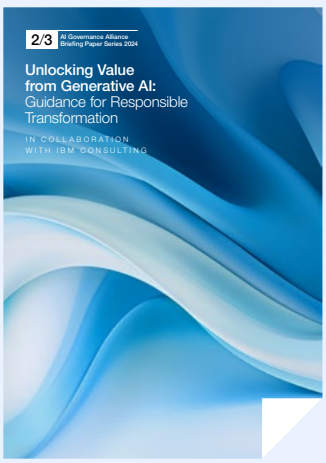


Theme 1
Safe Systems and Technologies

Theme 2
Responsible Applications and Transformation

Theme 3
Resilient Governance and Regulation





AI Governance Alliance Steering Committee

Nick Clegg
President, Global Affairs, Meta

Gary Cohn
Vice-Chairman, IBM

Sadie Creese
Professor of Cybersecurity, University of Oxford

Orit Gadiesh
Chairman, Bain & Company

Paula Ingabire
Minister of Information Communication Technology of Rwanda

Daphne Koller
Founder and Chief Executive Officer, Insitro

Xue Lan
Professor; Dean, Schwarzman College, Tsinghua University

Anna Makanju
Vice-President, Global Affairs, OpenAI

Durga Malladi
Senior Vice-President, Qualcomm

Andrew Ng
Founder, DeepLearning.AI

Sabastian Niles
President and Chief Legal Officer, Salesforce

Omar Sultan Al Olama
Minister of State for Artificial Intelligence, United Arab Emirates

Lynne Parker
Associate Vice-Chancellor and Director, AI Tennessee Initiative, University of Tennessee

Brad Smith
Vice-Chair and President, Microsoft

Mustafa Suleyman
Co-Founder and Chief Executive Officer, Inflection AI

Josephine Teo
Minister for Communications and Information Ministry of Communications and Information (MCI) of Singapore

Kent Walker
President, Global Affairs, Google

Glossary

Terminology in AI is a fast-moving topic, and the same term can have multiple meanings. The glossary below should be viewed as a snapshot of contemporary definitions.

Artificial intelligence system: a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.[1]

Causal AI: AI models that identify and analyse causal relationships in data, enabling predictions and decisions based on these relationships. Causal inference models provide responsible AI benefits, including explainability and bias reduction through formalizations of fairness, as well as contextualisation for model reasoning and outputs. The intersection and exploration of causal and generative AI models is a new conversation.

Fine-tuning: The process of adapting a pre-trained model to perform a specific task by conducting additional training while updating the model's existing parameters.

Foundation model: A foundation model is an AI model that can be adapted to a wide range of downstream tasks. Foundation models are typically large-scale (e.g. billions of parameters) generative models trained on a vast array of data, encompassing both labelled and unlabelled datasets.

Frontier model: This term generally refers to the most advanced or cutting-edge models in AI technology. Frontier models represent the latest developments and are often characterized by increased complexity, enhanced capabilities and improved performance over previous models.

Generative AI: AI models specifically intended to produce new digital material as an output (e.g. text, images, audio, video and software code), including when such AI models are used in applications and their user interfaces. These are typically constructed as machine learning systems that have been trained on massive amounts of data.[2]

Hallucination: Hallucinations occur when models produce factually inaccurate or untruthful information. Often, hallucinatory output is presented in a plausible or convincing manner, making detection by end users difficult.

Jurisdictional interoperability: The ability to operate within and across different jurisdictions governed by differing policy and regulatory requirements.[3]

Mis/disinformation: Misinformation involves the dissemination of incorrect facts, where individuals may unknowingly share or believe false information without the intent to mislead. Disinformation involves the deliberate and intentional spread of false information with the aim of misleading others.[4]

Model drift monitoring: The act of regularly comparing model metrics to maintain performance despite changing data, adversarial inputs, noise and external factors.

Model hyperparameters: Adjustable parameters of a model that must be tuned to obtain optimal performance (as opposed to fixed parameters of a model, defined based on its training set).

Multi-modal AI: AI technology capable of processing and interpreting multiple types of data (like text, images, audio, video), potentially simultaneously. It integrates techniques from various domains (natural language processing, computer vision, audio processing) for more comprehensive analysis and insights.

Prompt engineering: The process of designing natural language prompts for a language model to perform a specific task.

Retrieval augmented generation: A technique in which a large language model is augmented with knowledge from external sources to generate text. In the retrieval step, relevant documents from an external source are identified from the user's query. In the generation step, portions of those documents are included in the model prompt to generate a response grounded in the retrieved documents.

Parameter-efficient fine-tuning: An efficient, low-cost way of adapting a pre-trained model to new tasks without retraining the model or updating its weights. It involves learning a small number of new parameters that are appended to a model's prompt while freezing the model's existing parameters (also known as prompt-tuning).

AI red teaming: A method of simulating attacks by a group of people authorized and organized to identify potential weaknesses, vulnerabilities and areas for improvement. It should be integral from model design to development to deployment and application. The red team's objective is to improve security and robustness by demonstrating the impacts of successful attacks and by demonstrating what works for the defenders in an operational environment.

Reinforcement learning from human feedback (RLHF): An approach for model improvement where human evaluators rank model-generated outputs for safety, relevance and coherence, and the model is updated based on this feedback to broadly improve performance.

Release access – A gradient covering different levels of access granted.[5]

- **Fully closed:** The foundation model and its components (like weights, data and documentation) are not released outside the creator group or sub-section of the organization. The same organization usually does model creation and downstream model adaptation. External users may interact with the model through an application.
- **Hosted:** Creators provide access to the foundation model by hosting it on their infrastructure, allowing internal and external interaction via a user interface, and releasing specific model details.
- **Application programming interface (API):** Creators provide access to the foundation model by hosting it on their infrastructure and allowing adapter interaction via an API to perform prescribed tasks and release specific model details.
- **Downloadable:** Creators provide a way to download the foundation model for running on the adapters' infrastructure while withholding some of its components, like training data.
- **Fully open:** Creators release all model components, including all parameters, weights, model architecture, training code, data and documentation.

Responsible adoption: The adoption of individual use cases and opportunities within the responsible AI framework of an organization. It requires thorough evaluation to ensure that value can be realized and change management is successfully aligned with defined goals in a responsible framework.

Responsible AI: AI that is developed and deployed in ways that maximize benefits and minimize the risks it poses to people, society and the environment. It is often described by various principles and organizations, including but not limited to robustness, transparency, explainability, fairness and equity.[6]

Responsible transformation: The organizational effort and orientation to harness the opportunities and benefits of generative AI while mitigating the risks to individuals, organizations and society. Responsible transformation is strategic coordination and change across an organization's governance, operations, talent and communications.

Traceability: Determining the original source and facts of the generated output.

Transparency: The disclosure of details (decisions, choices and processes) in the documentation about the sources, data and model to enable informed decisions regarding model selection and understanding.

Usage restriction: The process of restricting the usage of the model beyond the intended use cases/purpose to avoid unintended consequences of the model.

Watermarking: The act of embedding information into outputs created by AI (e.g. images, videos, audio, text) for the purposes of verifying the authenticity of the output, identity and/or characteristics of its provenance, modifications and/or conveyance.[7]

Endnotes

1. "OECD AI Principles overview", *Organisation for Economic Co-operation and Development (OECD) AI Policy Observatory*, 2023, https://oecd.ai/en/ai-principles.
2. OECD, *G7 Hiroshima Process on Generative Artificial Intelligence (AI) Towards a G7 Common Understanding on Generative AI*, 2023, https://www.oecd.org/publications/g7-hiroshima-process-on-generative-artificial-intelligence-ai-bf3c0c60-en.htm.
3. World Economic Forum, *Interoperability In the Metaverse*, 2023, https://www.weforum.org/publications/interoperability-in-the-metaverse/.
4. World Economic Forum, *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, 2023, https://www.weforum.org/publications/toolkit-for-digital-safety-design-interventions-and-innovations-typology-of-online-harms/.
5. Solaiman, Irene, "The Gradient of Generative AI Release: Methods and Considerations", *Hugging Face*, 2023, https://arxiv.org/abs/2302.04844.
6. World Economic Forum, *The Presidio Recommendations on Responsible Generative AI*, 2023, https://www.weforum.org/publications/the-presidio-recommendations-on-responsible-generative-ai/.
7. The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 2023: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

1/3 AI Governance Alliance Briefing Paper Series 2024

Presidio AI Framework: Towards Safe Generative AI Models

IN COLLABORATION WITH IBM CONSULTING

Cover image: MidJourney

Contents

Executive summary

The Presidio AI Framework addresses generative AI risks by promoting safety, ethics, and innovation with early guardrails.

The rise of generative AI presents significant opportunities for positive societal transformations. At the same time, generative AI models add new dimensions to AI risk management, encompassing various risks such as hallucinations, misuse, lack of traceability and harmful output. Therefore, it is essential to balance safety, ethics and innovation.

This briefing paper identifies a list of challenges to achieving this balance in practice, such as lack of a cohesive view of the generative AI model life cycle and ambiguity in terms of the deployment and perceived effectiveness of varying safety guardrails throughout the life cycle. Amid these challenges, there are significant opportunities, including greater standardization through shared terminology and best practices, facilitating a common understanding of the effectiveness of various risk mitigation strategies.

This briefing paper presents the **Presidio AI Framework**, which provides a structured approach to the safe development, deployment and use of generative AI. In doing so, the framework highlights gaps and opportunities in addressing safety concerns, viewed from the perspective of four primary actors: AI model creators, AI model adapters, AI model users, and AI application users. Shared responsibility, early risk identification and proactive risk management through the implementation of appropriate guardrails are emphasized throughout.

The Presidio AI Framework consists of three core components:

1. **Expanded AI life cycle**: This element of the framework establishes a comprehensive end-to-end view of the generative AI life cycle, signifying varying actors and levels of responsibility at each stage.

2. **Expanded risk guardrails**: The framework details robust guardrails to be considered at different steps of the generative AI life cycle, emphasizing prevention rather than mitigation.

3. **Shift-left methodology:** This methodology proposes the implementation of guardrails at the earliest stage possible in the generative AI life cycle. While shift-left is a well-established concept in software engineering, its application in the context of generative AI presents a unique opportunity to promote more widespread adoption.

In conclusion, the paper emphasizes the need for greater multistakeholder collaboration between industry stakeholders, policy-makers and organizations. The Presidio AI Framework promotes shared responsibility, early risk identification and proactive risk management in generative AI development, using guardrails to ensure ethical and responsible deployment. The paper lays the foundation for ongoing safety-related work of the AI Governance Alliance and the Safe Systems and Technologies working group. Future work will expand on the core concepts and components introduced in this paper, including the provision of a more exhaustive list of known and novel guardrails, along with a checklist to operationalize the framework across the generative AI life cycle.

Introduction

The current AI landscape includes both challenges and opportunities for progress towards safe generative AI models.

This briefing paper outlines the Presidio AI Framework, providing a structured approach to addressing both technical and procedural considerations for safe generative artificial intelligence (AI) models. The framework centres on foundation models and incorporates risk-mitigation strategies throughout the entire life cycle, encompassing creation, adaptation and eventual retirement. Informed by thorough research into the current AI landscape and input from a multistakeholder community and practitioners, the framework underscores the importance of established safety guidelines and recommendations viewed through a technical lens. Notable challenges in the existing landscape impacting the development and deployment of safe generative AI include:

- **Fragmentation:** A holistic perspective, which covers the entire life cycle of generative AI models from their initial design to deployment and the continuous stages of adaptation and use, is currently missing. This can lead to fragmented perceptions of the model's creation and the risks associated with its deployment.

- **Vague definitions:** Ambiguity and lack of common understanding of the meaning of safety, risks[1] (e.g. traceability), and general safety measures (e.g. red teaming) at the frontier of model development.

- **Guardrail ambiguity:** While there is agreement on the importance of risk-mitigation strategies – known as guardrails – clarity is lacking regarding accountability, effectiveness, actionability, applicability, limitations and at what stages of the AI design, development and release life cycle varying guardrails should be implemented.

- **Model access:** An open approach presents significant opportunities for innovation, greater adoption and increased stakeholder population diversity. However, the availability of all the model components (e.g. weights, technical documentation and code) could also amplify risks and reduce guardrails' effectiveness. There is a need for careful analysis of risks and common consensus among the use of guardrails considering the gradient of release;[2] that is, varying levels at which AI models are accessible once released, from fully closed to fully open-sourced.

Simultaneously, there are some identified opportunities for progress towards safety, such as:

- **Standardization:** By linking the technical aspects at each phase of design, development and release with their corresponding risks and mitigations, there is the opportunity for bringing attention to shared terminology and best practices. This may contribute towards greater adoption of necessary safety measures and promote community harmonization across different standards and guidelines.

- **Stakeholder trust and empowerment:** Pursuing clarity and agreement on the expected risk mitigation strategies, where these are most effectively located in the model life cycle and who is accountable for implementation paves the way for stakeholders to implement these proactively. This improves safety, prevents adverse outcomes for individuals and society, and builds trust among all stakeholders.

While this briefing paper details the generative AI model life cycle along with some guardrails, it is by no means exhaustive. Some topics outside this paper's scope include a discussion of current or future government regulations of AI risks and mitigations (this is covered in the Resilient Governance working group briefing paper) or consideration of downstream implementation and use of specific AI applications.

1 Introducing the Presidio AI Framework

A structured approach that emphasizes shared responsibility and proactive risk mitigation by implementing appropriate guardrails early in the generative AI life cycle.

Those releasing, adapting or using foundation models often face challenges in influencing the original model design or setting up the necessary infrastructure for building foundation models. The combined need for regulatory compliance, the significant investments companies are making in AI, and the potential impacts the technology can have on society mean coordination among multiple roles and stakeholders becomes indispensable.

FIGURE 1 **The three elements of the Presidio AI Framework**

The Presidio AI Framework (illustrated in Figure 1) offers a streamlined approach to generative AI development, deployment and use from the perspective of four primary actors: AI model creators, AI model adapters, AI model users and AI application users. This human-centric framework harmonizes the activities of these roles to enable more efficient information transfer between upstream development and downstream applications of foundation models.

AI model creators are responsible for the end-to-end design, development and release of generative AI models. AI model adapters tailor generative AI models to specific generative tasks before integration into AI applications and can provide feedback to the AI model creator. AI model users interact with a generative AI model through an interface provided by the creator. AI application users interact indirectly with the adapted model through an application or application programming interface (API). These actors include secondary groups, for instance, AI model validators and AI model auditors, whose goal is to test and validate against defined metrics, perform safety evaluations or certify the conformity of the AI models pre-release. Validators are internal to AI creator or adapter organizations, while auditors are external entities pursuing model certification.

2

Expanded AI life cycle

The expanded AI life cycle encompasses risks and guardrails with varying safety benefits and challenges throughout each phase.

The expanded AI life cycle synthesizes elements from data management, foundation model design and development, release access, use of generative capabilities and adaptation to a use case. The expanded AI life cycle is introduced in Figure 2.

FIGURE 2 **Presidio AI Framework's expanded AI life cycle**

The **data management phase** describes the data foundations for responsible AI development, including the data access gradient and the catalogue of data source types. The latter aids the AI model creator in navigating various legal implications and challenges, where multiple data source types are typically considered in model creation.

In the **foundation model building phase**, the model moves through various stages from design to internal audit and approval. In contrast, each stage is accompanied by a set of distinct guardrails, detailed in the following section.

The **foundation model release phase** provides responsible model dissemination and risk mitigation, benefiting downstream users and adapters. Foundation models are classified based on how they are released, depending on the level of access granted to downstream actors. This gradient of access spans from fully closed to fully open access; each access type has its own set of norms, standards and release guardrails and has specific benefits and challenges, highlighted in Table 1.

In all phases, unexpected model behaviour could harm users and bring reputational risks or legal consequences to the user and the model creator or adapter. However, the chances of misuse – such as plagiarism, intentional non-disclosure, violation of intellectual property (IP) rights, deepfakes, creation of biologically harmful compounds, generation of toxic content, and misinformation generation – may increase if vigilant oversight processes are not adequately implemented going from fully closed to fully open model access.

TABLE 1 **Safety benefits and challenges of release types**

Release type	Safety benefits	Safety challenges
Fully closed	Creators control the model use and can provide safeguards for data privacy and the IP contained in the model. There is more clarity around responsibility and ownership.	Other actors have limited visibility into the model design and development process. Auditability and contributors' diversity are limited. Application users have minimal influence on model outputs.
Hosted	Creators can provide safeguards for model outputs, such as blocking model response for sensitive queries. They can streamline user support. Use can be tracked and used to improve model responses.	Similar challenges as "fully closed". Other actors have little insight into the model, limiting their ability to understand its decisions.
API	Creators retain control over the model while empowering users to adapt the model for specific use cases. They can provide user support. This level of access increases the "researchability" of the model. Increased access allows users to help identify risks and vulnerabilities.	Even though transparency is limited, model details can be inferred by third-party tools or attacks (in case of bad actors).
Downloadable	Along with creators, adapters and users are also empowered through the release of model components. This means more transparency, flexibility for model use and modification of the model.	Lowered barriers for misuse and potential bypassing of guardrails. Model creators have difficulties in tracking and monitoring model use. Users typically have less support when experiencing unexpected undesirable model outputs/outcomes.
Fully open	These models provide the highest levels of auditability and transparency. This level of access increases global participation and contribution to innovation – also in terms of safety and guardrails. Adapters and users are empowered to adapt models that better align with their specific task and improve existing model functionality and safety via fine tuning.	These models present a higher chance of possible misuse. Access to model weights means higher risk of model replication for unintended purposes by bad actors. Ambiguity around accountability and ownership.

The **model adaptation phase** describes several stages, techniques and guardrails for adapting a pre-trained foundation model to perform specific generative tasks. This phase precedes the **model integration phase**, involving the model's integration with an application, including developing APIs to serve downstream AI application users.

In the **model use phase**, users engage with hosted access models using natural language prompts through an interface provided by the model creator or test it for vulnerabilities. This phase highlights the importance of having necessary guardrails during the foundation model building and release phases as users directly interact with the model. In contrast, adapters can add additional guardrails based on the use case.

3 Guardrails across the expanded AI life cycle

Implementation of known and novel guardrails is necessary for safe systems to ensure technical quality, consistency and control.

Guardrails for safe AI systems refer to guidelines, principles and practices that are put in place to ensure the responsible development, deployment and use of generative AI systems and technologies. They are intended to mitigate risks, prevent harm and ensure AI systems operate according to specific standards and ethical and societal values. Guardrails are implemented from the model-building phase and onward throughout the expanded AI life cycle and may be technical or procedural. Technical guardrails involve tools or automated systems and controls, while procedural guardrails rely on human adherence to established processes and guidelines. A combination of both types is often needed to ensure safe systems. Technical guardrails ensure technical quality and consistency, while procedural guardrails provide process consistency and control.

The section below provides a snapshot of selected guardrails applicable at varying phases of the AI life cycle. Due to brevity, only two of the most widely used guardrails are highlighted, along with their phase placement.

TABLE 2 **Highlighted guardrails and their phase placement**

Highlighted guardrails	Phase placement
Red teaming and reinforcement learning from human feedback (RLHF)[3]	Building
Transparent documentation and use restriction	Release
Model drift monitoring and watermarking	Adaptation

3.1 Model building phase

Performing red teaming early, especially during fine-tuning and validation of the building phase, is crucial for preventing adverse outcomes and ensuring model safety. Addressing vulnerabilities and ethical concerns earlier in the life cycle demonstrates a commitment to security and ethics while building trust among stakeholders. For foundation models, tests should cover prompt injection, leaking, jailbreaking, hallucination, IP and personal information (PI) generation, as well as identifying toxic content. While red teaming is effective for known vulnerabilities, it may have limitations in identifying unknown risks, especially before mass release.

Incorporating reinforcement learning from human feedback (RLHF) early on provides a strategic advantage by enabling efficient learning, faster iterations and a strong foundation for subsequent phases, ultimately leading to improved model performance and alignment with human objectives. RLHF may be used here to train a reward model, which is then used to fine-tune the primary model, eliciting more desirable responses. This process ensures the reliability and alignment of the model outputs and improves performance, including an iterative feedback loop between human raters, a trained reward model and the foundation model. Although effective for ongoing improvement, there is a risk of introducing new biases with this method and data privacy and security considerations around the use of generated data.

Novel approaches to implement these guardrails include "red teaming language models with language models" and reinforcement learning from AI feedback (RLAIF).[4] Both techniques employ language models to generate test cases or provide safety-related feedback on the model. The automation significantly reduces the time needed to implement these guardrails. These may also be applied in later phases, but the advantage of using them earlier allows for adjustments to the model hyperparameters to enhance performance. However, they may come with new vulnerabilities that are not yet fully identified.

3.2 Model release phase

Guardrails implemented in the release phase include a combination of approaches designed to empower downstream actors (such as transparent documentation) and protect them (such as use restrictions).

Transparent documentation is a collection of details (decisions, choices and processes) about the AI model, including the data. It mitigates the risk of lack of transparency,[5] and therefore empowers downstream adapters and users to understand the model's limitations, evaluate its impact and make decisions on model use. This guardrail increases the auditability of the model and helps advance policy initiatives. Some best practices include understanding target consumers, their requirements, and expectations, developing persona-based (e.g. business owner, validator and auditors) templates with pre-defined fields and assigning responsibility for gathering information at every phase of the life cycle. Datasheets, data cards, model cards, factsheets and Stanford's foundation model transparency index indicators are a few examples of building templates. Automating fact collection, building documentation and auditing transparency could improve overall efficiency and effectiveness. Limitations include identifying the most useful facts and ambiguity in balancing the disclosure of proprietary and required information.

Use restriction limits the model use beyond intended purposes. It mitigates the risk of model misuse and other unintended harms like generating harmful content and model adaptation for problematic use cases. Some best practices involve using restrictive licences like responsible AI licences (RAIL), setting up model use and user tracking, and providing clear guidelines on allowed use while implementing feedback/incident reporting mechanisms. Additionally, integrating moderation tools to filter or flag undesirable content, disallowing harmful or sensitive prompts and blocking the model from responding to misaligned prompts must be considered. Limitations include having standards for model licences and guidelines and high-quality tools to help restrict the model response.

3.3 Model adaptation phase

A critical goal of the adaptation phase is to ensure that the adapted model remains effective and aligned with the selected use case. Model drift monitoring involves regularly comparing post-deployment metrics to maintain performance in the face of evolving data, adversarial inputs, noise and external factors. The goal is to mitigate the risk of model drift, where the model's output deviates from expectations over time. Best practices include systematically using data, algorithms, and tools for tracking data drift, and defining response protocols and adaptation techniques to sustain model performance and customer trust.

The decision to watermark model outputs depends on the use case, model nature and watermarking goals. Watermarking adds hidden patterns for algorithmic detection, mitigating mass production of misleading content. It aids in identifying AI-generated content for policy enforcement, attribution, legal recourse and deterrence. However, workarounds exist, such as removing watermarks or paraphrasing content. Watermarking can be applied earlier (during model creation for ownership) and adaptation for control over visibility.

4 Shifting left for optimized risk mitigation

The “shift-left” approach involves implementing safety guardrails earlier in the life cycle to mitigate risks and increase efficiency.

The term “shift-left”[6] describes implementing quality assurance and testing measures earlier in a product cycle. The core objective is proactively identifying and managing potential risks, increasing efficiency and cost-effectiveness. This well-established concept applies to various technologies and processes, including software engineering.

In the Presidio AI Framework, the concept of shift-left is extended and applied to generative AI models. It gains a new dimension of importance due to:

- Increased interest in foundation models where model creators are not always the model adapters.
- Increased accessibility of powerful models by users of varying skills and technical backgrounds, raising the demand for model transparency.
- Considerable risk for users using factually incorrect output without validation, model misuse (e.g. in disinformation campaigns) and adversarial attacks on the model (e.g. jailbreaking).

These considerations require understanding and coordination of the activities of different actors (creators, adapters and users) across the AI value chain to avoid significant effort in resolving issues during model adoption and use. For example, data subject rights in some countries allow people to request that their personal information be deleted from the model. The removal can be costly for model creators as they may need to retrain the model. It can also be challenging for adaptors to apply effective guardrails to prevent sensitive information from surfacing in the output.

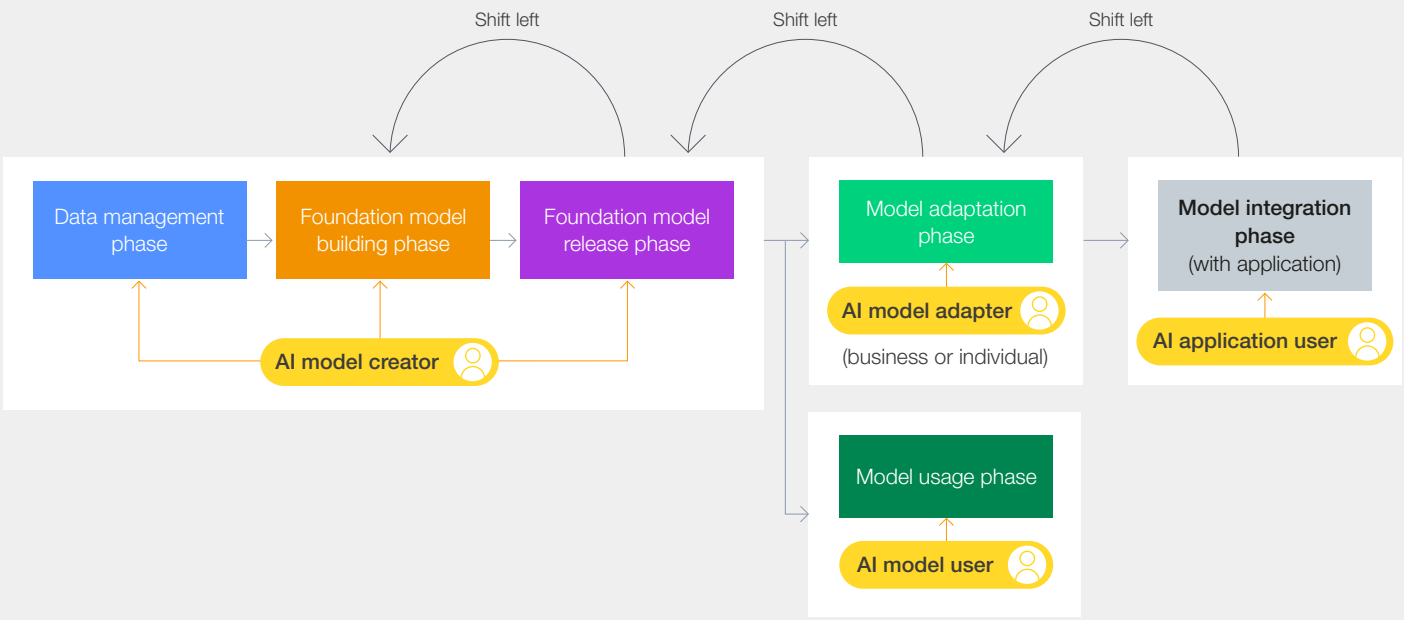
For generative AI, the shift-left methodology proposes guardrails earlier in the life cycle, considering their effectiveness in mitigating risk at a particular phase, along with essential foundation model safety features, the need for balancing safety with model creativity and implementation cost. Based on the model’s purpose, there could be a trade-off between guardrail placement and safety dimensions like privacy, fairness, accuracy and transparency.

Figure 3 illustrates three shift-left instances crucial for building safe generative AI models.

- **Release to build shift** occurs when an AI model creator proactively incorporates guardrails throughout the foundation model-building phase and collects necessary data and model facts and transparency surrounding these.
- **Adaptation/use to release shift** occurs during the foundation model release phase. The AI model creator incorporates additional guardrails, establishes norms and standards for use, and creates comprehensive documentation to help downstream actors understand and make informed decisions regarding model use.
- **Application to adaptation shift** occurs when the AI model adapter proactively incorporates guardrails considering the use case and considering the documentation from AI model creators about the foundation model. These would be documented for the downstream application user.

Some organizations have already integrated the shift-left approach into their responsible AI development process. However, it is vital to extend and emphasize the importance of this practice across all expanded phases of the generative AI life cycle and ensure its adoption by all organizations. Those that shift left to implement appropriate safety guardrails where most effective can minimize legal consequences and reputational risk, increase trusted adoption and positively impact society and users.

FIGURE 3 | **Presidio AI Framework with shift-left methodology for generative AI models**



Conclusion

The Presidio AI Framework promotes shared responsibility, early risk identification and proactive risk management in generative AI development, using guardrails to ensure ethical and responsible deployment. The AI Governance Alliance and the Safe Systems and Technologies working group encourage greater information exchange between industry stakeholders, policy-makers and organizations. This collaborative effort aims to increase trust in AI systems, ultimately benefiting society.

In addition to known guardrails, the group will continue to identify novel mechanisms for AI safety, including emerging technical guardrails such as red teaming language models,[7] liquid neural networks (LNN),[8] BarrierNets,[9] causal foundation models[10] and neurosymbolic learning,[11] among others. Additionally, the group will investigate the various guardrail options and introduce a checklist to operationalize the framework to assess AI model risks and guardrails across the generative AI life cycle.

Contributors

This paper is a combined effort based on numerous interviews, discussions, workshops and research. The opinions expressed herein do not necessarily reflect the views of the individuals or organizations involved in the project or listed below. Sincere thanks are extended to those who contributed their insights via interviews and workshops, as well as those not captured below.

World Economic Forum

Benjamin Larsen
Lead, Artificial Intelligence and Machine Learning

Cathy Li
Head, AI, Data and Metaverse; Deputy Head, Centre for the Fourth Industrial Revolution; Member of the Executive Committee

Supheakmungkol Sarin
Head, Data and Artificial Intelligence Ecosystems

AI Governance Alliance Project Fellows

Ravi Kiran Singh Chevvan
AI Strategy & Complex Program Executive, IBM

Jerry Cuomo
Executive Fellow and Vice-President, Technology, IBM

Steven Eliuk
Executive Fellow and Vice-President, AI & Governance, IBM

Jennifer Kirkwood
Executive Fellow, Partner, IBM

Eniko Rozsa
Distinguished Engineer, IBM

Saishruthi Swaminathan
Tech Ethics Program Adviser, IBM

Joseph Washington
Senior Technical Staff Member, IBM

Acknowledgements

Sincere appreciation is extended to the following working group members, who spent numerous hours providing critical input and feedback to the drafts. Their diverse insights are fundamental to the success of this work.

Uthman Ali
Senior Product Analyst, AI Ethics SME, BP

Animashree (Anima) Anandkumar
Bren Professor of Computing and Mathematical Sciences, California Institute of Technology (Caltech)

Amir Banifatemi
Co-Founder and Director, AI Commons

Michael Benton
Director, Responsible AI Practice, Microsoft

Stella Biderman
Executive Director, EleutherAI

Shane Cahill
Director, Privacy and AI Legislation and Policy Development, Meta Platforms

Suha Can
Chief Information Security Officer, Grammarly

Jennifer Chayes
Dean of the College of Computing, Data Science, and Society, University of California, Berkeley

Kevin Chung
Chief Operating Officer, Writer

Jeff Clune
Associate Professor, Department of Computer Science, Faculty of Science, Vector Institute

Cathy R Cobey
Global Responsible Co-Lead, EY

Umeshwar Dayal
Corporate Chief Scientist, Hitachi

Mona Diab
Director of Language Technologies Institute, Carnegie Mellon University

Mennatallah El-Assady
Professor, ETH Zurich

Gilles Fayad
Adviser, Institute of Electrical and Electronics Engineers (IEEE)

Jocelyn Goldfein
Managing Director, Zetta Venture Partners

Tom Gruber
Founder, Humanistic AI

Lan Guan
Global Data and AI Lead, Senior Managing Director, Accenture

Gillian Hadfield
Professor of Law and Professor of Strategic Management, University of Toronto

Peter Hallinan
Leader, Responsible AI, Amazon Web Services

Or Hiltch
Chief Data and AI Architect, JLL

Babak Hodjat
Chief Technology Officer AI, Cognizant Technology Solutions US

Sara Hooker
Head, Research, Cohere

David Kanter
Founder and Executive Director, MLCommons

Vijay Karunamurthy
Head of Engineering and Vice-President, Engineering, Scale AI

Sean Kask
Chief AI Strategy Officer, SAP

Robert Katz
Vice-President, Responsible AI & Tech, Salesforce

Michael Kearns
Founding Director, Warren Center for Network and Data Sciences, University of Pennsylvania

Steve Kelly
Chief Trust Officer, Institute for Security and Technology

Jin Ku
Chief Technology Officer, Sendbird

Sophie Lebrecht
Chief, Operations and Strategy, Allen Institute for Artificial Intelligence

Aiden Lee
Co-Founder and Chief Technology Officer, Twelve Labs

Stefan Leichenauer
Vice-President, Engineering, SandboxAQ

Tze Yun Leong
Professor of Computer Science; Director, NUS Artificial Intelligence Laboratory

Scott Likens
Global AI and Innovation Technology Lead, PwC

Shane Luke
Vice-President, Product and Engineering, Workday

Richard Mallah
Principal AI Safety Strategist, Future of Life Institute

Pilar Manchón
Senior Director, Engineering, Google

Risto Miikkulainen
Professor of Computer Science, University of Texas at Austin

Lama Nachman
Intel Fellow, Director of Human & AI Systems Research Lab, Intel

Syam Nair
Chief Technology Officer, Zscaler

Mark Nitzberg
Executive Director, UC Berkeley Center for Human-Compatible AI,

Vijoy Pandey
Senior Vice-President, Outshift by Cisco, Cisco Systems

Louis Poirier
Vice-President AI/ML, C3 AI

Victor Riparbelli
Co-Founder and Chief Executive Officer, Synthesia

Jason Ruger
Chief Information Security Officer, Lenovo

Daniela Rus
Director, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT)

Noam Schwartz
Chief Executive Officer and Co-Founder, Activefence

Jun Seita
Team Leader (Principal Investigator), Medical Data Deep Learning Team, RIKEN

Susannah Shattuck
Head, Product, Credo AI

Paul Shaw
Group Security Officer, Dentsu Group

Evan Sparks
Chief Product Officer, AI, Hewlett Packard Enterprise

Catherine Stihler
Chief Executive Officer, Creative Commons

Fabian Theis
Science Director, Helmholtz Association

Li Tieyan
Chief AI Security Scientist, Huawei Technologies

Kush Varshney
Distinguished Research Scientist and Senior Manager, IBM

Lauren Woodman
Chief Executive Officer, DataKind

Yuan Xiaohui
Senior Expert, Tencent Holdings

Grace Yee
Director, Ethical Innovation, AI Ethics, Adobe

Michael Young
Vice-President, Products, Private AI

Leonid Zhukov
Vice-President, Data Science, BCGX; Director of BCG Global AI Institute, Boston Consulting Group

World Economic Forum

John Bradley
Lead, Metaverse Initiative

Karyn Gorman
Communications Lead, Metaverse Initiative

Devendra Jain
Lead, Artificial Intelligence, Quantum Technologies

Jenny Joung
Specialist, Artificial Intelligence and Machine Learning

Daegan Kingery
Early Careers Programme, AI Governance Alliance

Connie Kuang
Lead, Generative AI and Metaverse Value Creation

Hannah Rosenfeld
Specialist, Artificial Intelligence and Machine Learning

Stephanie Teeuwen
Specialist, Data and AI

Karla Yee Amezaga
Lead, Data Policy and AI

Hesham Zafar
Lead, Digital Trust

IBM

Jesús Mantas
Global Managing Director

Christina Montgomery
Chief Privacy & Trust Officer

Production

Laurence Denmark
Creative Director, Studio Miko

Sophie Ebbage
Designer, Studio Miko

Martha Howlett
Editor, Studio Miko

Endnotes

1. IBM AI Ethics Board, *Foundation models: Opportunities, risks and mitigations*, 2023, https://www.ibm.com/downloads/cas/E5KE5KRZ.
2. Solaiman, Irene, "The Gradient of Generative AI Release: Methods and Considerations", *Hugging Face*, 2023, https://arxiv.org/abs/2302.04844.
3. Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic et al., "Deep Reinforcement Learning from Human Preferences", *arxiv*, 17 February 2023, https://arxiv.org/pdf/1706.03741.pdf.
4. Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard et al., "RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback", *Google Research*, 1 December 2023, https://arxiv.org/pdf/2309.00267.pdf.
5. Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayah Kapoor et al, "The Foundation Model Transparency Index", *Stanford Center for Research on Foundation Models and Stanford Institute for Human-Centered Artificial Intelligence*, 2023, https://arxiv.org/pdf/2310.12941.pdf.
6. Smith, Larry, "Shift-left testing", *Association for Computing Machinery Digital Library*, 2001, https://dl.acm.org/doi/10.5555/500399.500404.
7. Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai et al., "Red Teaming Language Models with Language Models", *Association for Computational Linguistics*, 2022, https://aclanthology.org/2022.emnlp-main.225.pdf.
8. Hasani, Ramin, Mathias Lechner, Alexander Amini, Daniela Rus et al., "Liquid Time-constant Networks", *arxiv*, 2020, https://arxiv.org/pdf/2006.04439.pdf.
9. Xiao, Wei, Ramin Hasani, Xiao Li and Daniela Rus, "BarrierNet: A Safety-Guaranteed Layer for Neural Networks", *Massachusetts Institute of Technology*, 2021, https://arxiv.org/pdf/2111.11277.pdf.
10. Willig, Moritz, Matej Zecevic, Devendra Singh Dhami and Kristian Kerting, "Can Foundation Models Talk Causality?", *arxiv*, 2022, https://arxiv.org/pdf/2206.10591.pdf.
11. Roy, Kaushik, Yuxin Zi, Vignesh Narayanan, Manas Gaur and Amit Seth, "Knowledge-Infused Self Attention Transformers", *arxiv*, 2023, https://arxiv.org/pdf/2306.13501.pdf.

2/3 AI Governance Alliance Briefing Paper Series 2024

Unlocking Value from Generative AI: Guidance for Responsible Transformation

IN COLLABORATION WITH IBM CONSULTING

Images: Getty Images, MidJourney

Contents

Executive summary

Organizations should emphasize responsible transformation with generative AI to build a sustainable future.

Generative AI entered the popular domain with the launch of OpenAI's ChatGPT in November 2022, igniting global fascination surrounding its capabilities and potential for transformative impact. As generative AI's technical maturity accelerates, its adoption by organizations seeking to capitalize on its potential is maturing at pace while also swiftly disrupting business and society and forcing leaders to rethink their strategies in real time. This paper addresses the impact of generative AI on industry and introduces best practices for responsible transformation.

Leaders have realized new generative AI opportunities for their organizations, from streamlining enterprise processes to supporting artists in reimagining furniture design or even aiding nations in addressing global climate challenges. From the public to the private sector, organizations are witnessing generative AI's ability to enhance enterprise productivity, create net new products or services, and redefine industries and societies. In adopting generative AI, leaders report a shift towards a use-case-based approach, focusing on evaluating and prioritizing use cases and structures that enable the successful deployment of generative AI technologies and compound value generation.

Organizations should evaluate potential use cases across the following domains: business impact, organisational readiness and investment strategy.

- Strategic alignment with the organization's goals, revenue and cost implications, and impact on resources are key factors when leaders prioritize use cases based on their potential for **business impact**.

- The requisite technical talent and infrastructure, the ability to track data and model lineage, and the governance structure to manage risk are considerations when leaders evaluate use cases against their **operational readiness**.

- Balancing upfront development cost with reusability potential, projected time to value and an increasingly complex regulatory environment are criteria when leaders select use cases in alignment with an organization's **investment strategy**.

Following use case selection, organizations weigh benefits against downstream impacts such as impact to the workforce, sustainability or inherent technology risk such as hallucinations. A multistakeholder approach helps leaders to mitigate risk and scale responsibly.

- Multistakeholder governance with distributed ownership is central to **addressing accountability**.

- Communications teams that shape a cohesive narrative are essential to **addressing trust** through transparency.

- Operational structures that roadmap and cascade use cases to extract, realize, replicate and amplify value across the entire organization are key to **addressing challenges to scale**.

- Value-based change management is critical to **addressing human impact** and ensuring the workforce remains engaged and upskilled.

The findings in this briefing paper provide leaders with insights on how to realise the benefits of generative AI while mitigating its downstream impacts. Future publications will build on these recommendations for responsible transformation as generative AI becomes increasingly able to mimic human skills and reasoning, and technology advances in pursuit of artificial general intelligence.

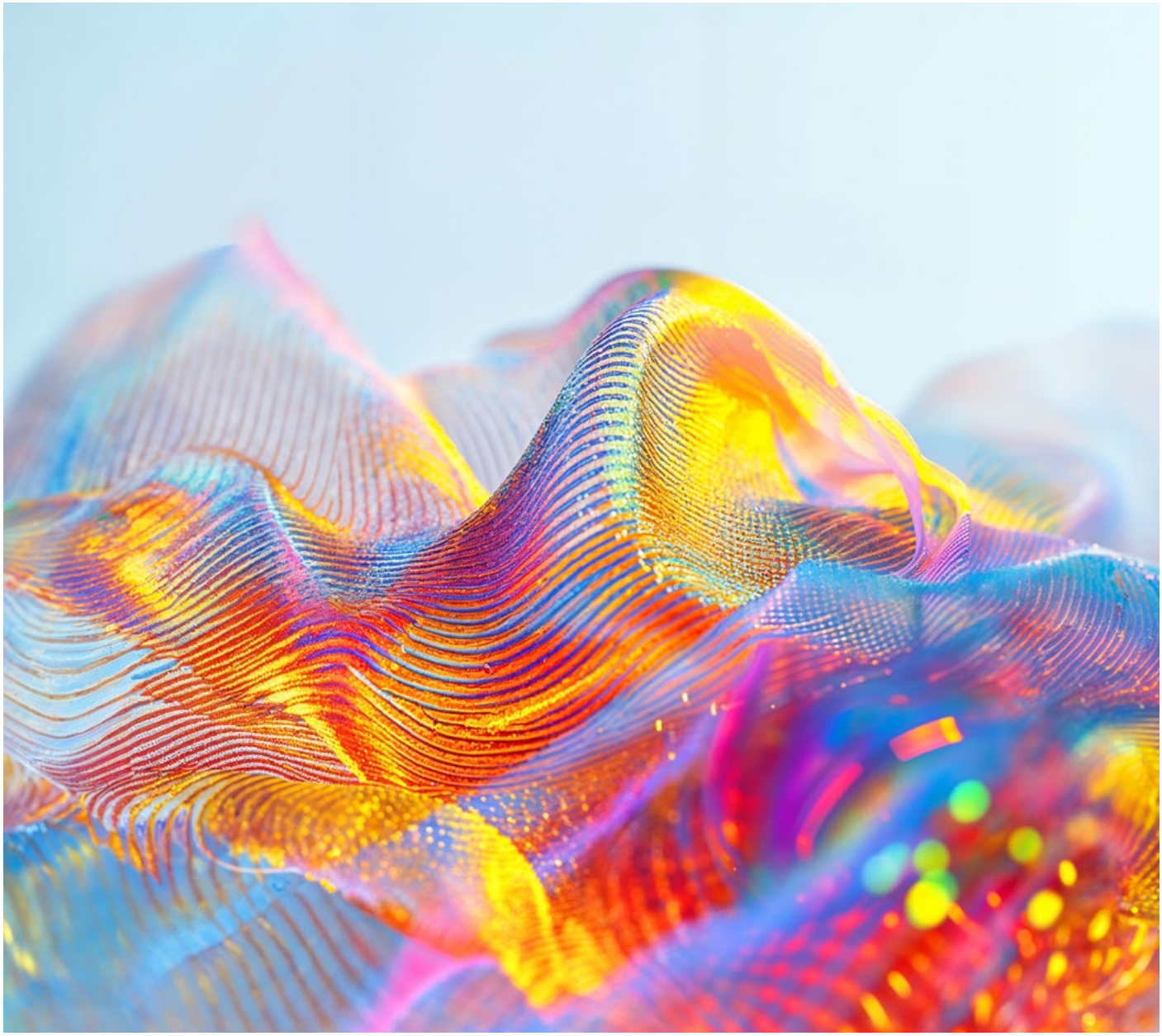
Introduction

Generative AI raises new questions about responsible transformation for industry executives, government leaders and academia.

Generative artificial intelligence (AI) has captured global imagination with its human-like capabilities and has shown the potential to elevate creativity, amplify productivity, reshape industries and enhance the human experience. As a result, cross-sector executives, government leaders and academia are considering the potential impact of this technology as they weigh answers to critical questions:

- Where are the growing opportunities and novel application areas to drive sustainable economic growth?
- What are the new challenges and downstream impacts?
- What are the best practices for scaling responsibly and bringing about exponential transformation?

Finally, as the curiosity to replicate or even exceed human intelligence grows in the future, what does this mean for organizations seeking to capitalize on the opportunities offered by this technology?



1 New opportunities with generative AI

Generative AI creates new opportunities but requires a distinctive approach to value generation focused on use cases and experimentation.

Generative AI is expected to unlock opportunities that will significantly impact the global economy. Organizations are already using generative AI to enhance existing products, services, operations and provide hyper-personalized customer experiences. While most use cases focus on boosting human capabilities, some have the potential to radically accelerate benefits to humanity. For example, novel synthetic protein structures generated to help fix DNA errors can significantly accelerate the creation of new cancer therapies.[1] Generative AI is also used to orchestrate deep synthesis of numerous data catalogues to enable work to protect the oceans.[2] These bolder bets have the potential to reshape not just entire industries but economies and societies at large. In general, use cases can be considered under different categories that include enhancing enterprise productivity, creating new products or services and, eventually, redefining industries and societies.

TABLE 1 **Snapshot of sample generative AI case studies in the market**

Category	Company	Challenge	Action	Impact
Enhancing enterprise productivity	**Brex:** automating corporate card expenses[3]	Support corporate card customers to categorize transactions and add notes to meet company policies and Internal Revenue Service (IRS) compliance.	Brex, with OpenAI and Scale, used generative AI to create the Brex Assistant to streamline expense reporting, automatically classify expenses and create IRS-compliant notes.	Brex Assistant fully handles 51% of card swipes, saving time and improving expense accuracy and compliance. It generated over 1.4 million receipts and 1 million receipt memos.
Enhancing enterprise productivity	**IKEA:** reimagining furniture design[4]	Seek creative solutions to aid furniture designers in crafting new designs inspired by their iconic past.	IKEA and SPACE10 used generative AI to explore furniture design concepts, training a model on 1970s and 1980s catalogues for students to create future-focused designs inspired by the past.	Furniture designers collaborate with AI, expanding design possibilities and speeding up cycles.
Enhancing enterprise productivity and net-new product or service	**Google:** streamlining software prototyping[5]	Reduce software development cycles internally and simplify access to generative AI models.	Google created Google AI Studio, a generative AI tool to simplify software prototyping and democratize access to their foundation models, which were first used internally.	Increased proactive UX and product prototyping, provided an efficient UI for easy model prompting and was later launched as a new product in 179 countries and territories.
Net-new product or service	**Synthesia and PepsiCo:** reinventing the football fan experience[6]	Connect brand and performance marketing efforts into one seamless experience.	Fans could generate and share personalized videos using Lionel Messi's AI avatar in eight languages, bypassing traditional production limits.	Seven million videos weregenerated, attracting over 38 million website visits in 24 hours.

TABLE 1 | **Snapshot of sample generative AI case studies in the market (continued)**

Category	Company	Challenge	Action	Impact
Redefining industries and societies	**Insilico Medicine:** accelerating drug discovery[7,8]	Discover and develop new treatments for serious diseases more quickly and cheaply compared to traditional processes.	Generative AI was used during the preclinical drug discovery process to identify a novel drug candidate for idiopathic pulmonary fibrosis.	A preclinical drug candidate was discovered in less than 18 months and at one-tenth of the cost of a conventional programme. The drug candidate has now entered phase two trials.
Redefining industries and societies	**NASA and IBM:** unique global planning for climate phenomena and sustainability[9]	Build a unique foundation model to generate insights from over 250 terabytes (TBs) of mission satellite imagery.	NASA and IBM created the first open-source geospatial foundation model, available via Hugging Face, using NASA data to enhance and democratize global environmental research and planning.	The model is estimated to increase geospatial analysis speed by four times with 50% less labelled data; used to solve global climate challenges, including reforestation in Kenya and other development efforts in the Global South.

Organizations are shifting towards smaller, use-case based approaches that emphasize ideation and experimentation.

The speed of adoption and implementation of generative AI is unparalleled to any other technological advancement. The technology is no longer dependent on the manual labelling of significant amounts of data – often the most time-consuming and costly part of traditional AI workflows.

Across the board, leaders report a new approach to generative AI opportunities that extends beyond rapid proofs of concept (POCs) based on large models. Instead, organizations are shifting towards smaller, use-case based approaches that emphasize ideation and experimentation. They are involving the workforce in the use case discovery and ideation process. Smaller use cases with low complexity are often applied first, allowing leaders to find value while minimizing downstream implications. In either case, leaders start with diverse POCs, which are scaled across the enterprise once value is proven.

In many instances, generative AI experiments may yield unexpected learnings about where value, and often also cost and challenges, truly lie. Organizations may realize the compound benefits of generative AI when implementing it in tandem with technologies such as causal AI models[10] to increase explainability, advances in quantum technologies to accelerate the generative AI life cycle, or 5G to increase reach. These compounding benefits will help organizations to prioritize use cases for adoption.

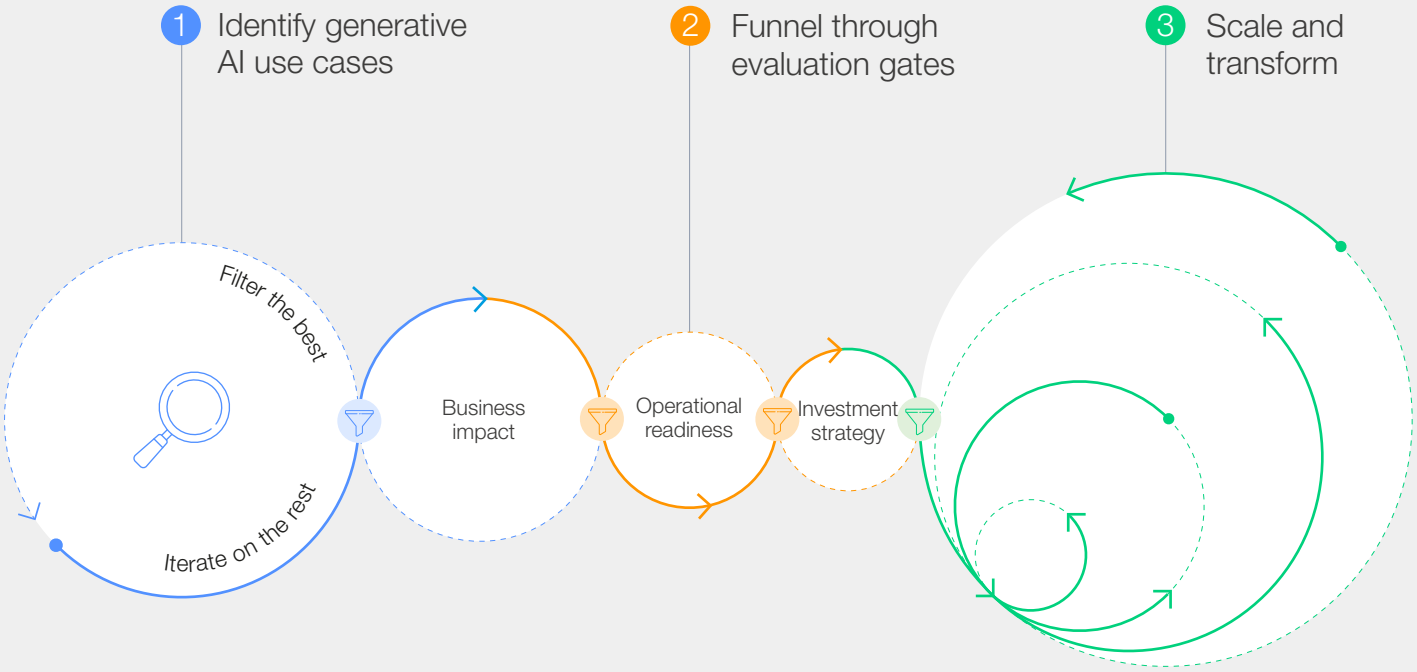
2

Assessing use cases for adoption

Generative AI use cases may be assessed by business impact, organizational readiness and investment strategy prior to adoption.

As organizations consider generative AI, they must assess all factors involved to move a use case from concept to implementation. Leaders need to ensure that each use case benefits the organization, its customers, its workforce and/or society. While evaluation criteria can differ between organizations, the following gates comprise the most common approaches adopted by industry leaders to evaluate the viability and value-generation potential of use cases. The order is not sequential and can differ depending on each organization and use case.

FIGURE 1 **Funnelling use cases through evaluation gates**



2.1 Evaluation gate: business impact

Leaders evaluate the use case's value alignment with the organization's strategic objectives and its stakeholder responsibility. After alignment on the outcomes and generative AI as the best technology to address a specific use case, the impact of each use case on an organization can be categorized as follows:

1. **Scaling human capability** by enhancing productivity and existing human skills (e.g. near instant new content generation for rapid idea iteration; creation of multiple versions of an advertising campaign).

2. **Raising the floor** by increasing accessibility to technologies and capabilities previously requiring specific resources, skills and expertise (e.g. giving everyone the ability to code).

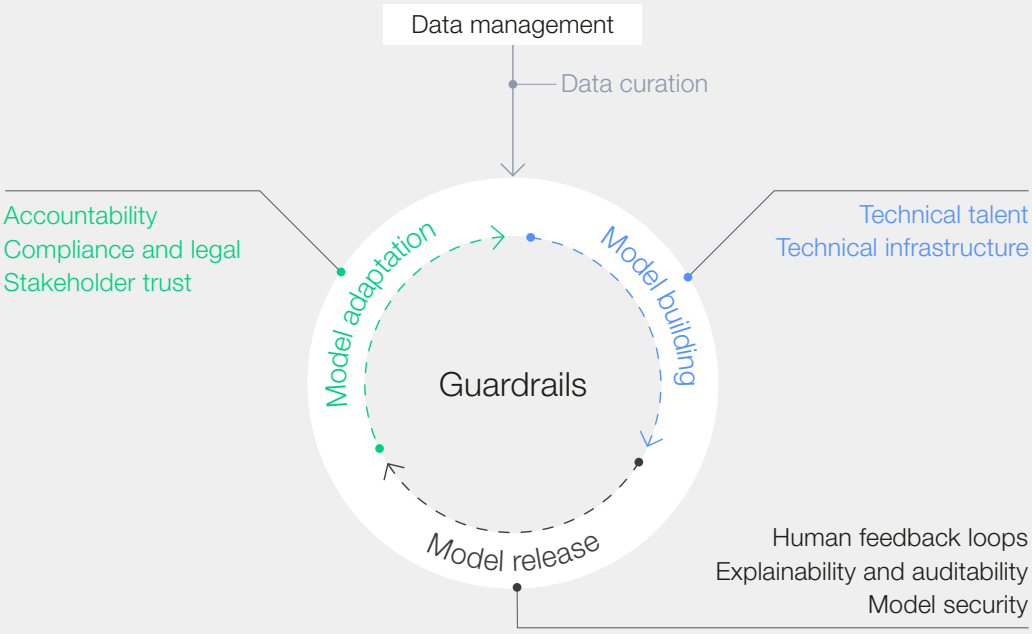
3. **Raising the ceiling** by solving problems thus far unsolvable by humans (e.g. generating new molecular structures, which could aid the creation of novel and more effective therapeutic agents.[11]

Generative AI opportunities have created strong competitive pressures and inaction can come with significant opportunity costs.[12] In industries such as marketing or consumer goods, understanding the criticality of time to market and improved experience for users, helps leaders prioritise use cases and resource allocation. Reputation is another important consideration – will the use case enhance the organization's brand as a pioneer of innovation? Enabling the workforce to access generative AI tools can be an important factor for talent attraction and retention. When generative AI performs administrative tasks that previously required significant time and effort, the workforce can repurpose their time from rote activities to those that allow them to explore their creativity and hone their unique skillset.

2.2 Evaluation gate: operational readiness

Responsible adoption of generative AI requires operational readiness for technological dependencies and outcomes. Before organizations expose generative AI to their data, data curation is essential to ensure it is accurate, secure, representative and relevant. In developing or implementing generative AI technologies, organizations must consider if they have the right technical talent and infrastructure, such as appropriate models and necessary computing power. In deploying generative AI technologies, organizations should ensure human feedback loops are in place to mitigate risks by ensuring user feedback is elicited, standardized and incorporated into the continuous fine-tuning of the model. Additionally, organizations require the ability to track model lineage and data sources that inform model outputs, as well as vet models and systems for cybersecurity robustness.

FIGURE 2 **Operational readiness considerations (non-exhaustive) across the model life cycle**



Organizations will be held responsible for the outcomes of their AI technology and must, therefore, ensure compliance with the global complexity of regulation and policies as cited in *Generative AI Governance: Shaping the Collective Global Future*.[13] This will require new skills and roles for accountability, compliance and legal responsibilities as a multistakeholder approach. Generative AI's evolutionary nature and its inherent potential for downstream implications create a greater need to continually evaluate even if the necessary guardrails are in place. Finally, organizations need a plan to enhance stakeholder trust with a technology that can elicit great scepticism to ensure their workforce, customers and other critical parties responsibly adopt generative AI.

2.3 Evaluation gate: investment strategy

While investment considerations are important to any organizational decision-making, they are particularly significant for generative AI opportunities. Use cases often require a higher upfront investment, the regulatory environment is becoming increasingly complex and the technology is evolving at a rapid pace.

When prioritizing use cases, leaders must consider if each merits the use of models adopted from open-source communities, acquired from other third parties or developed in-house. Model selection must account for alignment with the use case, speed to market, requisite resource investments, including capital and talent, licensing and acceptable use policies, risk exposure and competitive differentiation offered by each option.

Leaders evaluate the reusability potential of a use case across the organization, as it can offset development costs and curtail sustainability footprints. Additionally, they evaluate whether the use case can operate viably within the current regulatory environment and whether the organization can monitor compliance to minimize legal risk. This can require significant investment of capital and human resources, such as developers, lawyers, senior leadership and ethics boards.

Talent availability is central to an organization's investment strategy as well. Total investment may include upskilling, re-skilling or hiring additional employees with appropriate generative AI skills, such as content creation, model development or model tuning.

Following the evaluation of use cases by business impact, organizational readiness and investment strategy, the next step is to implement and scale selected use cases. How can they maximize opportunities while mitigating risks to ensure a responsible and successful transformation?

3 Responsible transformation

A multistakeholder approach creates value while balancing challenges of trust, accountability, scale and the workforce.

3.1 The case for responsible transformation

As *The Presidio Recommendations on Responsible Generative AI* detail, responsible transformation requires specific considerations for generative AI's unique capabilities, along with multistakeholder collaboration and proper steering during the transformation journey. Global generative AI regulations and standards (NIST et al.) are changing, and so the current need for self-governance is shared by organizations and leaders. There is also a need to ensure that the technology is accessible to all. Organizations are committed to aligning with global environmental and sustainability goals, pledging to adopt AI in a responsible and accessible manner.

The lack of responsibility in an organization's transformation can have many negative consequences, which are multi-fold and compounded for a technology as revolutionary as generative AI. From perpetuating biases, introducing security vulnerabilities and spreading misinformation – causing severe reputational damage – irresponsible generative AI applications and practices not only threaten the organization itself but can also negatively impact society at speed and scale.

Generative AI comes with several downstream implications associated with more traditional forms of AI, together with amplified and new ones. The following are most often noted for their potential impact, with a further list to be explored in future work.

1. **Workforce and talent impact**
   While AI is commonly used to automate tasks, the scale at which generative AI can accomplish this amplifies its impact on the workforce. The potential risk of job displacement presents significant challenges for society that can exacerbate inequality. Research indicates that generative AI's automation capabilities provide the greatest exposure for clerical jobs, which have traditionally been held by women. In some cases, particularly in developing countries, these types of jobs may cease to exist, removing an avenue that has historically served as an entry for women into the labour market.[14] Additionally, generative AI's novel capability to create, generate and simulate human-like interactions may now overlap with tasks in creative industries, and its ability to rapidly learn domain expertise may influence the roles of knowledge workers.

   Skills and workloads are changing, and organizational structures need to evolve at pace.[15] Generative AI is profoundly changing the way employees view their jobs and the value work brings. Nevertheless, the technology presents a unique opportunity for organizations to re-evaluate their working practices and skills: to inspire, incentivize, motivate, upskill and reskill workers, while evaluating the agility of their own organizational structures.

2. **Hallucination impact**
   Generative AI introduces the risk of hallucinations, which can propagate misinformation, leading to confusion, mistrust and even potential harm. Equally, hallucinations are a corollary of generative AI's capability to create net-new content, which is central to its power to accelerate creativity. Organizations need to understand whether the benefit of content creation outweighs the risk of hallucination for each use case.

   Hallucinations are particularly concerning when generative AI outputs appear authoritative but are factually inaccurate, especially when used to influence decision-making that may impact global communities in areas such as health, politics and science. Organizations that rely on digital content production or customer engagement face challenges as brand reputation and customer trust could be damaged. Guardrails from *Presidio AI Framework: Towards Safe Generative AI Models* need to be considered and embedded in the process.[16]

3. **Sustainability impact**
   Training and fine-tuning generative AI models demand very high energy consumption.[17] Growing global efforts to offset or mitigate their sustainability footprint are ongoing, such as advancements in model, runtime and hardware

optimization, as well as improved education on model choices. Algorithmic approaches like federated computing can further minimize the energy consumption of data collection and processing. Organizations also consider their choices in data needs as a growing move towards smaller, more targeted, and more energy-efficient models underlines.

In addition to ensuring generative AI models are more sustainable, the technology itself can be used to improve sustainability, for example, through use cases focussed on energy modelling and supply chain optimization.[18]

As the risks associated with generative AI amplify and expand, traditional organizational structures need to pivot with agility. Leaders need to ensure cross-functional connectivity from the board level down and across all impacted functions. The following are four interconnected and interdependent functions that support this organizational effort to balance the opportunities and benefits of generative AI with its downstream impacts as organizations implement and scale generative AI applications.

3.2 Addressing accountability: defined governance for immediate and downstream outcomes

An AI ethics council modelled on value-based principles is indispensable for any organization.

Multistakeholder governance with distributed ownership is central to responsible transformation in the age of generative AI. This approach is characteristic of industry leaders, with legal, governance, IT, cybersecurity, human resources (HR), as well as environmental and sustainability representatives requiring a seat at the table to ensure responsible transformation across the organization. The positive and negative externalities of generative AI expand the conventional responsibilities in governance towards a more holistic, human-centred and values-driven approach.

An AI ethics council modelled on value-based principles[19] is indispensable for any organization; larger organizations appoint members from their stakeholder and shareholder groups, while smaller organizations may need to rely on a limited committee or an external ethics council. Councils must collaborate with stakeholders on aspects such as workplace policies, even if they do **not** deploy generative AI, as the workforce is likely already using it at work on personal devices. The council should expand to incorporate a diverse set of members from across the entire organization to ensure the responsible adoption of not just individual use cases but also emerging and intersecting strategies on open technologies, artificial general intelligence (AGI), 5G and quantum technology.

The evolving nature of generative AI requires rigorous self-regulation and internal AI governance leads may serve as the sentinels of the organization. Generative AI supports human-led analysis in regulatory, environmental and sustainability efforts. It assists in algorithm monitoring and policy formulation, but crucially, it requires human oversight to ensure responsible and effective application, addressing potential risks and maintaining quality outcomes.

3.3 Addressing trust: enabling transparency through communication

Generative AI evokes mixed reactions from stakeholders, placing a high demand on communications teams. These teams shape a cohesive narrative to showcase how their organization optimizes transparency, explainability, coherence and trustworthiness on a use case basis. They play a role in educating stakeholders and shareholders on the capabilities and fallibilities of the technology while managing expectations. They can inspire and instruct end-users about the benefits on the horizon, thus building trust and increasing adoption.

External communications need to assuage stakeholders that seek innovation, but not at the cost of ethical behaviour, trust and actions that prove that the organization is committed to the greater good of humanity. Internal accountability and advocacy are needed from top leadership to obtain buy-in from the workforce and establish a culture that benefits from generative AI. Examples of effective trust programmes include taking a prominent ethics stance in policy or the executive community, buddy programmes for all employees seeking (generative) AI immersion and novel career pathways that can lead to increased trust and ownership from the workforce.

3.4 Addressing challenges to scale: diverse and agile operations structures

Initial adoption of generative AI across organizations has focused on targeted, often isolated, use cases. However, as leaders plan their strategic roadmaps, many are challenged with how to scale these use cases across their organizations to realize the compound benefits of generative AI.

Operations teams are the primary implementers of use cases. Data analysts, research and development teams, resource managers, HR executives and business leaders ensure use cases are roadmapped and cascaded across the organization for maximum benefit. In their initial development, use cases require a diverse operational structure to ensure a multistakeholder approach to extracting, realizing, replicating and amplifying value. However, as use cases become integrated and scale, an interlocking and agile operational structure is needed to understand how compound value can be unlocked, and corollary impacts to other parts of the workforce or other lines of business can be anticipated.

3.5 Addressing human impact: value-based change management

Technologies that develop as rapidly as generative AI require adoption by a workforce that evolves at pace. The implications of generative AI on the workforce are central to business and need to be managed well. The chief human resources officer, the chief information officer, and the chief financial officer teams should come together to support the workforce as needed when implementing and scaling generative AI use cases.

Leaders plan and implement talent transformation while ensuring staff have access to the necessary technological tools and training. This starts with communicating the vision for generative AI pilots that clearly states desired benefits for customers and employees alike, together with emerging professional development pathways for staff. Competencies, capabilities and skills are rapidly evolving as generative AI use cases are implemented across the organization.

Change management responsibilities across the organization are significant. HR professionals engage with the implementation of use cases from the beginning so they can proactively assess the impact on staff and put workforce transformation plans in place. Including employees in idea generation for use cases and encouraging them to own their career paths can increase engagement. Hackathons and company-wide training days are effective in upskilling the workforce while also encouraging experimentation and innovation.

The immense potential of generative AI for benefit as well as for harm requires that all four of these primary functions are dynamic, interlocked and in equilibrium. The effectiveness of this interlock correlates directly with the extent to which an organization scales generative AI applications responsibly.

“Technologies that develop as rapidly as generative AI require adoption by a workforce that evolves at pace.

Conclusion

New technologies driving productivity have always been positioned as repurposing workers to higher-value work, which has traditionally required human oversight and creativity. However, with generative AI becoming increasingly advanced in its ability to mimic human skills and capabilities, it opens more questions about its impact on the organizations choosing to adopt it. Technological advances towards human reasoning in the pursuit of artificial general intelligence demand ongoing discourse on the responsibility of organizations to their workforce, customers and wider society.

Future work through the World Economic Forum’s AI Governance Alliance will build on this foundation and address essential considerations, such as internal metrics for responsibility, understanding organizational barriers to responsible transformation, as well as broader issues such as intellectual property, regulatory alignment and workforce considerations. Generative AI is reimagining the status quo for every organization. Providing a roadmap for organizations that guides them to innovate responsibly is key to adopting and scaling this powerful technology.

Contributors

This paper is a combined effort based on numerous interviews, discussions, workshops and research. The opinions expressed herein do not necessarily reflect the views of the individuals or organizations involved in the project or listed below. Sincere thanks are extended to those who contributed their insights via interviews and workshops, as well as those not captured below.

World Economic Forum

Hubert Halopé
Lead, Artificial Intelligence and Machine Learning

Devendra Jain
Lead, Artificial Intelligence, Quantum Technologies

Daegan Kingery
Early Careers Programme, AI Governance Alliance

Connie Kuang
Lead, Generative AI & Metaverse Value Creation

Benjamin Larsen
Lead, Artificial Intelligence and Machine Learning

Cathy Li
Head of AI, Data and Metaverse; Deputy Head, Centre for the Fourth Industrial Revolution; Member of the Executive Committee

AI Governance Alliance Project Fellows

Ann-Sophie Blank
Managing Consultant, IBM

Alison Dewhirst
Senior Managing Consultant, IBM

Heather Domin
Executive Fellow, Director of Responsible AI Initiatives, IBM

Sophia Greulich
Senior Consultant, IBM

Michelle Hannah Jung
Senior Managing Consultant, IBM

Jennifer Kirkwood
Executive Fellow, Partner, IBM

Avi Mehra
Associate Partner, IBM

Sandra Misiaszek
Associate Partner, IBM

Acknowledgements

Sincere appreciation is extended to the following working group members, who spent numerous hours providing critical input and feedback to the drafts. Their diverse insights are fundamental to the success of this work.

Martin Adams
Co-Founder, METAPHYSIC

Basma AlBuhairan
Managing Director, Centre for the Fourth Industrial Revolution, Saudi Arabia

Uthman Ali
Senior Product Analyst, AI Ethics SME, BP

Mohamed Alsharid
Chief Digital Officer, Dubai Electricity and Water Authority (DEWA)

Stefan Badža
Director, Team for Special Projects, Office of the Prime Minister of Serbia

Ricardo Baptista Leite
Chief Executive Officer, Health AI, The Global Agency for Responsible AI in Health

Elisabeth Bechtold
Head, AI Governance, Zurich Insurance Group

Sébastien Bey
Senior Vice-President and Global Head of IT at Siemens Smart Infrastructure, Siemens

Lu Bo
Vice-President; Head, Corporate Strategy, Lenovo Group

Ting Cai
Group Senior Managing Executive Officer; Chief Data Officer, Rakuten Group

Cansu Canca
Director, Responsible AI Practice, Institute for Experiential AI, Northeastern University

Nadia Carlsten
Vice-President, Product, SandboxAQ

Will Cavendish
Global Digital Services Leader, Arup Group

Rohit Chauhan
Executive Vice President, AI & Security Solutions, Mastercard International

Adrian Cox
Managing Director, Thematic Strategist, Deutsche Bank Research, Deutsche Bank

Bhavesh Dayalji
Chief Executive Officer, Kensho Technologies

Evren Dereci
Chief Executive Officer, KocDigital

Dan Diasio
Global Artificial Intelligence Consulting Leader, EY

P. Murali Doraiswamy
Professor of Psychiatry and Medicine, Duke University School of Medicine

Elena Fersman
Vice-President and Head of Global AI Accelerator, Ericsson

Ryan Fitzpatrick
Senior Vice-President, Strategy, Vindex

Lucas Glass
Vice-President, Analytics Center of Excellence, IQVIA

Mark Gorenberg
Chair, Massachusetts Institute of Technology (MIT)

Mark Greaves
Executive Director, AI2050, Schmidt Futures

Olaf Groth
Professional Faculty, Haas School of Business, University of California, Berkeley

Sandeep Grover
Trust & Safety Leadership, TikTok

Sangeeta Gupta
Senior Vice-President, National Association of Software and Services Companies (NASSCOM)

Bill Higgins
Vice-President, watsonx Platform Engineering and Open Innovation, IBM

Matissa Hollister
Assistant Professor of Organizational Behaviour, McGill University

Michael G. Jacobides
Professor of Strategy; Sir Donald Gordon Professor of Entrepreneurship and Innovation, London Business School

Fariz Jafarov
Executive Director, Centre for the Fourth Industrial Revolution, Azerbaijan

Reena Jana
Head, Content & Partnership Enablement, Responsible Innovation, Google

Jeff Jarvis
Professor, Graduate School of Journalism, City University of New York

Emilia Javorsky
Director, Futures Program, Future of Life Institute

Siddhartha Jha
AI and Digital Innovation Lead, Botnar Foundation

Shailesh Jindal
Vice-President of Corporate Strategy, Palo Alto Networks

Athina Kanioura
Executive Vice-President, Chief Strategy and Transformation Officer, PepsiCo

Vijay Karunamurthy
Head and Vice-President, Engineering, Scale AI

Sean Kask
Chief AI Strategy Officer, SAP

Faisal Kazim
Head, Centre for the Fourth Industrial Revolution, United Arab Emirates

Rom Kosla
Chief Information Officer, Hewlett Packard Enterprise

Nikhil Krishnan
Chief Technology Officer, Products, C3 AI

Sebastien Lehnherr
Chief Information Officer, SLB

Giovanni Leoni
Head, Business Development and Strategy, Credo AI

Art Levy
Chief Strategy Officer, Brex

Leland Lockhart
Director, Artificial Intelligence & Machine Learning, Vista Equity Partners

Harrison Lung
Group Chief Strategy Officer, e&

Manny Maceda
Chief Executive Officer, Bain & Company

Jim Mainard
Chief Technology Officer and Executive Vice-President, Deep Technology, XPRIZE Foundation

Naveen Kumar Malik
Chief of Staff, Office of the Chief Technology Officer, HCL Technologies

Thomas W. Malone
Professor of Management and Director, Center for Collective Intelligence, MIT Sloan School of Management

Darren Martin
Chief Digital Officer, AtkinsRéalis

Francesco Marzoni
Chief Data & Analytics Officer, Ingka Group (IKEA)

Darko Matovski
Chief Executive Officer, causaLens

Andrew McMullan
Chief Data and Analytics Office, Commonwealth Bank of Australia

Nicolas Miailhe
Founder and President, The Future Society (TFS)

Steven Mills
Partner and Chief Artificial Intelligence Ethics Officer, Boston Consulting Group

Angela Mondou
President and Chief Executive Officer, TECHNATION

Yao Morin
Chief Technology Officer, JLL

Mashael Muftah
International and Regional Organizations Adviser, Ministry of Information and Communication Technology (ICT) of Qatar

Abhishek Pandey
Global Head of Services Business Development, GEP

Charna Parkey
Real-Time AI Product and Strategy Leader, DataStax

Cyril Perducat
Senior Vice-President and Chief Technology Officer, Rockwell Automation

Andreas Prøsch
Vice-President and Head, Aker AI Unit, Aker ASA

Philippe Rambach
Chief AI Officer, Schneider Electric

Mary Rozenman
Chief Financial Officer and Chief Business Officer, Insitro

Crystal Rugege
Managing Director, Centre for the Fourth Industrial Revolution, Rwanda

Prasad Sankaran
Executive Vice-President, Software and Platform Engineering, Cognizant Technology Solutions US

Isa Scheunpflug
Head, Automation Office, UBS

Mikkel Skovborg
Senior Vice-President, Innovation, Novo Nordisk Foundation

Genevieve Smith
Founding Co-Director, Responsible & Equitable AI Initiative, Berkeley Artificial Intelligence Research Lab (UC Berkeley)

Eric Snowden
Vice-President, Design, Digital Media, Adobe

Jim Stratton
Chief Technology Officer, Workday

Murali Subbarao
Vice-President, Generative AI Solutions, ServiceNow

Norihiro Suzuki
Chairman of the Board, Hitachi Research Institute, Hitachi

Behnam Tabrizi
Co-Director and Teaching Faculty of Executive Program, Stanford University

Amogh Umbarkar
Vice-President, SAP Product Engineering, SAP

Ingrid Verschuren
Executive Vice-President, Data and AI; General Manager, Europe, Middle East and Africa, Dow Jones

Daniel Verten
Strategy Partner, Synthesia

Judy Wade
Managing Director, CPP Investments

Anna Marie Wagner
Senior Vice-President, Head of AI, Ginkgo Bioworks

Min Wang
Chief Technology Officer, Splunk

Amy Webb
Chief Executive Officer, Future Today Institute

Chaoze Wu
Head of R&D Department, Managing Director, China Securities

Joe Xavier
Chief Technology Officer, Grammarly

Alice Xiang
Global Head, AI Ethics, Sony

Zhang Ya-Qin
Chair Professor and Dean, Tsinghua University

Zhang Ying
Professor of Marketing and Behavioral Science, Guanghua School of Management, Peking University

Zhang Yuxin
Chief Technology Officer, Huawei Cloud, Huawei Technologies

Yijie Zeng
Chief Technology Officer, Beijing Langboat Technology

World Economic Forum

John Bradley
Lead, Metaverse Initiative

Karyn Gorman
Communications Lead, Metaverse Initiative

Jenny Joung
Specialist, Artificial Intelligence and Machine Learning

Hannah Rosenfeld
Specialist, Artificial Intelligence and Machine Learning

Supheakmungkol Sarin
Head, Data and Artificial Intelligence Ecosystems

Stephanie Teeuwen
Specialist, Data and AI

Karla Yee Amezaga
Lead, Data Policy and AI

Hesham Zafar
Lead, Digital Trust

IBM

Phaedra Boinodiris
Associate Partner

Frank Madden
Privacy and Regulatory Risk Adviser

Jesús Mantas
Global Managing Director

Christina Montgomery
Chief Privacy & Trust Officer

Catherine Quinlan
Vice-President, AI Ethics

Sencan Sengal
Distinguished Engineer

Jamie VanDodick
Director AI Ethics and Governance

Production

Laurence Denmark
Creative Director, Studio Miko

Sophie Ebbage
Designer, Studio Miko

Martha Howlett
Editor, Studio Miko

Endnotes

1. Gordon, Rachel, "Generative AI imagines new protein structures", *MIT News*, 12 July 2023, https://news.mit.edu/2023/generative-ai-imagines-new-protein-structures-0712#:~:text=.
2. "Navigating the Ocean of Data: Harnessing the Power of Knowledge Graphs in Data Catalogs", *HUB Ocean*, n.d., https://www.hubocean.earth/blog/ocean-knowledge-graph.
3. "Brex Gives Every Employee an Expense Assistant with AI", *Brex*, September 2023, https://www.brex.com/journal/press/brex-gives-every-employee-an-expense-assistant-with-ai.
4. "Exploring the future with vintage designs in AI", *IKEA Newsroom*, 20 April 2023 https://www.ikea.com/global/en/stories/design/to-nytillverkad-and-beyond-ikea-space10-and-designers-of-tomorrow-explore-future-with-ai-230420/.
5. "Make with MakerSuite – Part 1: An Introduction", *Google for Developers*, 26 September 2023, https://developers.googleblog.com/2023/09/make-with-makersuite-part1-introduction.html. [Note: Google renamed the product to Google AI Studio on December 6, 2023].
6. "See our personalized Messi video campaign", *Synthesia Newsroom*, 30 October 2023, https://www.synthesia.io/post/messi.
7. "New Milestone in AI Drug Discovery: First Generative AI Drug Begins Phase II Trials with Patients", *Insilico Newsroom*, 1 July 2023, https://insilico.com/blog/first_phase2.
8. "Insilico: linking target discovery and generative chemistry AI platforms for a drug discovery breakthrough", *Nature Research Media*, n.d., https://www.nature.com/articles/d43747-021-00039-5.
9. "IBM and NASA are building an AI foundation model for weather and climate", *IBM Newsroom*, 30 November 2023, https://research.ibm.com/blog/weather-climate-foundation-model.
10. Definition for causal AI taken from: Forney, Andrew, "Casual Inference in AI Education: A Primer", *Journal of Causal Inference*, 2022, https://ftp.cs.ucla.edu/pub/stat_ser/r509.pdf.
11. Centre for Trustworthy Technology, *A New Frontier for Drug Discovery and Development: Artifical Intelligence and Quantum Technology*, n.d., https://c4tt.org/1155-2/.
12. "Building a Value-Driving AI Strategy for Your Business", *Gartner*, n.d., https://www.gartner.com/en/information-technology/topics/ai-strategy-for-business.
13. World Economic Forum, *Generative AI Governance: Shaping the Collective Global Future*, 2024.
14. International Labour Organization (ILO), *Generative AI and jobs: A global analysis of potential effects on job quantity and quality*, 2023, https://www.ilo.org/wcmsp5/groups/public/---dgreports/---inst/documents/publication/wcms_890761.pdf.
15. World Economic Forum, *The Future of Jobs Report 2023*, 2023, https://www.weforum.org/publications/the-future-of-jobs-report-2023/.
16. World Economic Forum, *Presidio AI Framework: Towards Safe Generative AI Models*, 2024.
17. Strubell, Emma, Ananya Ganesh and Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, Cornell University Department for Computer Science, Computation and Language, 5 June 2019, https://arxiv.org/abs/1906.02243.
18. "Generative AI: The Next Frontier in Energy & Utilities and Oil & Gas Innovation", *BirlaSoft Newsroom*, 26 October 2023, https://www.birlasoft.com/articles/generative-ai-frontier-energy-utilities-oilgas-innovation.
19. OECD AI Principles overview adopted in May 2019: "OECD AI Principles overview", *Organisation for Economic Co-operation and Development*, n.d., https://oecd.ai/en/ai-principles.

3/3 AI Governance Alliance Briefing Paper Series 2024

Generative AI Governance:
Shaping a Collective Global Future

IN COLLABORATION
WITH ACCENTURE

Images: Getty Images, MidJourney

Contents

Executive summary

Shaping a prosperous and equitable global future with AI depends on international cooperation, jurisdictional interoperability and inclusive governance.

The global landscape for artificial intelligence (AI) governance is complex and rapidly evolving, given the speed and breadth of technological advancements, as well as social, economic and political influences. This paper examines various national governance responses to AI around the world and identifies two areas of comparison:

1. **Governance approach:** AI governance may be focused on risk, rules, principles or outcomes; and whether or not a national AI strategy has been outlined.

2. **Regulatory instruments:** AI governance may be based on existing regulations and authorities or on the development of new regulatory instruments.

Lending to the complexity of AI governance, the arrival of generative AI raises several governance debates, two of which are highlighted in this paper:

1. **How to prioritize addressing current harms and potential risks of AI.**

2. **How governance should consider AI technologies on a spectrum of open-to-closed access.**

International cooperation is critical for preventing a fracturing of the global AI governance environment into non-interoperable spheres with prohibitive complexity and compliance costs. Promoting international cooperation and jurisdictional interoperability requires:

- **International coordination:** To ensure legitimacy for governance approaches, a multistakeholder approach is needed that embraces perspectives from government, civil society, academia, industry and impacted communities and is grounded in collaborative assessments of the socioeconomic impacts of AI.

- **Compatible standards:** To prevent substantial divergence in standards, relevant national bodies should increase compatibility efforts and collaborate with international standardization programmes. For international standards to be widely adopted, they must reflect global participation and representation.

- **Flexible regulatory mechanisms:** To keep pace with AI's fast-evolving capabilities, investment in innovation and governance frameworks should be agile and adaptable.

Equitable access and inclusion of the Global South in all stages of AI development, deployment and governance is critical for innovation and for realizing the technology's socioeconomic benefits and mitigating harms globally.

- **Access to AI**: Access to AI innovations can empower jurisdictions to make progress on economic growth and development goals. Genuine access relies on overcoming structural inequalities that lead to power imbalances for the Global South, including in infrastructure, data, talent and governance.

- **Inclusion in AI**: To adequately address unique regional concerns and prevent a relegation of developing economies to mere endpoints in the AI value chain, there must be a reimagining of roles that ensure Global South actors can engage in AI innovation and governance.

The findings of this briefing paper are intended to inform actions by the different actors involved in AI governance and regulation. These findings will also serve as a basis for future work of the World Economic Forum and its AI Governance Alliance that will raise critical considerations for resilient governance and regulation, including international cooperation, interoperability, access and inclusion.

Introduction

Generative AI promises economic growth and social benefits but also poses challenges.

The rapid onset of generative artificial intelligence (AI) is promising socially and economically,[1] including the potential to raise global gross domestic product (GDP) by 7% over a 10-year period.[2] At the same time, a range of complex challenges has emerged, such as the impact on employment, education and the environment, as well as the potential amplification of online harms.[3] Additionally, there are increased demands for corporate transparency of AI systems[4] and for clarity on data provenance and ownership.[5] Governance authorities worldwide face the daunting task of developing policies that harness the benefits of AI while establishing guardrails to mitigate its risks. Additionally, they are attempting to reconcile AI governance approaches with existing legal structures such as privacy and data protection, human rights, including rights of the child, intellectual property and online safety.

1 Global developments in AI governance

The nascent and fragmented global AI governance landscape is further complicated by challenges posed by generative AI.

The complex and fast-evolving AI governance landscape is marked by diverse national responses: risk-based, rules-based, principles-based and outcomes-based, as delineated in Table 1. It is important to note the difficulty of neatly attributing singular approaches to individual jurisdictions, as elements of multiple approaches can complement each other and are likely to be incorporated into hybrid responses.[6]

TABLE 1 **Summary of AI governance approaches (not mutually exclusive)**

	Risk-based	Rules-based	Principles-based	Outcomes-based
Definition	Focuses on classifying and prioritizing risks in relation to the potential harm AI systems could cause	Lays out detailed and specific rules, standards and/or requirements for AI systems	Sets out fundamental principles or guidelines for AI systems, leaving the interpretation and exact details of implementation to organizations	Focuses on achieving measurable AI-related outcomes without defining specific processes or actions that must be followed for compliance
Benefits	– Tailored to application area – Proportional to risk profile – Flexible to changing risk levels	– Potential reduction of complexity – Consistent enforcement possible	– Intended to foster innovation – Adaptable to new developments – Can encourage sharing of best practices	– Can support efficiency – Flexible to change – Intended to foster innovation – Compliance can be cost-effective
Challenges	– Risk assessments can be complex – May create barriers to market entry in high-risk areas – Assessment and enforcement can be complex	– Rigidity can increase compliance costs – May be unreliable to enforce	– Potential inconsistencies with interpretation of principles – Unpredictable compliance and impractical enforcement – Potential for abuse by bad actors	– Scope of measurable outcomes can be vague – Potential for diffused accountability – Limited control over process and transparency
Example	**EU:** *Artificial Intelligence Act,* 2023 (provisional agreement)	**China:** *Interim Measures for the Management of Generative AI Services*, 2023	**Canada:** *Voluntary Code of Conduct for Artificial Intelligence*, 2023	**Japan:** *Governance Guidelines for Implementation of AI Principles Ver. 1.1*, 2022

The recent provisional agreement reached on the EU AI Act represents the world's first attempt at enacting comprehensive and binding AI regulation applicable to AI products and services within a risk-based and use case-driven structure.[7] Other AI-specific regulatory efforts are also under development in various jurisdictions, such as in Canada,[8] Brazil,[9] Chile[10] and the Philippines.[11] Meanwhile, the Indian government has weighed a non-regulatory approach, emphasizing the need to innovate, promote and adapt to the rapid advancement of AI technologies.[12] In direct response to the rapid progress and widespread use of generative AI foundation models, China enacted regulations related to the use of generative AI. The EU AI Act also incorporates specific obligations for foundation models underpinning general-purpose AI (GPAI) systems.[13]

Additional countries such as Singapore,[14] Malaysia,[15] Saudi Arabia,[16] Japan,[17] and Rwanda [18] are responding to the transformative potential of AI by developing national polices[19] that outline governance intentions and explore a range of regulatory instruments, ranging from hard laws and mandatory compliance rules to soft guidance and voluntary best practices. Lending to the intricacy of the governance landscape, regulatory responses are spread across a matrix of sector-specific considerations and cross-sectorial requirements. The recently issued US Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence directs federal agencies to develop new standards and includes sector-specific guidance driven by risk management.

In addition to government regulatory efforts, there is a growing awareness of the importance of industry-responsible AI governance practices[20] in safeguarding societal interests. For example, in response to the US Executive Order the National Institute of Standards and Technology (NIST) has established the AI Safety Consortium, which intends to collaborate closely with industry, among other stakeholders, to inform risk management best practices.[21]

1.1 Evolving AI governance tensions

The existence of a spectrum of AI governance approaches considers debates arising from new and amplified challenges[22] introduced by the scale, power and design of generative AI technologies. Table 2 provides a snapshot of two prominent debates taking place with a sample of divergent positions regarding the nature of risks and access to AI models. Other emerging tensions include how generative AI will impact employment,[23] its intersection with copyright protections,[24] data transparency requirements,[25] allocation of responsibility among actors within the generative AI life cycle[26] and addressing misinformation and disinformation concerns amplified by generative AI.[27]

Many of these emerging tensions have their roots in data governance issues,[28] such as privacy concerns, data protection, embedded biases,[29] identity and security challenges from the use of data to train generative AI systems, and the resultant data created by generative AI systems. There is a need to re-examine existing legal frameworks that provide legal assurance to the ownership of AI-generated digital identities.[30]

TABLE 2 | **Areas of debate in AI governance (non-exhaustive)**

Debate and context	Sample position	Policy arguments for	Policy arguments against
Policy focus on long-term existential risks[31] vs present AI harms.[32] AI poses present harms and a spectrum of potential near- to long-term risks. Diverse positions exist regarding how to identify and prioritize the harms and risks from AI as well as the timeframe over which risks should be considered.	Advanced autonomous AI systems pose an existential threat to humanity.[33]	– Without sufficient caution, humans could irreversibly lose control of autonomous AI systems.[34] – Starting with the biggest questions around existential risk supports the development of trustworthy AI and could prevent overregulation.[35]	– Existential risks are speculative and uncertain.[36] – Can redirect the flow of valuable resources from scientifically studied present harms.[37] – Misdirects regulatory attention.[38]
	Effective regulation of AI needs grounded science that investigates present harms.[39]	– In terms of urgency, there are immediate problems and emerging vulnerabilities with AI that disproportionately impact marginalized and vulnerable populations. – Contending with known harms will address long-term hypothetical risks.[40]	– Focus on known harms may lead to neglecting long-term risks not well considered by traditional policy goals.
Policy treatment of open-source vs closed-source AI.[41] Governance consideration is being given regarding where an AI technology may sit on a spectrum of open-to-closed access.[42]	Open-source AI is critical to AI adoption and mitigating current and future harms from AI systems.[43]	– Increased access to AI and democratization of its capabilities. – Spurs innovation and stimulates competition. – Enables study of risks that can reduce bias and disparate performance for marginalized populations.	– Increased access exposes AI models to greater malicious use and unintentional misuse. – Difficulties in patching vulnerabilities can leave the AI system unsecured.[44]
	Closed-source AI is necessary to protect against misuse of powerful AI technology.[45]	– Protects commercial intellectual property. – Safeguards against potentially harmful future capabilities. – Identified vulnerabilities can be fixed and safety features can be implemented.[46]	– Concentration of power and knowledge within high-resource organizations.[47] – Increased dependency on a few foundation model providers with the risk of monopoly-related consequences.

2 International cooperation and jurisdictional interoperability

International cooperation to facilitate jurisdictional interoperability is vital to ensure global cohesion and trust in AI.

International cooperation is critical to ensure societal trust in generative AI and to prevent a fracturing of the global AI governance environment into non-interoperable spheres with prohibitive complexity and compliance costs. Facilitating jurisdictional interoperability requires international coordination, compatible standards and flexible regulatory mechanisms. For example, the US has taken the initiative to enable cooperation with Europe through the US-EU Trade and Technology Council, while Chile, New Zealand and Singapore have signed a Digital Economy Partnership Agreement. Indicative of a growing consensus on the need for AI regulation, delegate nations at the 2023 UK AI Safety Summit signed the Bletchley Declaration with a commitment to establish a shared understanding of AI opportunities and risks.

2.1 International coordination and collaboration

To ensure enduring legitimacy for governance proposals, global regulatory interoperability must adopt a multistakeholder approach that embraces a diversity of perspectives from government, civil society, academia, industry and impacted communities. Effective grounding of efforts in a comprehensive assessment of the socioeconomic impacts of AI and the efficacy of regulatory responses demands collaboration in identifying and prioritizing critical issues. Examples of international coordination efforts in drafting AI policy guidance include UNICEF's 2021 Policy guidance on AI for children and INTERPOL's 2023 Toolkit for Responsible AI Innovation in Law Enforcement developed in collaboration with the United Nations Interregional Crime and Justice Research Institute (UNICRI).

Efforts like the Organisation for Economic Co-operation and Development's OECD.AI to map interoperability gaps between national governance frameworks[48] are crucial to reducing conflicting regulatory requirements and establishing predictability and clarity for companies and people. At the intergovernmental level, coordination efforts to address international AI governance matters are currently under way at the Council of Europe's Committee on AI, OECD's Working Party on Artificial Intelligence Governance, the African Union High-Level Panel on Emerging Technologies (APET), the Association of Southeast Asian Nations (ASEAN) workshops[49] and the Guide on AI Governance and Ethics,[50] the G7[51] and the G20, among others.[52] In May 2023, G7 leaders published a report on the Hiroshima Process on Generative AI to study the rapidly evolving technology and help guide discussions on common policy priorities related to generative AI.[53] Additionally, international efforts like the United Nations High-Level Advisory Body on AI and the World Economic Forum's AI Governance Alliance are playing a critical role in coordinating multistakeholder dialogue and knowledge sharing to inform governance interoperability conversations.

2.2 Compatible AI standards

Creating the capacity and space for broader participation in the AI standards-making process is needed.

Governing bodies around the world are turning to standards as a method for governing AI. The British Standards Institution launched an AI Standards Hub aimed at helping AI organizations in the UK understand, develop and benefit from international AI standards. The European Telecommunications Standards Institute (ETSI) and the European Committee for Electrotechnical Standardization (CENELEC) have published the European Standardization agenda that includes the adoption of external international standards already available or under development, in part stimulated by the proposed EU AI Regulation's framework for standards. In the US, NIST has developed an AI Risk Management Framework to support technical standards for trustworthy AI.[54]

Despite criticisms regarding the instrumentalization of standards to shift regulatory powers from governments to private actors,[55] they are increasingly recognized as an important tool in international trade, investment, competitive advantage and national values. There is concern that substantial divergences in approaches to setting AI standards threaten a further fragmentation of the international AI governance landscape, lending to downstream social, economic and political implications internationally.

International standardization programmes are being developed by the Joint Technical Committee of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC1/SC42)[56] as well as by the Institute of Electrical and Electronic Engineers Standards Association (IEEE SA). For their part, the US, EU and China, have signalled commitments to undertake best efforts to align with internationally recognized standardization efforts.[57] Despite these signals, there is no guarantee that every country will follow these standards, especially if there is concern that their development has not been inclusive of local interests. Creating the capacity and space for broader participation in the standards-making process is thus needed.

2.3 Flexible regulatory mechanisms

The fast-evolving capabilities of generative AI require investment in innovation and governance frameworks that are agile and adaptable. This includes ongoing assessment of opportunity and risk emanating from applied practice and feedback from those directly impacted by the technology. Flexible regulatory mechanisms, beyond statutory instruments, are needed to account for societal implications and regulatory challenges that will emerge as generative AI technologies continue to advance and be adopted across various cultures and sectors. For example, Singapore,[58] the United Arab Emirates,[59] Brazil,[60] the UK,[61] the EU,[62] and Mauritius[63] have pioneered "regulatory sandboxes" that allow organizations to test AI in a safe and controlled environment. Such policy innovations must be coupled with additional efforts to clarify regulatory intent and the associated requirements for compliance. For flexible mechanisms to scale, supervisory authorities will need to consider how they provide industry participants confidence to participate and help establish agile best practice approaches while addressing the fear of regulatory capture through participation.

3 Enabling equitable access and inclusive global AI governance

The Global South's role in AI development and governance is critical to shaping a responsible future.

The need for diversity and more equitably deployed generative AI systems is of significant global concern. Inclusive governance that consults with diverse stakeholders, including from developing countries, can help surface challenges, priorities and opportunities to make generative AI technologies work better for everyone[64] and address widening inequalities associated with the pre-existing digital divide. By ensuring the inclusion of underrepresented countries from Sub-Saharan Africa, the Caribbean and Latin America, the South Pacific, as well as some from Central and South Asia (collectively referred to as the Global South) in international discussions on AI governance, a more diverse and equitable deployment of generative AI systems and compatibility of governance regimes can be achieved.

3.1 Structural limitations and power imbalances

The Global South's priorities in areas such as healthcare, education or food security often force trade-offs, hampering investments in long-term digital infrastructure. However, access to AI innovations can empower countries to make progress on economic growth and development goals[65] where needs are greatest – transforming health services, improving education quality, increasing agricultural productivity, etc. to improve lives.[66] Successfully deploying generative AI solutions at scale relies on overcoming several structural inequalities lending to power imbalances as detailed in Table 3.

TABLE 3 | **Sources of global disparities and exclusion in generative AI (non-exhaustive)**

Dimension	Context	Governance considerations
Infrastructure Access to compute, cloud providers and energy resources	Training generative AI systems, supporting experimentation and solution development and maintaining physical data centres[67] requires extensive compute and cloud infrastructure that is financially and environmentally costly[68] and results in high energy intensity.[69]	The level of computing infrastructure required for research and development of generative AI models is primarily accessible to just a few industry laboratories with sufficient funding.[70] This puts at risk the participation of the vast majority in the development of these advanced models.
Data Low resource languages and representation	Generative AI's outputs inherently reflect the data and design of a model's training. Current major generative AI models are primarily developed in the US and China and trained on data from North America, Europe and China.	Active inclusion of developing nations and diverse voices in generative AI development and governance is critical to ensure global inclusion in a future influenced by generative AI.
Talent Access to education and technical expertise	Students from the Global South often do not have access to the education and mentorship required to develop emerging technologies, such as generative AI. This can contribute to a lack of global representation among generative AI researchers and engineers, with potential downstream effects of unintended algorithmic biases and discrimination in generative AI products.	Local access to high-quality education and generative AI expertise is key to creating a sustainable talent pipeline and widening the locations where generative AI research is done. Further, more researchers and engineers from the Global South will lead to more diversity in generative AI ideas, enhanced innovation and increased opportunities for local experts to build and wield generative AI with local issues in mind.
Governance Institutional capacity and policy development	Economically disadvantaged countries often lack the financial, political and technical resources needed to develop effective AI governance policies, and regulators within these jurisdictions remain severely underfunded. According to a 2023 study of 193 countries, 114 countries, almost exclusively from the Global South, lack any national AI strategy.[71]	Disparity in AI governance capabilities can reinforce existing power imbalances and hinder global participation in the benefits of generative AI. The absence of governance policies for data and AI can lead to privacy violations, potential misuse of AI and a missed opportunity to harness AI for positive socioeconomic development, among others. Further, underfunded regulatory institutions may be ill-equipped to address the ethical, legal and social implications of AI.

3.2 Inclusion of the Global South in AI governance

In addition to equitable access, inclusion of the Global South in all stages of the development and governance of AI is essential to prevent a reinforced power imbalance whereby developing economies are relegated to mere endpoints in the global generative AI value chain, either as extractive digital workers or as consumers of the technology. Though AI policy and governance frameworks are predominantly being developed in China, the EU and North America (46%), compared to 5.7% in Latin America and 2.4% in Africa,[72] it is important to recognize the significant activities of different national bodies such as Colombia,[73] Brazil,[74] Mauritius,[75] Rwanda,[76] Sierra Leone,[77] Viet Nam[78] and Indonesia,[79] the recently introduced Digital Forum of Small States (FOSS) chaired by Singapore, as well as the emergence of AI research and industry ecosystems out of the Global South.

The absence of historical and geopolitical contexts of power and exploitation from dominant AI governance debates underscores the necessity for diverse voices and multistakeholder perspectives. The significant differences between some concerns of the Global South and those elevated within more dominant discourses of AI risks[80] warrant a restructuring of AI governance processes, moving beyond current frameworks of inclusion.[81] To adequately address regional concerns there must be a reimagining of roles that ensure Global South actors can engage in co-governance.

Conclusion

The global governance landscape for AI is complex, fragmented and rapidly evolving, with new and amplified challenges presented by the advent of generative AI. To effectively harness the global opportunities of generative AI and address its associated risks, there is a critical need for international cooperation and jurisdictional interoperability. Coordinated multistakeholder efforts, including government, civil society, academia, industry and impacted communities, are essential.

As humans drive the development of this technology and policy, responses must be developed to increase equity and inclusion in the development of AI, including with the countries of the Global South. It is up to stakeholders to take concrete action on access and inclusion. The World Economic Forum and its AI Governance Alliance are committed to driving this change, using its unique platform as a catalyst to convene diverse voices from around the world and urge them to act on vital issues, promote shared learnings and advance novel solutions.

Contributors

This paper is a combined effort based on numerous interviews, discussions, workshops and research. The opinions expressed herein do not necessarily reflect the views of the individuals or organizations involved in the project or listed below. Sincere thanks are extended to those who contributed their insights via interviews and workshops, as well as those not captured below.

World Economic Forum

Benjamin Larsen
Lead, Artificial Intelligence and Machine Learning

Cathy Li
Head of AI, Data and Metaverse; Deputy Head, Centre for the Fourth Industrial Revolution; Member of the Executive Committee

Karla Yee Amezaga
Lead, Data Policy and AI

AI Governance Alliance Project Fellows

Arnab Chakraborty
Senior Managing Director, Global Responsible AI Lead, Accenture

Rafi Lazerson
GenAI Policy Manager, Accenture

Valerie Morignat
Global Responsible AI Lead for Life Sciences, Accenture

Manal Siddiqui
Responsible AI Manager, Accenture

Ali Shah
Global Principal Director for Responsible AI, Accenture

Kathryn White
Global Principal Director for Innovation Incubation, Accenture

Acknowledgements

Sincere appreciation is extended to the following working group members, who spent numerous hours providing critical input and feedback to the drafts. Their diverse insights are fundamental to the success of this work.

Lovisa Afzelius
Chief Executive Officer, Apriori Bio

Hassan Al-Darbesti
Adviser to the Minister and Director, International Cooperation Department, Ministry of Information and Communication Technology (ICT) of Qatar

Uthman Ali
Senior Product Analyst, AI Ethics SME, BP

Erich David Andersen
General Counsel; Head, Corporate Affairs, TikTok

Jason Anderson
General Counsel, Vice-President and Corporate Secretary, DataStax

Norberto Andrade
Professor and Academic Director, IE University

Richard Benjamins
Chief AI and Data Strategist, Telefonica

Saqr Binghalib
Executive Director, Artificial Intelligence, Digital Economy and Remote Work Applications Office, United Arab Emirates

Anu Bradford
Professor of Law, Columbia Law School

Michal Brand-Gold
Vice-President General Counsel, Activefence

Adrian Brown
Executive Director, Center for Public Impact

Winter Casey
Senior Director, SAP

Simon Chesterman
Senior Director of AI Governance, AI Singapore, National University of Singapore

Melinda Claybaugh
Director, Privacy Policy, Meta Platforms

Amanda Craig
Senior Director, Responsible AI Public Policy, Microsoft

Renée Cummings
Data Science Professor and Data Activist in Residence, University of Virginia

Nicholas Dirks
President and Chief Executive Officer, The New York Academy of Sciences

Nita Farahany
Robinson O. Everett Professor of Law and Philosophy; Director, Duke Science and Society, Duke University

Max Fenkell
Vice-President, Government Relations, Scale AI

Kay Firth-Butterfield
Senior Research Fellow, University of Texas at Austin

Katharina Frey
Deputy Head, Digitalisation Division, Federal Department of Foreign Affairs, Federal Department of Foreign Affairs (FDFA) of Switzerland

Alice Friend
Head, Artificial Intelligence and Emerging Tech Policy, Google

Tony Gaffney
Chief Executive Officer, Vector Institute

Eugenio Garcia
Deputy Consul-General, San Francisco, Ministry of Foreign Affairs of Brazil

Urs Gasser
Dean, TUM School of Social Sciences and Technology, Technical University of Munich

Avi Gesser
Partner, Debevoise & Plimpton

Debjani Ghosh
President, National Association of Software and Services Companies (NASSCOM)

Danielle Gilliam-Moore
Director, Global Public Policy, Salesforce

Brian Green
Director, Technology Ethics, Santa Clara University

Samuel Gregory
Executive Director, WITNESS

Koiti Hasida
Director, Artificial Intelligence in Society Research Group, RIKEN Center for Advanced Intelligence Project, RIKEN

Dan Hendrycks
Executive Director, Center for AI Safety

Benjamin Hughes
Senior Vice-President, Artificial Intelligence (AI) & Real World Data (RWD), IQVIA

Dan Jermyn
Chief Decision Scientist, Commonwealth Bank of Australia

Jeff Jianfeng Cao
Senior Research, Tencent Research Institute

Sam Kaplan
Assistant General Counsel, Public Policy & Government Affairs, Palo Alto Networks

Kathryn King
General Manager, Technology & Strategy, Office of the eSafety Commissioner, Australia

Edward S. Knight
Executive Vice-Chairman, Nasdaq

Andrew JP Levy
Chief Corporate and Government Affairs Officer, Accenture

Caroline Louveaux
Chief Privacy and Data Responsibility Officer, Mastercard

Shawn Maher
Global Vice-Chair, Public Policy, EY

Gevorg Mantashyan
First Deputy Minister of High-Tech Industry, Ministry of High-Tech Industry of Armenia

Gary Marcus
Chief Executive Officer, Center for Advancement of Trustworthy AI

Gregg Melinson
Senior Vice-President, Corporate Affairs, Hewlett Packard Enterprise

Nicolas Miailhe
Founder and President, The Future Society (TFS)

Robert Middlehurst
Senior Vice-President, Regulatory Affairs, e& International

Casey Mock
Chief Policy and Public Affairs Officer, Center for Humane Technology

Chandler Morse
Vice-President, Corporate Affairs, Workday

Miho Naganuma
Senior Executive Professional, Digital Trust Business Strategy Department, NEC

Dan Nechita
Head, Cabinet, MEP Dragoș Tudorache, European Parliament

Michael Nunes
Head, Government Advisory, Visa

Bo Viktor Nylund
Director, UNICEF Innocenti Global Office of Research and Foresight, United Nations Children's Fund (UNICEF)

Madan Oberoi
Executive Director, Technology and Innovation, International Criminal Police Organization (INTERPOL)

Michael Ortiz
Senior Director, Policy, Sequoia Capital Operations

Florian Ostmann
Head, AI Governance and Regulatory Innovation, The Alan Turing Institute

Marc-Etienne Ouimette
Lead, Global AI Policy, Amazon Web Services

Timothy Persons
Principal, Digital Assurance and Transparency of US Trust Solutions, PwC

Tiffany Pham
Founder and Chief Executive Officer, Mogul

Valerie Pisano
President and Chief Executive Officer, MILA, Quebec Artificial Intelligence Institute

Oreste Pollicino
Professor, Constitutional Law, Bocconi University

Catherine Quinlan
Vice-President, AI Ethics, IBM

Martin Rauchbauer
Co-Director and Founder, Tech Diplomacy Network

Alexandra Reeve Givens
Chief Executive Officer, Center for Democracy and Technology

Philip Reiner
Chief Executive Officer, Institute for Security and Technology

Andrea Renda
Senior Research Fellow, Centre for European Policy Studies (CEPS)

Sam Rizzo
Head, Global Policy Development, Zoom Video Communications

John Roese
Global Chief Technology Officer, Dell Technologies

Arianna Rufini
ICT Adviser to the Minister, Ministry of Enterprises and Made in Italy

Crystal Rugege
Managing Director, Centre for the Fourth Industrial Revolution, Rwanda

Nayat Sanchez-Pi
Chief Executive Officer, INRIA Chile

Thomas Schneider
Ambassador, Director of International Affairs, Swiss Federal Office of Communications, Federal Department of the Environment, Transport, Energy and Communications (DETEC)

Robyn Scott
Co-Founder and Chief Executive Officer, Apolitical

Var Shankar
Director, Policy, Responsible Artificial Intelligence Institute

Navrina Singh
Founder and Chief Executive Officer, Credo AI

Irina Soeffky
Director, National, European and International Digital Policy, Federal Ministry for Digital and Transport of Germany

Uyi Stewart
Chief Data and Technology Officer, data.org

Chizuru Suga
Director, Digital Economy, Ministry of Economy, Trade and Industry of Japan

Arun Sundararajan
Harold Price Professor, Entrepreneurship and Technology, Stern School of Business, New York University

Nabiha Syed
Chief Executive Officer, The Markup

Patricia Thaine
Co-Founder and Chief Executive Officer, Private AI

V Valluvan Veloo
Director, Manufacturing Industry, Science and Technology Division, Ministry of Economy, Malaysia

Rishi Varma
Senior Vice-President and General Counsel, Hewlett Packard Enterprise

Ott Velsberg
Government Chief Data Officer, Ministry of Economic Affairs and Information Technology of Estonia

Miriam Vogel
President and Chief Executive Officer, Equal AI

Arif Zeynalov
Transformation Chief Information Officer, Ministry of Economy of the Republic of Azerbaijan

World Economic Forum

John Bradley
Lead, Metaverse Initiative

Karyn Gorman
Communications Lead, Metaverse Initiative

Devendra Jain
Lead, Artificial Intelligence, Quantum Technologies

Jenny Joung
Specialist, Artificial Intelligence and Machine Learning

Daegan Kingery
Early Careers Programme, AI Governance Alliance

Connie Kuang
Lead, Generative AI and Metaverse Value Creation

Hannah Rosenfeld
Specialist, Artificial Intelligence and Machine Learning

Supheakmungkol Sarin
Head, Data and Artificial Intelligence Ecosystems

Stephanie Teeuwen
Specialist, Data and AI

Hesham Zafar
Lead, Digital Trust

Accenture

Patrick Connolly
Research Manager

Charlie Moskowitz
Senior Manager, Government Relations

Anna Schilling
Data & AI – Strategy Manager

Sekhar Tewari
Associate Research Manager

Dikshita Venkatesh
Research Senior Analyst, Responsible AI

Japan External Trade Organization

Genta Ando
Executive Director; Project Fellow, World Economic Forum

Production

Laurence Denmark
Creative Director, Studio Miko

Sophie Ebbage
Designer, Studio Miko

Martha Howlett
Editor, Studio Miko

Endnotes

1. World Economic Forum, Unlocking value from Generative AI: Guidance for responsible transformation, 2024.
2. "Generative AI could raise global GDP by 7%", Goldman Sachs, 05 April 2023, https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html.
3. World Economic Forum, Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms, 2023, https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf.
4. Schaake, Marietje, "There can be no AI regulation without corporate transparency", Financial Times, 31 October 2023 https://cyber.fsi.stanford.edu/publication/there-can-be-no-ai-regulation-without-corporate-transparency.
5. Appel, Gil, Juliana Neelbauer and David A. Schweidel, "Generative AI Has an Intellectual Property Problem", Harvard Business Review, 7 April 2023, https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.
6. These approaches can be complementary. For example, a jurisdiction may decide to govern predictable risks with a risk-based approach, while leaving unpredictable risks governed by an outcomes-based approach.
7. Council of the EU and the European Council, Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world [Press release], 9 December 2023, https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/.
8. "The Artificial Intelligence and Data Act (AIDA) – Companion document", Government of Canada, 2023, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.
9. "Committee of jurists approves text with rules for artificial intelligence", Senado Noticias, 1 December 2022, https://www12.senado.leg.br/noticias/materias/2022/12/01/comissao-de-juristas-aprova-texto-com-regras-para-inteligencia-artificial.
10. "Legal Alert: Chile takes first steps towards regulation of Artificial Intelligence", DLA PIPER, 15 June 2023, https://www.dlapiper.cl/en/2023/06/15/legal-alert-chile-takes-first-steps-towards-regulation-of-artificial-intelligence/.
11. Republic of the Phillipines, House Bill 7396, 1 March 2023, https://hrep-website.s3.ap-southeast-1.amazonaws.com/legisdocs/basic_19/HB07396.pdf.
12. Liu, Shoashan, "India's AI Regulation Dilemma", The Diplomat, 27 October 2023, https://thediplomat.com/2023/10/indias-ai-regulation-dilemma/.
13. European Parliament, Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI [Press release], 9 December 2023, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-actdeal-on-comprehensive-rules-for-trustworthy-ai.
14. Government of Singapore, AI for the Public Good For Singapore and the World, 2023, https://file.go.gov.sg/nais2023.pdf.
15. Malaysia Ministry of Science, Technology & Innovation (MOSTI), Malaysia National Artificial Intelligence Roadmap 2021-2025, August 2022 https://airmap.my/wp-content/uploads/2022/08/AIR-Map-Playbook-final-s.pdf
16. National Strategy for Data & AI (NSDAI), Kingdom of Saudi Arabia, Realizing our Best Tomorrow, 2020, https://ai.sa/Brochure_NSDAI_Summit%20version_EN.pdf.
17. Cabinet Office, Government of Japan, AI Strategy 2022, 2022, https://www8.cao.go.jp/cstp/ai/aistrategy2022_gaiyo.pdf.
18. Republic of Rwanda Ministry of ICT and Innovation, The National AI Policy, 2022, https://rura.rw/fileadmin/Documents/ICT/Laws/Rwanda_national_Artificial_intelligence_Policy.pdf.
19. For a live repository of over 1,000 AI policy initiatives see: OECD.AI Policy Observatory, National AI policies & strategies [Infographic and live repository], https://oecd.ai/en/dashboards/overview/policy.
20. World Economic Forum, Unlocking Value from Generative AI: Guidance for Responsible Transformation, 2024.
21. "NIST Seeks Collaborators for Consortium Supporting Artificial Intelligence Safety", National Institute of Standards and Technology (NIST), 2 November 2023, https://www.nist.gov/news-events/news/2023/11/nist-seeks-collaborators-consortium-supporting-artificial-intelligence.
22. World Economic Forum, Data Equity: Foundational Concepts for Generative AI, 2023, pp. 10, https://www3.weforum.org/docs/WEF_Data_Equity_Concepts_Generative_AI_2023.pdf.
23. World Economic Forum, Jobs of Tomorrow: Large Language Models and Jobs, 2023, https://www.weforum.org/publications/jobs-of-tomorrow-large-language-models-and-jobs/.
24. Henderson, Peter, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, et al., "Foundation Models and Copyright Questions", Stanford University Human-Centered Artificial Intelligence, November 2023, https://hai.stanford.edu/policy-brief-foundation-models-and-copyright-questions; see also: D'Auria, Giuseppina and Arun Sundararajan, "Rethinking Intellectual Property Law in an Era of Genarative AI", TechREG Chronicle, November 2023, pp. 3-11, https://www.pymnts.com/cpi_posts/rethinking-intellectual-property-law-in-an-era-of-generative-ai/.
25. Workday, Workday Position on Foundation Models and Generative AI for the EU AI Act's Trilogue Negotiations, 2023.
26. World Economic Forum, Presidio AI Framework: Towards Safe Generative AI Models, 2024.

27. Leibowicz, Claire, "Why watermarking AI-generated content won't guarantee trust online", MIT Technology Review, 9 August 2023, https://www.technologyreview.com/2023/08/09/1077516/watermarking-ai-trust-online/.
28. For in-depth analysis on data equity and generative AI see: World Economic Forum, Data Equity: Foundational Concepts for Generative AI, 2023, https://www.weforum.org/publications/data-equity-foundational-concepts-for-generative-ai/.
29. Talat, Zeerak, Aurélie Névéol, Stella Biderman, Miruna Clinciu, et al., "You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings", in Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, eds. Angela Fan, Suzana Ilic, Thomas Wolf and Matthias Gallé, Association for Computational Linguistics, 2022, pp. 26-41.
30. Treat, David and Marie Wallace, "3 urgent questions to ask as we navigate a new digital identity", World Economic Forum, 28 September 2023, https://www.weforum.org/agenda/2023/09/3-urgent-questions-digital-identity/.
31. Hendrycks, Dan, Mantas Mazeika and Thomas Woodside, "An overview of catastrophic ai risks", arXiv, 9 October 2023, https://arxiv.org/pdf/2306.12001.pdf.
32. For a live crowd-sourced repository of AI-related harms see: Artificial Intelligence Incident Database (AIID) [live repository], AI Incident Database, https://incidentdatabase.ai/.
33. Center for AI Safety, Statement on AI Risk, 2023, https://www.safe.ai/statement-on-ai-risk#open-letter.
34. Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, et al., "Managing AI Risks in an Era of Rapid Progress", arXiv, 2023, https://arxiv.org/pdf/2310.17688.pdf.
35. Frank, Michael, "Managing Existential Risk from AI without Undercutting Innovation", Center for Strategic and International Studies (CSIS), 10 July 2023, https://www.csis.org/analysis/managing-existential-risk-ai-without-undercutting-innovation.
36. Thornhill, John, "AI will never threaten humans, says top Meta scientist", Financial Times, 18 October 2023, https://www.ft.com/content/30fa44a1-7623-499f-93b0-81e26e22f2a6.
37. Buolamwini, Joy, "Chapter 12", Unmasking AI, Penguin Random House, 2023.
38. Gebru, Timnit, Emily M. Bender, Angelina McMillan-Major and Margaret Mitchell, "Statement from the listed authors of Stochastic Parrots on the "AI pause" letter", DAIR Institute, 31 March 2023, https://www.dair-institute.org/blog/letter-statement-March2023/.
39. Hanna, Alex and Emily M. Bender, "AI Causes Real Harm. Let's Focus on That over the End-of-Humanity Hype", Scientific American, 12 August 2023, https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/.
40. Buolamwini, Joy, "No One is Immune to AI Harms with Dr. Joy Buolamwini", Your Undivided Attention [podcast transcript], episode 77, 26 October 2023, https://assets-global.website-files.com/5f0e1294f002b1bb26e1f304/653fdb3dda89d000e063ab75_77-your-undivided-attention-dr.-joy-buolamwini-transcript-corrected-title.docx.pdf.
41. Ge, Ling, "Achieving Balance in Generative AI: Open-Source Versus Proprietary Models", Tencent, 19 October 2023, https://www.tencent.com/en-us/articles/2201720.html; Sutskever, Ilya, "Open-Source vs. Closed-Source AI", Stanford eCorner, 26 April 2023, https://ecorner.stanford.edu/clips/open-source-vs-closed-source-ai/.
42. World Economic Forum, Presidio AI Framework: Towards Safe Generative AI Models, 2024.
43. Mozilla, Joint Statement on AI Safety and Openness, 31 October 2023, https://open.mozilla.org/letter/.
44. Harris, David Evan, "How to Regulate Unsecured "Open-Source" AI: No Exemption", Tech Policy Press, 3 December 2023, https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.
45. Sutskever, Ilya, "Open-Source vs. Closed-Source AI", Stanford eCorner, 26 April 2023, https://ecorner.stanford.edu/clips/open-source-vs-closed-source-ai/.
46. Harris, David Evan, "How to Regulate Unsecured "Open-Source" AI: No Exemption", Tech Policy Press, 3 December 2023, https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.
47. Sutskever, Ilya, "Open-Source vs. Closed-Source AI", Stanford eCorner, 26 April 2023, https://ecorner.stanford.edu/clips/open-source-vs-closed-source-ai/.
48. "OECD Artifical Intelligence Papers", OECD Library, n.d., https://www.oecd-ilibrary.org/science-and-technology/common-guideposts-to-promote-interoperability-in-ai-risk-management_ba602d18-en.
49. "ASEAN initiates regional discussion on generative AI Policy", Association of Southeast Asian Nations, 7 December 2023, https://asean.org/asean-initiates-regional-discussion-on-generative-ai-policy/.
50. Ministry of Communications and Information, The 3rd ASEAN Digital Ministers Meeting and Related Meetings at the Philippines [Press release], 9 February 2023, https://www.mci.gov.sg/media-centre/press-releases/the-3rd-asean-digital-ministers-meeting-at-the-philippines/.
51. "G7 Leaders' Statement on the Hiroshima AI Process", Eurpoean Commission, 30 October 2023, https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process.
52. For more examples of international collaboration, see: Oxford Insights, 2023 Government AI Readiness Index, 2023, pp. 9-10, https://oxfordinsights.com/wp-content/uploads/2023/12/2023-Government-AI-Readiness-Index-1.pdf.
53. Organisation for Economic Co-operation and Development (OECD), G7 Hiroshima Process on Generative Artificial Intelligence (AI), 2023, https://read.oecd-ilibrary.org/science-and-technology/g7-hiroshima-process-on-generative-artificial-intelligence-ai_bf3c0c60-en#page1.

54. "AI Risk Management Framework", National Institute of Standards and Technology (NIST), n.d., https://www.nist.gov/itl/ai-risk-management-framework.

55. "Standardisation Strategy Consultation - Feedback From ETUC", European Trade Union Confederation (ETUC), 28 July 2021, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13099-Standardisation-strategy/F2663296_en.

56. "ISO/IEC JTC 1/SC 42", International Organization for Standardization (ISO), 2017, https://www.iso.org/committee/6794475.html.

57. **EU**: The EU AI Act will also rely on compliance with harmonized standards aligned with international standardization efforts as a means to demonstrate conformity with its requirements. **US**: The long-standing Circular No. A-119 on federal development and use of voluntary consensus standards and conformity assessment outlines a commitment to using international standards whenever possible. **China**: 2021 National Standardization Development Outline reiterates Beijing's investment in AI standards and conformity assessment, laying out standards for AI development and deployment, and aligning these standards with international ones.

58. Infocomm Media Development Authority (IMDA), First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA [Press release], 31 October 2023, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox.

59. United Arab Emirates Government, "Regulatory Sandboxes in the UAE", n.d., https://u.ae/en/about-the-uae/digital-uae/regulatory-framework/regulatory-sandboxes-in-the-uae.

60. "ANPD's Call for Contributions to the regulatory sandbox for artificial intelligence and data protection in Brazil is now open", Autoridade Nacional de Proteção de Dados, 3 October 2023, https://www.gov.br/anpd/pt-br/assuntos/noticias/anpds-call-for-contributions-to-the-regulatory-sandbox-for-artificial-intelligence-and-data-protection-in-brazil-is-now-open.

61. "Regulatory Sandbox", Information Commissioner's Office, n.d., https://ico.org.uk/sandbox.

62. European Parliament, Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI [Press release], 9 December 2023, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

63. Ministry of Public Service, Administrative and Institutional Reforms, Sandbox Framework for Adoption of Innovative Technologies in the Public Service, 2021, https://civilservice.govmu.org/Documents/Circulars%202021/Booklet%20Sandbox%20framework.pdf.

64. Brookings Institution, "Why the Global South has a stake in dialogues on AI governance", YouTube, 23 October 2023. https://www.youtube.com/live/SbVW6lj786w?si=3uxgxonjDWwygsxj.

65. Okolo, Chinasa T., "AI in the Global South: Opportunities and challenges towards more inclusive governance", Brookings, 1 November 2023, https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/.

66. African Union High-Level Panel on Emerging Technologies, AI for Africa: Artificial Intelligence for Africa's Socio-Economic Development, 2021, https://www.nepad.org/publication/ai-africa-artificial-intelligence-africas-socio-economic-development.

67. Li, Pengfei, Jianyi Yang, Mohammad A. Islam and Shaolei Ren, "Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models", arXiv, 29 October 2023, https://arxiv.org/pdf/2304.03271.pdf.

68. OECD, AI language models: Technological, socio-economic and policy considerations, 2023, https://doi.org/10.1787/13d38f92-en.

69. Ludvigsen, Kasper Groes Albin, "The Carbon Footprint of ChatGPT", Towards Data Science, 21 December 2022, https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d?gi=e2bf91b0f208.

70. Li, Fei-Fei, Governing AI Through Acquisition and Procurement, 14 September 2023, Testimony presented to the U.S. Senate Committee on Homeland Security and Governmental Affairs, Washington DC. https://hai.stanford.edu/sites/default/files/2023-09/Fei-Fei-Li-Senate-Testimony.pdf.

71. Oxford Insights, 2023 Government AI Readiness Index, 2023, https://oxfordinsights.com/wp-content/uploads/2023/12/2023-Government-AI-Readiness-Index-1.pdf.

72. OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, https://oecd.ai.

73. Department of National Planning, Republic of Colombia, Política Nacional para la Transformación Digital e Inteligencia Artificial, 2019, https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf.

74. Shimoda Uechi, Cristina Akemi and Thiago Guimarães Moraes, "Brazil's path to responsible AI", OECD, 27 July 2023, https://oecd.ai/en/wonk/brazils-path-to-responsible-ai.

75. Mauritius Working Group on AI, Mauritius Artificial Intelligence Strategy, 2018, https://ncb.govmu.org/ncb/strategicplans/MauritiusAIStrategy2018.pdf.

76. Republic of Rwanda Ministry of ICT and Innovation, The National AI Policy, 2022, https://www.minict.gov.rw/index.php?eID=dumpFile&t=f&f=67550&token=6195a53203e197efa47592f40ff4aaf24579640e.

77. Sierra Leone Directorate of Science Technology & Innovation, Sierra Leone National Innovation & Digital Strategy, 2019, https://www.dsti.gov.sl/wp-content/uploads/2019/11/Sierra-Leone-National-Innovation-and-Digital-Strategy.pdf.

78. Government of the Socialist Republic of Viet Nam, National Strategy on R&D and Application of Artificial Intelligence, 2021, https://wp.oecd.ai/app/uploads/2021/12/Vietnam_National_Strategy_on_RD_and_Application_of_AI_2021-2030.pdf.

79. "AI Towards Indonesia's Vision 2045", Indonesia Center for Artificial Intelligence Innovation, n.d., https://ai-innovation.id/strategi.

80. Thomson Reuters Foundation, AI Governance for Africa, Part 1 and 2, 2023, https://www.trust.org/dA/97390870db/pdfReport/AI%20Governance%20for%20Africa%20Toolkit%20-%20Part%201%20and%202.pdf.

81. Chatham House, Reflections on Building More Inclusive Global Governance, 2021,https://www.chathamhouse.org/sites/default/files/2021-04/2021-04-15-reflections-building-inclusive-global-governance.pdf.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org