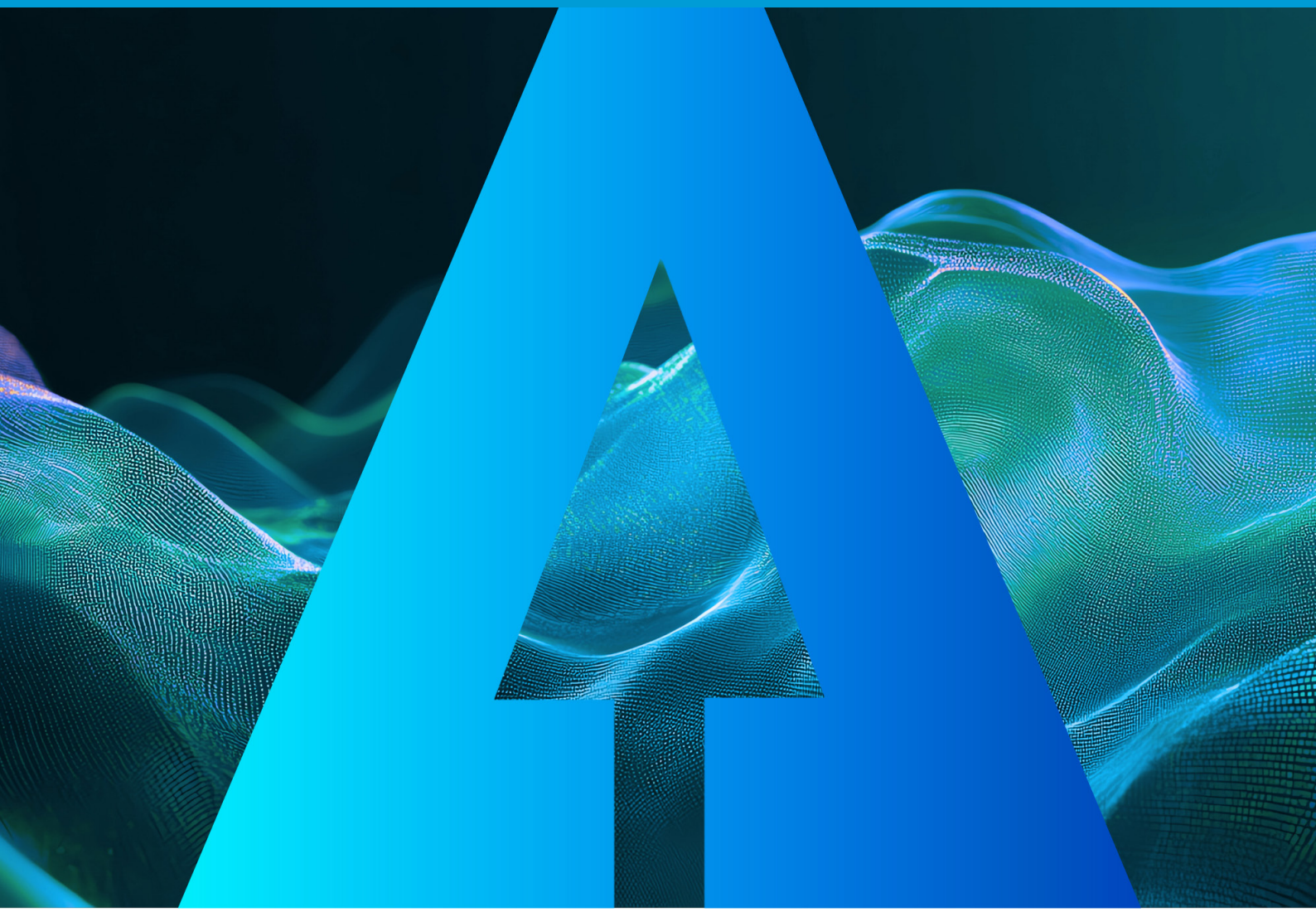
ITUPublications

International Telecommunication Union
Telecommunication Standardization Sector

The Annual AI Governance Report 2025: Steering the Future of AI

2025 Report



In partnership with:

AI GOVERNANCE INITIATIVE

OXFORD MARTIN SCHOOL

UNIVERSITY OF OXFORD

ITU

Disclaimer

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of ITU concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The mention of specific companies or certain manufacturer products does not imply that they are endorsed or recommended by ITU in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by ITU to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader.

The opinions, findings and conclusions expressed in this publication do not necessarily reflect the views of ITU or its membership.

ISBN

978-92-61-41441-2 (Electronic version)
978-92-61-41451-1 (Paper version)

The Annual AI Governance Report 2025: Steering the Future of AI

2025 Report

In partnership with:

Foreword by Doreen Bogdan-Martin, ITU Secretary-General

As the global AI race continues to accelerate, humanity stands at a unique, transformative moment.

Our collective challenge is not whether to govern artificial intelligence, but to understand and ensure governance steers AI in the right direction.

This is at the heart of ITU's mission to offer a neutral, global platform for artificial intelligence where everyone has a voice and seat at the table.

Our second annual **AI Governance Dialogue** provided a timely opportunity for exactly this kind of multi-stakeholder discussion among governments, the private sector, academia, civil society organizations, the technical community, and United Nations colleagues – each of whom has a key role to play.

We were especially privileged this year to have welcomed **H.E. Alar Karis, President of Estonia**, one of the most technologically advanced countries in the world, and to have received a spiritual message from **His Holiness Pope Leo XIV**.

We also heard from leading AI thinkers, including **Yoshua Bengio** and **Daniela Rus,** on how to maximize AI's benefits while minimizing its risks**.** I encourage you to read these and other outcomes of our discussions throughout this report.

On behalf of ITU, I thank both co-chairs of the AI Governance Dialogue: His Excellency Engineer **Majed Al Mesmar**, Director-General of the Telecommunications and Digital Government Regulatory Authority of the United Arab Emirates, and Madame **Anne Bouverot**, France's Special Envoy for Artificial Intelligence.

The "Ten Pillars for AI Governance" (see Part 2, Chapter 4.2) captures the co-chairs' vision for AI while reflecting the commitment of diverse stakeholders to translate dialogue into action for an AI that can deliver benefits for everyone, everywhere.

ITU is also grateful to Professor **Robert Trager** and his team at the University of Oxford for their advice and support in curating this year's Dialogue. Their white paper **"Themes and Trends in AI Governance**" offers an excellent glimpse into the current state of play when it comes to AI governance approaches around the world – you will find it as Part 1 of this report.

Finally, I extend my gratitude to the **over 11,000 people** from **169 countries** who attended this year's AI for Good Global Summit and World Summit on the Information Society (WSIS+20) High-Level Event. The value of such gatherings lies not only in the ideas shared on stage, but in the connections formed and the collaborations strengthened.

In **August 2025**, the United Nations General Assembly **approved the resolution** "Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance". I am also thrilled that the Resolution further specified that the Dialogue will initially

be held back-to-back in the margins of ITU's AI for Good Global Summit in Geneva in 2026. We gladly offer this report, and in particular the conclusions of the co-chairs of this year's AI Governance Dialogue, as input to Global Dialogue on AI Governance next year.

Dialogue helps us set direction and chart a path towards a better inclusive digital future.

We do not need to sail in the same ship, or at the same speed, but we do need to navigate the same oceans by the same compass.

As one Minister said during the Dialogue: we must look forward, together.

I hope this report serves in helping humanity navigate towards a horizon where AI becomes a force for shared, positive change for all.

Doreen Bogdan-Martin
ITU Secretary-General

Figure 1: Doreen Bogdan-Martin, Secretary-General, International Telecommunication Union (ITU)

Figure 2: Challenging discussions

Table of contents

Part I – White Paper: Themes and Trends in AI Governance[1]

Executive Summary

An area of emerging focus in AI governance over the past year have been **AI Agents**. The rise of AI agents – systems that can autonomously perform multi-step tasks, interact with software environments, and make decisions with minimal human input – has introduced new governance concerns, including traceability, agent coordination, and security vulnerabilities. As these agents begin to be integrated into consumer tools and enterprise workflows, questions of oversight and liability have become more urgent.

AI is also transforming science and innovation itself. AlphaFold - recognized with a Nobel Prize in 2024 – continues to evolve and demonstrates the transformative role of AI in accelerating protein structure discovery and fundamental research. Recent iterations of AlphaFold and other emerging "AI scientist" systems are driving progress in molecular design, materials science, and bioengineering. Meanwhile, autonomous experimentation platforms are beginning to integrate robotics with language models to iteratively design, test, and refine scientific hypotheses, offering an early glimpse into new research workflows shaped by AI-driven planning and reasoning.

On the socioeconomic front, **AI deployment is reshaping labor markets** not just through automation, but by reorganizing the value of human work. While early concerns focused on displacement in gig sectors – such as translation, design, and transcription – recent studies suggest a more complex picture. More recent research shows that generative AI is augmenting mid-skill roles, particularly in customer service and professional writing, leading to increased productivity and wage compression rather than pure substitution. However, the displacement effect remains concentrated in routine, low-mobility roles, with freelancers on digital platforms reporting declining demand in certain categories. In parallel, large employers are experimenting with hybrid workflows that pair AI systems with human oversight, suggesting a shift from full substitution to "recomposition" of labor. Public sentiment increasingly supports automation that removes repetitive or low-value tasks, but resists its use in sensitive or relational domains (e.g., education, healthcare). To manage this transition, governments are piloting new forms of digital public infrastructure—from India's AI Skilling Stack to targeted worker transition funds in the EU—aimed at embedding human-centered design and retraining pathways into the future of work.

Over the past year, **AI infrastructure, with compute resources being a critical pillar,** has become a central arena for strategic competition, industrial policy, and digital sovereignty. According to an analysis of researchers at the University of Oxford, the United States, China and the European Union account for over half of the world's most powerful data centers. American and Chinese companies operate more than 90 percent of the data centers used globally by other organizations for AI work. Africa and South America have almost no computing hubs. India has at least five; Japan at least four. More than 150 countries have none at all.

[1] The white paper was sent to the participants of the AI for Good Dialogue luncheon (10 July 2025) prior to the luncheon.

2025 also saw increased attention to **frontier model development** outside the traditional hubs. The release of DeepSeek-V2, China's highest-performing open-weight model, marks a significant milestone in the global development of frontier-scale AI systems. DeepSeek's capabilities prompted reflection on the minimum levels of compute, engineering expertise, and energy resources needed to build frontier systems. At the same time, models like *Mistral* and *Falcon* reinforced that open-source leadership is emerging across a range of geographies.

AI systems consume energy across two main stages: (a) **Training**, which requires high energy input at a single location over a limited time. (b) **Inference**, which involves ongoing energy use as models are deployed and used by millions daily. Current typical chatbot interactions have a negligible per-use energy footprint (~0.2–0.34 Wh per query): assuming 100 interactions with a chatbot (corresponding to 10 chats with 10 back-and-forth messages) every single day of a year, corresponds to roughly 10 kWh for the annual chatbot use, or the equivalent of a 10 km car drive, or 5 hot showers of 5 minutes, or 2 hot baths. However, future scenarios – where billions of users each employ multiple AI agents – could dramatically increase global energy demand, potentially reaching on the order of TWh/day or more. This would represent a significant share of the world's total electricity consumption (~80 TWh/day, which itself is about 20% of the world's total energy consumption). The energy infrastructure for AI is rapidly expanding: Globally, AI data centers may need 68 GW by 2027 and 327 GW by 2030, comparable to major U.S. states like California.

The global **AI standards landscape** is shaped by international Standards Development Organizations – ITU, ISO, IEC – and practitioner-led bodies like IEEE. Together, they are building a layered system of standards, spanning technical, managerial, and socio-technical domains. The **AI and Multimedia Authenticity Standards (AMAS)** initiative, launched under the World Standards Cooperation (the partnership of ITU, ISO and IEC), is developing tools and guidance to support transparency, accountability, and human rights in AI. Its initial deliverables – mapping the current landscape and identifying standardization gaps – are being published at the AI for Good Summit 2025, with more to follow. A new **AI Standards Exchange Database** will be introduced at the same summit, consolidating standards from ITU, ISO, IEC, IEEE, and IETF. ITU, ISO, and IEC also launched the **International AI Standards Summit** series in 2024, with the second edition planned for Seoul in December 2025, aligning with the goals of the UN Global Digital Compact. Regionally, CEN-CENELEC's JTC 21 in Europe is advancing sector-specific AI standards, notably in healthcare and mobility, supporting implementation of the EU AI Act and promoting cross-border interoperability through trustworthiness metrics and regulatory alignment. Nevertheless, adoption of formal AI standards remains the exception rather than the rule as firms are faced with a patchwork of faster – but unofficial – industry frameworks, resulting in inconsistency and forum-shopping risks.

The **Global AI Summit Series** (UK 2023, Seoul 2024, France 2025, India 2026) has evolved from ad hoc gatherings to structured international cooperation. Nevertheless, **AI governance** efforts remain fragmented and politically uneven. While a number of countries have launched AI safety institutes and initiated risk assessment frameworks, global coordination remains limited to a handful of Track 2 dialogues – informal, non-government channels who meet to explore ideas and issues that their governments (the "Track 1" diplomats) may be unable or unwilling to discuss in public – and regional initiatives. Track 2 diplomacy between the US and China has grown more substantive. The *Ditchley Statement* (2023) and the *Beijing Dialogue* (2024) reflect expert consensus around compute thresholds, model registration, and dual-use risk evaluation.

At the AI Action Summit in Paris (2025), Countries from the Global Majority including India, along regional bodies such as the African Union, played an active role in shaping the Summit's outcomes. Still, only a small number of countries dominate agenda-setting, while any countries – over 100 – have no meaningful voice in AI governance.

Risk assessment has become a focus of AI governance, with growing efforts to institutionalize evaluation practices and build shared safety infrastructure. AI Safety Institutes and equivalent bodies are tasked with model testing, red teaming, and developing national safety protocols. Alongside these institutions, multilateral bodies have begun coordinating on baseline risk assessment methodologies, including proposals to standardize thresholds for system classification, misuse potential, and post-deployment monitoring. However, approaches to risk classification remain uneven, and few tools currently address real-world incidents at scale. Governance of open-weight models remains contested, with proposals ranging from licensing regimes to tiered release based on misuse potential.

Verification – the ability to confirm or validate another party's actions or claims – is playing an increasingly important role in AI governance by enabling trust, accountability, and enforcement. For frontier AI, this includes validating training details, safety measures, system behavior, usage during inference, and compute resources. By improving transparency, verification helps build confidence and supports the development of shared global norms – much as it has in arms control and climate agreements – turning voluntary principles into credible, coordinated action.

Introduction

The artificial intelligence landscape has experienced rapid technological advancement throughout 2024 and early 2025. These developments include advanced capabilities in autonomous agents, multimodel AI, robotics and expanded international coordination mechanisms, significant economic deployment across sectors, and evolving governance frameworks.

This white paper analyzes seven key themes emerging from recent AI developments: autonomous agent deployment, verification systems, socioeconomic transformation, international coordination, technical standards, infrastructure requirements, and risk management. The analysis provides stakeholders with evidence-based insights into current AI trajectories and their implications for policy and implementation decisions.

Theme 1: The Year of AI Agents

Since the end of 2024, the field of AI has grown to encompass not only chat-style tools, but also so-called AI agents. Whereas earlier systems waited for a prompt, the new wave couples large-language-model reasoning with tool-use scaffolds, letting software plan, decide and act across multi-step workflows – booking trips, debugging code, even negotiating purchases – with minimal human supervision. The rapid improvement in capabilities witnessed in what many are now referring to as 'the year of AI agents' raises new questions about governance and safe deployment.

1.1 Rapid Capability Improvements

Current AI agent capabilities. Since late-2024 leading labs have begun shipping *production-ready* AI agents that can operate a full desktop or browser on the user's behalf. AI agents have evolved from simple task executors to complex systems capable of autonomous decision-making and multi-step reasoning, with minimal human oversight. OpenAI's Operator preview books holidays, fills out forms and completes online purchases end-to-end, while Salesforce's AgentForce platform promises "a billion enterprise agents" by 2026.[2] These systems couple large-language-model "brains" with tool-use scaffolding, letting them read screens, click buttons, and call APIs—marking a qualitative jump from passive copilots to autonomous digital workers.

Performance benchmarks. Progress has been impressive, but it's still patchy. On a benchmark using real GitHub bugs, the best agent increased its success rate from fixing roughly one-third of the problems to just over half. However, whenever the bug is difficult enough that a human coder would require an hour or more to resolve it, the agent almost never finds the solution.[3] The same pattern repeats outside coding. Household robot tests show agents completing only about one job in seven, whereas people succeed almost every time. In broader autonomy suites, where each task runs for an hour or longer, agents succeed fewer than one in five times. In a tough web navigation test, they finish only about one mission in seven. The upside is that the longest task an agent can complete is increasing rapidly–roughly doubling every seven months–so their capabilities are improving, even if their performance is still inconsistent.[4]

Policy considerations. This steep capability curve gives policymakers a narrow window to act. Researchers have shown that it is possible to boost scores on AI leaderboards just by using more computing power or trying lots of options, rather than actually making the AI smarter. This can hide the true costs and weaknesses of the systems; today's benchmarks rarely track factors such as energy use, bias or labour displacement. Forward-looking governance structures are needed that pair technical scores with *societal* metrics – safety, fairness, ecological footprint and economic impact – while requiring cost-controlled, reproducible evaluations. Forecasts suggest human-level performance on key autonomy suites could arrive before 2027, so establishing guard-rails now is prudent.[5]

---

2 Kraprayoon, J. (2025, April 17). *AI Agent Governance: A Field Guide*. Institute for AI Policy and Strategy.
3 Kraprayoon, J. (2025, April 17), Fn. 1
4 Kraprayoon, J. (2025, April 17), Fn. 1
5 Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Manyika, J. (2024, April 24). *The ethics of advanced AI assistants*. arXiv.org.

1.2 Agent Governance Frameworks

New governance structures are emerging to ensure AI agents operate safely and within regulatory boundaries.

Governance gap. Large-scale deployment of autonomous agents could unlock productivity gains but also introduce systemic risks—from labour disruption to cascading security failures. Yet the field of *agent governance* is still 'in its infancy,' with only a handful of researchers working on interventions while investment in building agents accelerates. This mismatch leaves both governments and industry poorly prepared to steer what could quickly become billions of autonomous digital actors. Information about AI agent deployment, for instance, is not well developed.

Legal-framework adaptation. Legal scholars are repurposing principal-agent theory and common law agency, which traditionally govern human agents acting on behalf of others, to tackle problems of information asymmetry, discretionary authority, and loyalty posed by AI systems that now operate with similar delegated authority. Yet, classical legal fixes—bonuses, monitoring, and punitive sanctions—assume understandable actions and human-paced decision-making.[6] Proposals now range from visibility-based safe-harbour regimes that reduce liability for deployers who log and disclose agent activity[7] to mandates for "law-following" models, yet courts still struggle with foreseeability and deterrence when an opaque agent acts contrary to its designer's intent at machine speed.[8]

Governance principles. Three pillars of governance are important to consider when it comes to AI agents–*inclusivity, visibility, and liability*: visibility is about making what agents do and what trained them *legible*; liability allocates responsibility and redress; inclusivity gives all affected communities a meaningful say. Technically, *visibility* can be delivered through agent identifiers, real-time monitoring, and tamper-evident activity logs, giving regulators and civil society a live audit trail without freezing innovation.[9] Legally, *liability* could be calibrated by tying safe-harbour protection or reduced fines to compliance with these visibility standards, mirroring precedents in cybersecurity and health data law. Such dual infrastructures aim to give innovators room to experiment while ensuring harms are traceable and compensable. *Inclusivity* means that the same rails that make agents visible and liable must also give voice and leverage to the people and regions they will affect—including workers, civil-society groups and governments in the Global South—so that they can shape what data are logged, who can query it and when agents may operate. Scholars point to participatory data trusts[10] and 'democratising AI' oversight boards[11] as concrete mechanisms for achieving this shared control to ensure that agent governance does not become the sole domain of a few cloud hubs in the Global North but rather reflects a variety of social interests and contexts.

---

6 Kraprayoon, J. (2025, April 17). *AI Agent Governance: A Field Guide*. Institute for AI Policy and Strategy.

7 Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 958–973.

8 O'Keefe, Cullen and Ramakrishnan, Ketan and Tay, Janna and Winter, Christoph, Law-Following AI: Designing AI Agents to Obey Human Laws (May 02, 2025). 94 Fordham L. Rev. (forthcoming 2025).

9 Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 958–973.

10 Delacroix, S., & Lawrence, N. D. (2019). Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law*.

11 Delacroix, S., Pineau, J., & Montgomery, J. (2021). Democratising the digital Revolution: The role of data Governance. In *Lecture notes in computer science* (pp. 40–52).

Multi-agent security. When agents interact, new attack surfaces appear, such as secret collusion that distorts markets, error cascades that spread misinformation and self-replicating 'agent worms' like Morris II, which can infect entire application networks.[12] Surveys show that current defences, such as sandboxing (running code in a restricted environment to limit what it can access or do), cross-examination (one AI agent to test another model's responses for safety, accuracy, or signs of manipulation) and cooperative 'AutoDefense' agents (agents working together to detect and block harmful prompts), can reduce the success of jailbreaks, but they remain basic.[13] Future frameworks will require standards for isolation, authenticated communication, and incident response across distributed agent ecosystems. This will entail moving beyond the robustness of a single model to achieve system-level resilience.

1.3 Infrastructure for Agent Deployment

Critical infrastructure is being developed to support the deployment, monitoring, and control of AI agents at scale.

Agent-infrastructure framework. Think of a future "agent-net" layered on top of today's internet: shared rails that let autonomous software act while giving humans levers to supervise it. Chan *et al.* (2024) propose three core functions for this infrastructure: (i) attribution: attaching a persistent identifier and "agent card" to every action,[14] (ii) interaction shaping: real-time monitors and permission systems that can pause or roll back risky behaviour,[15] and (iii) harm remedy: tamper-evident logs that regulators or courts can inspect after an incident to trace responsibility.[16] This moves governance toward prevention—from punishing bad outcomes to designing environments that make good conduct the default.[17]

Economic Infrastructure. E-commerce was designed for human fingertips—password boxes, CAPTCHAs, and card numbers—meaning agents still struggle with the basics: proving identity, discovering services, and processing payments. Researchers discussing agentic finance suggest that, without verifiable credentials and transparent loss-allocation rules, merchants have nothing to base their trust around when it comes to code they have never encountered before. Fintechs are rushing to retrofit the infrastructure: Stripe's open-source Agent Toolkit enables an LLM to generate one-time virtual cards or initiate bank transfers with a single command.[18] Meanwhile, Visa has announced pilots that connect autonomous shopping agents directly to its global network, indicating that a fully agent-driven checkout process is now a priority for the industry.[19] Still missing, policy analysts note, are interoperable identity proofs (e.g., verifiable credentials) and liability frameworks that allocate losses when an agent misfires.[20]

---

12 Cohen, S., Bitton, R., & Nassi, B. (2024, March 5). *Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications.* arXiv.org.

13 Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. (2025). AI Agents Under Threat: A survey of key security challenges and future pathways. *ACM Computing Surveys*.

14 Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 958–973. Page 963.

15 Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., & Anderljung, M. (2024). Visibility into AI Agents. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 958–973. Page 961.

16 *ibid.*

17 *See also:* Kraprayoon, J. (2025, April 17). *AI Agent Governance: A Field Guide*. Institute for AI Policy and Strategy.

18 *Add Stripe to your agentic workflows*. (n.d.). Stripe Documentation.

19 *Visa wants to give artificial intelligence "agents" your credit card* | *AP News*. (2025, April 30). AP News.

20 Birch, D. G. (2025, May 24). *Agentic commerce does not work without agent identities*. Forbes.

Communication Protocols. Just as HTTP and TLS are essential for the web, autonomous software requires lightweight, machine-first protocols for negotiation, task handover and authentication. The Model Context Protocol (MCP)[21], supported by Microsoft and Anthropic, and Google's developing A2A specification[22], both define a USB-C-style 'handshake' that enables itinerary-planning agents to book trains, hotels and payments across dozens of providers without the need for bespoke integrations.

[21] *Introducing the Model Context Protocol* | Anthropic. (2024, November 25).
[22] Surapaneni, R., Jha, M., Vakoc, M., & Segal, T. (2025, April 9). *Announcing the Agent2Agent Protocol (A2A)*.

Theme 2: AI and Socioeconomic Impact

2.1 Labor Market Transformation

Displacement of Labor

The displacement effect of AI is particularly pronounced in roles involving routine and repetitive tasks, where automation technologies can substitute for human labor. Workers in sectors such as manufacturing, clerical work, and low-skill services are especially vulnerable, as AI systems become more capable of performing structured, rule-based tasks efficiently and at scale. This can lead to significant job losses or restructuring, particularly for workers with limited digital or adaptive skills. While some jobs may be redefined rather than eliminated, the net effect without adequate reskilling and transition support is likely to deepen labor market inequality and exacerbate economic vulnerability among lower-income groups.[23]

We have already seen the effects of AI on the labor market. Generative AI has led to a measurable decline in the number of posted tasks and earnings for freelancers on online labor platforms, particularly in categories such as writing, translation, and graphic design. Importantly, prior strong performance does not appear to shield freelancers from these impacts; in fact, top-rated workers may be disproportionately affected.[24] Similar effects are found in implementations of AI in customer service; AI assistance increases worker productivity, measured by issues resolved per hour by 15% on average, where less experienced and lower-skilled workers improve both skill and quality, while most experienced and skilled workers see small gains in speed and small declines in quality.[25]

Anthropic's analysis of 500,000 coding-related interactions revealed that 79% of Claude Code conversations involved automation, where the AI directly completed coding tasks, rather than augmentation where AI collaborates with the user to perform a coding task. This indicates a strong trend toward AI taking over routine programming work, especially in front-end development, with implications for job disruption in roles focused on building user interfaces and simple web applications.[26]

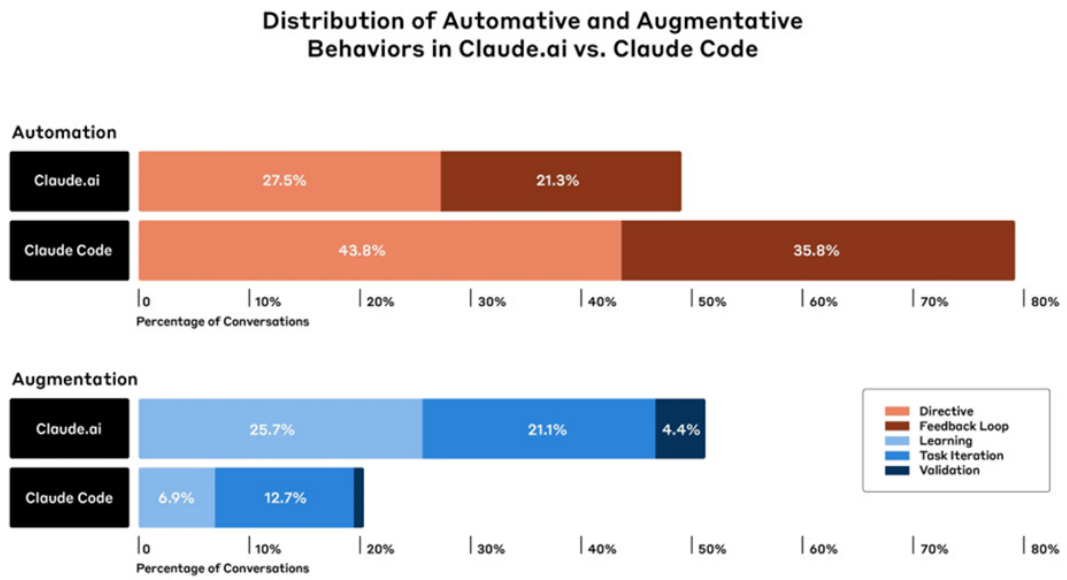
---

[23] Acemoglu, D., & Restrepo, P. (2019). Artificial Intelligence, Automation, and Work. In Ajay Agrawal, Joshua Gans, and Avi Goldfarb, editors, *The Economics of Artificial Intelligence: An Agenda* (pp. 197–236).

[24] Hui, X., Reshef, O., & Zhou, L. (2024). The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market. *Organization Science*.

[25] Brynjolfsson, E., Li, D., Raymond, L., (2025) Generative AI at Work, *The Quarterly Journal of Economics*

[26] *Anthropic Economic Index*. (2025).



Source: Anthropic[27]

Future of Jobs. There are growing mismatches in the labor market, with simultaneous shortages and surpluses across different roles and sectors. Labor shortages are expected in high-skill areas such as technology, healthcare, and green energy, driven by rising demand and insufficient supply of qualified talent. Conversely, labor surpluses are projected in roles involving routine or manual tasks, especially in manufacturing, administration, and low-skill services, as these are increasingly subject to automation and declining demand.[28] These imbalances are likely to intensify unless proactive reskilling and workforce transition strategies are implemented.

Research from the Stanford Social and Language Technologies Lab takes a worker-centric approach to understanding opportunities for human-agent collaboration. They find that for 46.1% of tasks, workers currently performing them express a "positive attitude" towards agent automation, even after considering concerns around job loss; this motivation comes from "freeing up time for high-value work," reducing "task-repetitiveness" and "stressfulness" and increasing opportunities of quality improvement.[29] This research classifies automation research into four useful categories:

1. Automation "Green Light" Zone: Tasks with both high automation desire and high capability. These are prime candidates for AI agent deployment with the potential for broad productivity and societal gains.
2. Automation "Red Light" Zone: Tasks with high capability but low desire. Deployment here warrants caution, as it may face worker resistance or pose broader negative societal implications.
3. R&D Opportunity Zone: Tasks with high desire but currently low capability. These represent promising directions for AI research.
4. Low Priority Zone: Tasks with both low desire and low capability.[30]

---

27 Anthropic, "Anthropic Economic Index: AI's Impact on Software Development," *Anthropic Research*, April 28, 2025, https://www.anthropic.com/research/impact-software-development.

28 Strack, R., Carrasco, M., Kolo, P., Nouri, N., Priddis, M., George, R., Boston Consulting Group, & Faethm. (2021). *The Future of Jobs in the Era of AI.*

29 Shao, T., Zope, H., Jiang, H., Pei, J., Nguyen, F., Brynjolfsson, E., Yang, D. (2024) Future of work with AI Agents, *Stanford University Social and Language Technologies Lab.*

30 Shao, T., Zope, H., Jiang, H., Pei, J., Nguyen, F., Brynjolfsson, E., Yang, D. (2024) Future of work with AI Agents, *Stanford University Social and Language Technologies Lab.*

2.2 Steps towards addressing the AI Divide

AI companies are making efforts to promote global inclusion by collaborating with diverse stakeholders, expanding research beyond Western-centric contexts, and supporting AI development in the Global South. These initiatives include partnering with regional universities, funding localized data collection, and promoting access to open-source tools and educational resources for underrepresented communities.[31] By doing so, companies aim to reduce bias, enhance cultural relevance, and broaden access to AI technologies worldwide. Despite these efforts, there are significant limitations. Many inclusion initiatives are still driven by Global North institutions, often sidelining local voices and reinforcing top-down dynamics. Challenges such as unequal resource distribution, language barriers, and the dominance of commercial interests often hinder meaningful engagement.

Dataset Inclusion

Efforts to improve dataset inclusion in AI development focus on expanding the geographic, linguistic, and cultural diversity of training data to reduce bias and enhance global relevance. Initiatives include sourcing data from underrepresented regions, incorporating non-English languages, and capturing context-specific information that reflects local realities. Some organizations support community-driven data collection and promote open datasets that are accessible to researchers and developers in the Global South. However, these efforts often face challenges such as limited infrastructure, uneven data governance, and ethical concerns around consent and representation. Ensuring meaningful dataset inclusion requires sustained investment, local collaboration, and safeguards that prioritize fairness and accountability.

Research Labs

AI research labs have begun expanding their global reach by establishing partnerships with institutions in the Global South, opening satellite offices, and funding regional AI hubs aimed at fostering local talent and innovation. These efforts are intended to decentralize AI development and bring more diverse perspectives into research and deployment. Google, Microsoft, and IBM have established research labs in the Global South, as well as development centers, customer support hubs, or data centers in these regions. However, the distribution of AI research facilities remains uneven. In Southeast Asia, lab representation is limited solely to India; in South America, to Brazil. Sub-Saharan Africa shows slightly more geographic diversity, with AI labs located in Accra (Ghana), Nairobi (Kenya), and Johannesburg (South Africa). Grassroots AI education and training initiatives by communities such as Deep Learning Indaba, Data Science Africa, and Khipu AI in Latin America aim to increase local AI talent. However, inclusion remains limited, as decision-making power and core research agendas often remain concentrated in the Global North. Many collaborations still operate within asymmetrical power structures, where local contributors have little influence over priorities or outcomes.

2.3 Economic Growth and Productivity Gains

Calculations on the economic growth and productivity gains of generative AI rely on two types of assumptions: task replaceability and new innovation capabilities. Experts have varying

[31] Chan, A., Okolo, C. T., Terner, Z., & Wang, A. (2021, February 2). *The Limits of Global Inclusion in AI Development*. arXiv.org.

predictions on what percentage of tasks can be reliably and profitably replaced by generative AI and at what timeline. Similarly, experts are uncertain what types of productivity gains can be achieved by new discoveries, such as "new materials, new drugs, or new services."[32] McKinsey research sizes the long-term global AI opportunity at $4.4 trillion annually, or a contributed 3.7% of global GDP growth, in added productivity from corporate use cases by enhancing productivity across industries, especially in customer operations, software engineering, marketing, and R&D.[33] In contrast, studies from MIT indicate a much more conservative growth estimate of 0.7% annually, suggesting a "nontrivial, but modest effect."[34]

New Value Streams in Education

There are emerging value streams in education by enabling more personalized, adaptive, and data-informed learning experiences. Tools such as Intelligent Tutoring Systems (ITSs), Natural Language Processing (NLP), and Automated Performance Enhancement (APE) systems offer the capacity to individualize instruction and assessment, allowing educators to respond more precisely to diverse student needs.[35] These technologies may contribute to improved educational outcomes by supporting real-time feedback and streamlining administrative tasks such as grading, thereby reallocating educators' time toward more complex pedagogical and mentoring activities. Furthermore, AI-enabled analytics and planning tools are beginning to inform curriculum development and institutional decision-making by aligning educational content with student performance patterns, local priorities, and anticipated labor market demands.

Additionally, AI is contributing to the advancement of immersive learning through technologies such as augmented and virtual reality (AR/VR), which are being explored for their potential to simulate hands-on experiences in disciplines ranging from science to the humanities. These tools may facilitate more engaging and context-rich educational environments, particularly in remote or under-resourced settings. AI also supports the development of lifelong and flexible learning pathways, potentially expanding access to upskilling and reskilling opportunities in rapidly evolving sectors such as data science, healthcare, and cybersecurity.

New Value Streams in Science

In 2024, for the first time and likely not the last, the Nobel Prize was awarded to a discovery that was enabled by AI: AlphaFold. AlphaFold was made to predict the structure of virtually all the 200 million proteins that researchers have identified, and it enabled the development of new protein structures; new iterations of AlphaFold predict the structure and interactions of all of life's molecules, which can unlock new materials, crops, drugs, and research.[36]

Generative AI can speedup the discovery of new materials and molecules using a process called generative design. For example, by training AI models on simulations from quantum physics, we can make more accurate predictions about how materials behave; these models, when paired with more traditional approaches, help scientists get a better understanding of

---

32 Walsh, D. (2025). A new look at the economics of AI. MIT Sloan School of Management.
33 Mayer, H., Yee, L., Chui, M., & Roberts, R. (2025, January 28). *Superagency in the workplace: Empowering people to unlock AI's full potential*. McKinsey & Company.
34 Acemoglu, D., (2024), The Simple Macroeconomics of AI. *MIT*
35 Escotet, M. Á. (2023). The optimistic future of Artificial Intelligence in higher education. *Prospects*.
36 Deepmind (2024), AlphaFold 3 predicts the structure and interactions of all of life's molecules.

how electrons interact or how materials respond to light.[37] Another major stream is autonomous experimentation, where robotics and large language models (LLMs) are combined to design and execute experiments dynamically. This creates a closed-loop system of AI-guided synthesis, real-time feedback, and iterative refinement, dramatically shortening the development cycle for new materials—from decades to potentially months. AI scientists are specialized scientific agents that can reason, plan experiments, interpret data, and collaborate with human researchers in iterative discovery processes.[38] Their value extends to designing modular AI tools that can be reused across disciplines, enabling scalability and adaptability in scientific workflows.[39]

New Value Streams in Transportation

Emerging value streams associated with autonomous vehicles (AVs) span across multiple sectors and are poised to reshape the transportation, logistics, urban planning, and public service landscapes. In logistics and delivery services, AVs create opportunities to reduce labour costs and enhance delivery speed, particularly in addressing the "last-mile" delivery challenge. For instance, the integration of autonomous drones and robots can improve service efficiency and reach underserved or hard-to-access locations. In the transport sector more broadly, AVs can reduce operational costs through improved fuel efficiency, predictive maintenance, and reduced accident rates—yielding potential economic benefits estimated at £51 billion annually in the UK and up to $936 billion per year in the U.S. Additionally, AVs are expected to improve personal mobility for individuals with limited access to transport—such as older adults or those with disabilities—thus generating new demand for inclusive mobility services. Fleet-based shared AVs may also disrupt traditional public transit models, offering dynamic, low-cost alternatives that could be particularly effective in first- and last-mile connections. Moreover, AVs can enable productivity gains by allowing passengers to work during travel and by reducing traffic congestion, which in turn can boost regional economic integration and increase real estate values in peripheral urban areas.[40]

New Value Streams in Agriculture

New value streams in agriculture are emerging through the combined use of analytical and generative AI, which together have the potential to unlock an estimated $250 billion in economic value globally—$100 billion on-farm ("on the acre") and $150 billion at the enterprise level. On the farm, AI supports yield optimization through virtual agronomy advisors that integrate weather, soil, and pest data, while also enabling labor efficiencies and input cost reductions via precision agriculture. At the enterprise level, generative AI is driving innovation across R&D, marketing, operations, and supply chain management—generating hypotheses for crop development, personalizing customer outreach, and automating regulatory processes.[41]

---

37 Chen, C. (2024). AI in materials science: Charting the course to Nobel-worthy breakthroughs. *Matter*, *7*(12), 4123–4125.

38 Gridach, M., Nanavati, J., Abidine, K. Z. E., Mendes, L., & Mack, C. (2025, March 12). *Agentic AI for Scientific Discovery: A survey of progress, challenges, and future directions*. arXiv.org.

39 Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024, August 12). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. arXiv.org.

40 Thomas, F. (2024, December 17). What Might be the Economic Implications of Autonomous Vehicles? *NIESR*.

41 Nuscheler, D., Fiocco, D., Prabhala, P., Perdur, R. M., Brennan, T., & Gautam, Y. (2024). *From bytes to bushels: How gen AI can shape the future of agriculture*. McKinsey & Company.

Theme 3: International AI Governance Coordination

3.1 Global AI Summits

Formalizing Global AI Summit Series: The Global AI Summit Series is a major international gathering bringing together governments, industry, and civil society to establish AI norms. The Global AI Summit Series, launched at Bletchley Park (1-2 November 2023)[42], continued in Seoul (21-22 May 2024)[43] and France AI Action Summit[44,45] (10-11 February 2025), establishing itself as the central venue for international collaboration on AI governance, producing foundational documents like the Bletchley Declaration[46] and the Seoul Declaration.[47] These summits have shifted from ad hoc gatherings to a more structured, recurring process, with recommendations to further institutionalize the series for sustained impact and clearer agenda-setting. India will host the next summit on 19-20 February 2026.

Advancing Concrete Commitments: At the AI Seoul Summit (21–22 May 2024), sixteen frontier-model developers, including Amazon, Anthropic, Google, Microsoft, and Samsung, signed the Frontier AI Safety Commitments, pledging to conduct red-team testing and implement model-weight security measures.[48] They also committed to publishing public "severe-risk" frameworks ahead of the Paris AI Action Summit in February 2025. A civil society Seoul Commitment Tracker released after the 10 February 2025 deadline found that only six firms had fully met the commitments, four were partially compliant, and six had fallen short, highlighting uneven corporate engagement.[49]

3.2 Track 2 Diplomacy Initiatives

Track 2 Diplomacy in AI Governance: Track 2 diplomacy refers to informal, non-government channels—such as academics, former government officials or civil society leaders—who meet to explore ideas and issues that their governments (the "Track 1" diplomats) may be unable or unwilling to discuss in public. A series of recurring Track 2 dialogues between US and Chinese experts—such as those held in Thailand (2024) and organized by institutions like Tsinghua University and Brookings—have become essential forums for candid, technical discussions on AI safety, military AI, and risk mitigation, especially where official channels are constrained.[50] These dialogues have addressed sensitive scenarios, using simulations and tabletop exercises to move from abstract principles to practical risk management.

---

42 UK Prime Minister's Office. (2023, June 7). *UK to host first global summit on Artificial Intelligence* [Press release]. GOV.UK. Retrieved June 25, 2025.

43 *About the AI Seoul Summit 2024*. (2024, May 7). GOV.UK.

44 Présidence de la République (France). (2025, March 13). *Sommet pour l'action sur l'intelligence artificielle* [Web page]. Présidence de la République française. Retrieved June 25, 2025.

45 *Dennis et al.* (2025). *What success looks like for the French AI Action Summit*. Centre for the Governance of AI.

46 UK Prime Minister's Office, Foreign, Commonwealth & Development Office, & Department for Science, Innovation and Technology. (2025, February 13). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023* (Policy paper). GOV.UK. Retrieved June 25, 2025.

47 Ministry of Foreign Affairs, Republic of Korea. (2024, May 23). *Seoul Declaration for safe, innovative and inclusive AI by participants attending the leaders' session of the AI Seoul Summit, 21 May 2024* [Press release]. Retrieved June 25, 2025.

48 Department for Science, Innovation and Technology. (2025, February 7). *Frontier AI Safety Commitments, AI Seoul Summit 2024*. GOV.UK.

49 *Seoul Commitment Tracker*.

50 Kahl, C., & Hass, R. (2024, April 5). *Laying the groundwork for US-China AI dialogue*. *Brookings*.

Building Common Ground: Track 2 initiatives have produced consensus statements (e.g., the Ditchley Statement,[51] Beijing Dialogue 2024[52]) that underline the need for coordinated global action to prevent unacceptable AI risks, such as misuse and loss of control, and have informed subsequent official (Track 1) negotiations.[53] Dialogues have focused on developing best practices for AI governance, including compute thresholds, model registration, and risk evaluation, with technical experts from both global north and south institutions sharing approaches.[54]

3.3 Regional AI Partnerships

Bilateral Agreements for AI Safety and Innovation: Bilateral agreements are quickly becoming the responsive layer of AI safety governance. In April 2024, the US and UK AI Safety Institutes signed the first formal partnership, pledging to carry out joint red-team exercises and share evaluation datasets for frontier models.[55] This model has since been adopted by other countries: In November 2024, Singapore signed a Memorandum of Cooperation with the UK,[56] and in June 2025, Canada followed suit, linking its new Safety Institute[57] and the firm Cohere to the same testing protocols.[58] At the strategic end of the spectrum, in November 2024, Washington and Beijing agreed that only humans, not AI, will hold nuclear launch authority, highlighting a shared interest in hard safety boundaries despite their rivalry.[59]

Interregional Cooperation: There is growing momentum from bilateral agreements to transregional initiatives. At their summit in Kuala Lumpur in May 2025, ASEAN, the Cooperation Council for the Arab States of the Gulf (GCC) and China agreed to develop a joint framework for data security cooperation and shared skills programmes.[60] In December 2024, the EU and the Smart Africa Alliance launched a Global Gateway partnership that pairs secure digital networks with an African-owned data governance flagship to promote trustworthy AI across the continent.[61] The African Union's Continental AI Strategy, endorsed in July 2024 and augmented in March 2025,[62] sets continent-wide benchmarks on data sovereignty, trusted digital infrastructure and ethical use of AI; by April 2025 AU technical assistance had enabled 22 member states to begin drafting aligned national strategies.

---

51 Russell, S., Bengio, Y., Yao, A., Felten, E., Grosse, R., Hadfield, G., Khareghani, S., Hadfield-Menell, D., Perset, K., Song, D., Zhang, Y.-Q., Chen, X., Tegmark, M., Seger, E., Zeng, Y., Zhang, H., He, Y.-H., Gleave, A., & Heide, F. (2023, October 18–20). *International Dialogue on AI Safety (Ditchley Statement)*. Ditchley Park, UK.

52 International Dialogues on AI Safety. (2024, March 10–11). *Consensus Statement on Red Lines in Artificial Intelligence*. Beijing, China.

53 Siddiqui, S., Loke, K., Clare, S., Lu, M., Richardson, A., Ibrahim, L., McGlynn, C., Ding, J. (2025). *Promising topics for US–China dialogues on AI safety and governance.*

54 Guest, O. (2025, April 11). *International AI Safety Dialogues: Benefits, Risks, and Best Practices – Institute for AI Policy and Strategy*.

55 Alder, M. (2024, April 2). *U.S., Britain announce partnership on AI safety, testing*. Reuters.

56 Ministry of Digital Development and Information, Singapore; & Prime Minister's Office, UK. (2024, November 6). *New Singapore–UK agreement to strengthen global AI safety and governance*. Government of Singapore.

57 Prime Minister's Office, UK & Prime Minister's Office, Canada. (2025, June 15). *Joint statement between the Prime Minister of the United Kingdom and the Prime Minister of Canada*. GOV.UK.

58 Cohere Labs. (2025, June 15). *Cohere partners with Canada and UK Governments on secure AI*. Cohere.

59 Alder, M. (2024, November 17). *Biden, Xi agree that humans, not AI, should control nuclear arms*. Reuters.

60 Association of Southeast Asian Nations; Cooperation Council for the Arab States of the Gulf; & People's Republic of China. (2025, May 27). *Final Joint Statement of the ASEAN-GCC-China Summit, Kuala Lumpur, Malaysia*.

61 Directorate-General for International Partnerships (European Commission). (2024, December 3). *Global Gateway: EU and Smart Africa strengthen partnership for Africa's digital transformation*. European Commission.

62 African Union, (2024, August 9) *Continental Artificial Intelligence Strategy*, African Union.

Multilateral and Cross-Regional Agreements: Regional and global forums are beginning to incorporate AI-safety provisions into wider economic and governance agreements. Within Southeast Asia, the ASEAN Digital Economy Framework Agreement (DEFA) will be the first trade pact to require trusted data flows, shared AI testbeds, and common safeguards for 'high-impact' models across all ten members,[63] intending to give the region a unified baseline for responsible deployment.[64] Across other major economies, the G7 Hiroshima AI Process has evolved into a voluntary code of conduct that is now endorsed by 49 jurisdictions. This commitment involves conducting risk-based evaluations and public reporting for frontier systems, thereby facilitating regulatory interoperability beyond the G7 itself.[65] For binding law, the Council of Europe AI Convention (open for signature since September 2024) establishes the legal duty to respect human rights, democracy, and the rule of law throughout the AI life cycle.[66] It has already been endorsed by the EU, the US, and the UK. At global level, the UN Global Digital Compact, aims to set global norms and rules for digital cooperation.[67]

In the Annex please find a list of multilateral and national initiatives as of June 2025 (the list is indicative and non-exhaustive).

63 https://www.weforum.org/stories/2025/05/asean-digital-economy-framework-agreement-a-gamechanger/

64 Kalash, S. Y. (2024, November 11). *AI and the Digital Economy Framework Agreement: Preparing ASEAN for the Next Tech Wave*. Centre for International Governance Innovation (CIGI).

65 Assiciated Press. (2024, May 2). *Japan's Kishida unveils a framework for global regulation of generative AI*. The Associated Press.

66 Council of Europe. (2024, September 5). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. *Council of Europe*.

67 United Nations. (n.d.). *Global Digital Compact* [Web page]. Office for Digital and Emerging Technologies.

Theme 4: AI Standards and Best Practices

4.1 Landscape of AI Standard Setting Initiatives

Institutional Landscape and Global Cooperation on AI Standards: The global architecture for AI standard-setting includes the international Standards Development Organizations ITU (International Telecommunication Union), ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission), and the practitioner-driven IEEE (Institute of Electrical and Electronics Engineers). Together, their combined portfolios now span telecom-centric specifications, horizontal management-system norms and socio-technical practice guides, forming an emerging layered stack of standards.

ITU commands the broadest constituency, with a membership including 194 Member States and more than 1'000 companies, universities, research institutes, and international and regional organizations[68], giving it both intergovernmental relevance and sector-specific technical depth. ISO has 48 participating members in its AI committee (JTC 1/SC 42).[69]

The AI and multimedia authenticity standards (AMAS) collaboration, under the World Standards Cooperation (the partnership of ITU, ISO and IEC), announced at the AI for Good Summit 2024, is exploring how standards can serve as practical tools to promote transparency and accountability, uphold human rights, and guide responsible innovation in AI systems.

AMAS has delivered a mapping of the standardization landscape in the area of AI and multimedia authenticity, including the identification of gaps where standards are needed. It has also developed related guidance for policymakers and regulators. These first two deliverables, being published in conjunction with the AI for Good Global Summit 2025, will be followed by future iterations.

ITU, ISO and IEC organized the first edition of an International AI Standards Summit alongside the World Telecommunication Standardization Assembly (WTSA-24) in New Delhi in October 2024 and will convene the second International AI Standards Summit in Seoul from 2 to 3 December 2025. The new summit series aims to help realize the objectives of the UN Global Digital Compact.[70]

At the AI for Good Global Summit in Geneva 2025, a new AI Standards Exchange Database was launched, which includes standards from ITU, ISO, IEC, IEEE and IETF (Internet Engineering Task Force), with other international and regional standards development organizations considering joining.

Sector-Specific and Regional Standardization Initiatives: Regional initiatives are moving from principle to implementation. In Europe, CEN-CENELEC Joint Technical Committee 21 on Artificial Intelligence (JTC 21)—established in 2021 and bringing together 300-plus experts from over 20 countries—is crafting sector-specific standards for healthcare, mobility and other high-risk domains.[71] The presumption-of-conformity routes under the EU AI Act are established by the standards they have set out, as stated in Commission Decision C(2023)3215.[72] The

---

68 ITU/UN. (2024, November 27). *Our members - ITU*.
69 *ISO/IEC JTC 1/SC 42 - Artificial intelligence*. (2023, September 11). ISO.
70 ITU. (2025, April 2). *World Telecommunication Standardization Assembly (WTSA-24)*. WTSA-24.
71 Skov, K. (2025, January 29). *About the Joint Technical Committee* | *CEN-CENELEC JTC 21* |. European AI Standardization | CEN-CENELEC JTC 21.
72 Register of Commission Documents. (2023)

committee is simultaneously mapping ISO/IEC texts to EU regulatory clauses and developing new metrics (e.g., trustworthiness indicators for autonomous vehicles) to secure cross-border interoperability within the single market.

4.2 Technical Standards Development

Types of Technical Standards: AI standards fall into four practical categories. *Management-system standards*, such as ISO/IEC 42001[73], require organisations to establish roles, risk processes, and continuous improvement loops for each stage of AI development. A similar approach is adopted at the network level by the ITU-T Y.3061 architecture for autonomous telecom networks.[74] *Assessment and assurance standards* address individual systems. ISO/IEC 42005 establishes an eight-step method for impact assessment[75], while ITU-T Y.3173 provides metrics for rating the 'intelligence' of 5G nodes and services.[76] *Conformity and audit standards* lay the foundations for third-party certification. ISO/IEC 42006 sets out the rules for the competence of auditors of AI management systems[77], while ITU-T F.781.2 specifies the test protocols and evidence thresholds for AI-driven medical software.[78] At the interface layer, technical implementation standards such as ITU-T Y.3172 define the data formats and control points that enable machine-learning functions to be integrated into 5G networks.[79] Meanwhile, *socio-technical practice guides* such as IEEE 7000 provide engineers with guidance on value elicitation and traceability, ensuring that ethical considerations are incorporated into software development.[80]

Gaps in Current Technical Standard Setting: Adoption of formal AI standards remains the exception rather than the rule as firms are faced with a patchwork of faster—but unofficial—industry frameworks, resulting in inconsistency and forum-shopping risks.[81] Participation is likewise skewed: civil-society groups, SMEs and many Global-South delegations lack the money and time to engage, while Big Tech staff attend up to 80 hours a week, giving them disproportionate influence over committee outcomes.[82] Traditional standards development procedures, which often take 18–36 months from proposal to publication, cannot keep pace with the rapid evolution of agentic AI models.[83] Such procedures also struggle to codify value-laden issues, such as fairness, across diverse legal regimes. A recent study by the Oxford Martin School thus argues that traditional Standard Development Organisations should shift their focus

---

73 *ISO/IEC 42001:2023.* (2023). ISO

74 ITU-T Recommendation database. (2023, December 14). *ITU-T Y.3061*. ITU.

75 ISO/IEC. (2025) *42005:2025.* ISO.

76 *OECD.AI*. (2024, July 2). *ITU-T Y.3173 - Framework for evaluating intelligence levels of future networks including IMT-2020*.

77 ISO (2025). *ISO/IEC 42006*.

78 ITU. (2024, June 13). *F.781.2- Quality assessment requirements for artificial intelligence/machine learning-based software as a medical device*.

79 ITU. (2019, June 22). *Y.3172: Architectural framework for machine learning in future networks including IMT-2020*.

80 IEEE Standard. (2021, September 15). *7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerns during System Design* | IEEE Xplore.

81 Huw R. and Ziosi M. (2025, June 9) *Can we standardise the frontier of AI?* Oxford Martin AI Governance Initiative.

82 Huw R. and Ziosi M. (2025, June 9) *Can we standardise the frontier of AI?* Oxford Martin AI Governance Initiative. Page 9.

83 Huw R. and Ziosi M. (2025, June 9) *Can we standardise the frontier of AI?* Oxford Martin AI Governance Initiative. Page 10.

to "trailing-edge" work—high-level management and process standards for well-established problems—rather than trying to codify every frontier risk.[84]

One practical way to address this "standardisation gap" would be to adopt ITU's pre-standardisation focus group model[85], which has been used in areas such as health and disaster management in collaboration with other UN agencies. This could involve establishing a joint ITU–ISO/IEC group to develop a reference architecture, risk taxonomy, and sandbox test framework. Meanwhile, IEEE could develop additional ethics-by-design guidance.

Standards for AI Agents: Agentic AI will require new standards that go beyond today's model-centric norms. No formal standard yet specifies agent-to-agent communication protocols, secure tool APIs, or guardrails for autonomous, self-modifying behaviour. Early prototypes, such as Microsoft's open Agent2Agent (A2A) protocol[86] and the community-led Model Context Protocol (MCP)[87], demonstrate how multi-agent workflows and context exchange could function. However, they remain ad hoc specifications outside the remit of any formal standards development organisation. Safety researchers warn that existing test methods do not capture the emergent risks of agents. The newly published Multi-Agent Emergent Behavior (MAEBE) framework is one of the first attempts to measure collective behaviours in the absence of harmonised evaluation methods.[88]

4.3 Ethical AI Frameworks

UNESCO Recommendation on the Ethics of Artificial Intelligence: On 23 November 2021, UNESCO adopted the Recommendation on the Ethics of Artificial Intelligence – the first multilateral agreement on AI ethics. It commits all 194 Member States to the principles of protecting human rights, transparency, fairness, and human oversight.[89] To turn these high-level values into practice, UNESCO created a Readiness Assessment Methodology (RAM), and according to the agency's Social Sciences directorate, it has 'worked with nearly 60 countries – largely in Africa, the Caribbean and Latin America – to run baseline diagnostics and draft action plans'.[90] While implementation efforts are still in early stages, and some concepts (such as "fairness" and "accountability") remain open to context-specific interpretation[91], the Recommendation serves as a valuable ethical framework. As metrics and regulatory pathways continue to evolve, the Recommendation provides important guidance and supports international convergence on AI ethics, especially in contexts where formal regulatory regimes are still emerging.

OECD AI Principles: Adopted by OECD ministers on 22 May 2019 and revised on 3 May 2024 to address the risks posed by generative AI, the OECD Recommendation on Artificial Intelligence

---

84 Huw R. and Ziosi M. (2025, June 9) *Can we standardise the frontier of AI?* Oxford Martin AI Governance Initiative.

85 International Telecommunication Union. (2025, May 5). *AI/ML (Pre-) Standardization - AI for Good*. AI For Good.

86 Arenas, Y., & Brekelmans, B. (2025, May 22). *Empowering multi-agent apps with the open Agent2Agent (A2A) protocol* | The Microsoft Cloud Blog.

87 Model Context Protocol. (2025, March 26). *Specification - model context protocol*.

88 Erisken, S., Gothard, T., Leitgab, M., & Potham, R. (2025, June 3). *MAEBE: Multi-Agent Emergent Behavior Framework*. arXiv.org.

89 https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

90 UNESCO. (2023). *Readiness assessment methodology. A tool of the Recommendation on the Ethics of Artificial Intelligence*.

91 AllahRakha, N. (2024). *UNESCO's AI Ethics Principles: Challenges and Opportunities*. International Journal of Law and Policy, 2(9), 24–36.

now has 47 adhering governments (38 OECD members and nine partner economies).[92] Branded the 'first intergovernmental AI standard', it sets out five values-based principles – human-centred values, transparency, robustness, accountability, and inclusive growth – and five policy pillars that provide officials with practical tools such as R&D investment and regulatory sandboxes. To help states translate these aims into practice, the OECD Policy Observatory promotes an 'hourglass' governance model that links high-level principles to organisational processes and system-level controls, emphasising stakeholder engagement and continual monitoring.[93]

Stakeholder Engagement and Inclusive Governance: Ethical-AI initiatives now embed structured multi-stakeholder consultations: the EU's draft Code of Practice for general-purpose AI, for example, drew almost 430 submissions from industry, academia and civil-society groups in late 2024, shaping both the text and its monitoring plan.[94] The Global Partnership on AI (GPAI) pairs government representatives with experts from science, business and NGOs to co-chair its working groups.[95] Under India's 2024 chairmanship, GPAI's New Delhi Declaration called for "pursuing a diverse membership, with a particular focus on low and middle-income countries to ensure a broad range of expertise, national and regional views"[96] and a side-meeting at the Global INDIAai Summit highlighted mechanisms to "overcome the global AI divide," a message applauded by Global-South delegates.[97] New memberships—such as Morocco's decision to join after the 2024 Belgrade ministerial—illustrate that these outreach efforts are beginning to translate into broader geographic representation.[98]

4.4 Safety Standards and Red-Teaming

Building a Shared Scientific and Policy Understanding: The 2025 International AI Safety Report, authored by experts from 33 countries and major intergovernmental organizations, synthesizes the latest evidence on AI risks and mitigation strategies. It serves as a scientific foundation for standards development and informed policymaking, with input from both global north and south experts.[99]

Establishing AI Safety Testing Protocols: AI Safety Testing Protocols are a set of methods and processes used to ensure that AI systems operate as intended and without causing harm or unintended consequences. The global community has prioritized the development of robust safety standards for AI systems, particularly for general-purpose and high-risk applications. The 2025 International AI Safety Report synthesizes current knowledge on AI risks and mitigation techniques, providing a scientific foundation for informed policymaking and shared international understanding. The EU AI Act, for instance, mandates strict risk assessment, mitigation systems, and robustness requirements for high-risk AI systems, including detailed documentation, traceability, and human oversight.[100]

---

92 *OECD.AI*. (2019). *AI Principles Overview*.
93 Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022, June 1). *Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance*. arXiv.org.
94 European Commission. (2024, September 24). *Industry, academia and civil society contribute to the work on Code of practice for general-purpose artificial intelligence*. Shaping Europe's Digital Future.
95 Global IndiaAI Summit. *Global Partnership on Artificial Intelligence*.
96 GPAI. (2024). *GPAI New Delhi Declaration*.
97 GPAI. (2024). *Two days' Global INDIAai Summit 2024 concludes*.
98 The North Africa Post. (2025). Morocco to join Global Partnership on AI.
99 UK Department for Science, Innovation and Technology. (2025, February 18). *International AI Safety Report 2025*.
100 European Commission (2024). *Regulation - EU - 2024/1689*.

Red-Teaming: Red-teaming is a structured process where expert teams simulate adversarial attacks and real-world threat scenarios against AI systems to identify vulnerabilities, test limits, and enhance resilience.[101] Unlike traditional testing, red-teaming adopts the perspective of potential attackers, probing for weaknesses such as harmful outputs, bias, data leaks, or system manipulation. Red-teaming is now recognized as essential for AI deployed in critical sectors—finance, healthcare, infrastructure—where failure could have severe consequences.

Adversarial Testing and Continuous Risk Assessment: Adversarial testing, a core component of red-teaming, involves designing challenging inputs—such as nonsensical prompts or attempts to bypass safety guardrails—to expose flaws in AI models. This approach uncovers issues like bias, harmful content, and security vulnerabilities before deployment, ensuring models are robust under unpredictable, real-world conditions.

Global Collaboration and Tool for Safety Evaluation: The UK's AI Security Institute has launched the Inspect evaluations platform, making advanced safety testing tools available to the international community and accelerating the adoption of consistent safety standards worldwide.[102]

4.5 Certification and Accreditation Programs

Practitioner Certification Pathways: A wide array of AI practitioner certifications now exists, targeting different expertise levels and career paths. The Certified Artificial Intelligence Practitioner (CAIP) is a cross-industry certification accredited under ISO/IEC 17024:2012, emphasizing practical skills in AI and machine learning for system design, implementation, and deployment.[103] Alongside CAIP, globally recognized certifications such as the Certified Artificial Intelligence Scientist (CAIS), ARTIBA AI Certification, and platform-specific credentials like the Microsoft Azure AI Engineer Associate and NVIDIA Jetson AI Certification support the development of both foundational and advanced AI competencies across diverse technological environments. These certifications are designed to validate not only technical proficiency but also understanding of ethical, legal, and governance aspects of AI, reflecting the growing demand for responsible practitioners.

Accreditation Programs for AI Systems in Key Sectors: New sector-specific accreditation initiatives are emerging, such as the URAC Health Care AI Accreditation (launching Q3 2025), which will provide a verifiable framework for safe, ethical, and equitable AI implementation in clinical environments.[104] In education, the Global AI Ethics in Education Charter (2025) led by UNESCO and partners, is establishing standardized codes of ethics for AI use in academia, with accreditation agencies encouraged to align local standards with this global charter.[105]

---

101 Wisbey, O. (2024, November 21). *What is AI red teaming?* Search Enterprise AI.

102 Department for Science, Innovation and Technology. (2024, May 10). *AI Safety Institute releases new AI safety evaluations platform.* *GOV.UK*.

103 *Certified Artificial Intelligence Practitioner - CAIP Training - CerTNexus.* (2025, March 14). CertNexus.

104 *URAC to launch First-Ever Health Care AI Accreditation Program in Q3 2025.* (2025, May 19). URAC.

105 International Education Accreditation Council. (2025, May 6). *Accrediting Ethical AI Integration in Higher Education: A Roadmap*. https://www.ieac.org.uk/46-Accrediting-Ethical-AI-Integration-in-Higher-Education-A-Roadmap-blog.php.

Theme 5: AI Infrastructure and Compute

5.1 Global Compute Distribution

Compute Needs: Access to computing power and storage ("compute" for short) –alongside data and skilled talent–is a crucial ingredient in building artificial intelligence systems. Scaling laws in AI, which describe how the performance of AI systems improves as the size of training data and computational resources improve, has been a dominant force in the field (although critics caution that diminishing returns may have been reached); these scaling laws have driven further development of compute resources. This access is heavily concentrated, with one or a few firms dominating critical points in the supply chain. Control over compute has become the central factor concentrating power in AI and raising barriers to entry.[106]

Compute Distribution: According to an analysis of researchers at the University of Oxford[107], only 32 countries host data centers that provide the compute power essential for advanced AI development. The United States, China[108], and the European Union account for over half of the world's most powerful data centers. American and Chinese companies operate more than 90 percent of the data centers used globally by other organizations for AI work. Africa and South America have almost no computing hubs. India has at least five; Japan at least four. More than 150 countries have none at all.

Implication for AI Development Equity: AI development is highly spatially uneven, with a small number of global hubs capturing most of the economic gains while other regions face growing disparities. Effective policies must ensure broader distribution of AI's benefits to prevent deepening regional inequality and economic polarization.[109]

Advances in both hardware (e.g., GPUs doubling in price-performance every ~2 years) and algorithms (e.g., image classification compute needs halving every 9 months) have drastically reduced the cost of training powerful AI models. As compute becomes more efficient, more actors can train models to a given performance level, and those with existing resources can train even more powerful models, maintaining a frontier advantage.[110] Large compute investors (like major tech firms) discover new AI capabilities first and maintain a lead, especially in high-performance. Smaller actors benefit from diffusion but face barriers due to economies of scale, proprietary technologies, and strategic integration.[111]

5.2 Energy and Sustainability

Calculations on energy use of generative AI vary on two axes of energy demand: (1) training, the amount of energy necessary to host large training runs for models, and (2) inference,

---

106 Vipra, J. (2025, April 23). *Computational power and AI*. AI Now Institute.

107 Hawkins, Zoe and Lehdonvirta, Vili and Wu, Boxi, AI Compute Sovereignty: Infrastructure Control Across Territories, Cloud Providers, and Accelerators (June 20, 2025). Available at SSRN: https://ssrn.com/abstract=5312977

108 Lewis, J. A. (2024). *An overview of global cloud competition.* Centre for Strategy and International Studies.

109 https://www.researchgate.net/profile/Uchechukwu-Ajuzieogu/publication/391430918_The_Economic_Geography_of_AI_Spatial_Distribution_of_Benefits_and_Costs/links/6817004960241d5140226eed/The-Economic-Geography-of-AI-Spatial-Distribution-of-Benefits-and-Costs.pdf

110 Pilz, K. F., Heim, L., & Brown, N. (2025). Increased compute efficiency and the diffusion of AI capabilities. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(26), 27582–27590.

111 Ahmed, N., & Wahed, M. (2020, October 22). *The de-democratization of AI: deep learning and the compute divide in artificial intelligence research*. arXiv.org.

the energy used each time someone interacts with the model, such as tying a request into a chatbot and receiving a response, i.e., the cost of using models and data centers. Additional energy considerations extend beyond AI model use to include the raw materials needed for development of chips and data centers, and water and energy needs for production. Growth in AI computation is already driving significant power demands, and we have seen mobilization to produce more energy across the AI industry, with leading labs investing significantly in nuclear power.

In the United States, AI data center power demand grew tenfold over the last three years—from 0.4 gigawatts (GW) in 2020 to 4.3 GW in 2023 (Patel, Nishball, and Ontiveros, 2024). In 2025, total AI data center demand will likely reach about 21 GW of total power capacity, more than a fourfold increase from 2023 and twice the total power capacity of the state of Utah.[112] From RAND's research, Global AI data centers may need 68 gigawatts (GW) of power by 2027 and up to 327 GW by 2030, driven by continued exponential growth in AI chip supply and training demands—comparable to the total power capacity of major U.S. states like California.

Furthermore, looking beyond direct usage. chip manufacturing relies on scarce minerals like cobalt and tungsten, and construction of AI infrastructure involves carbon-intensive materials like concrete.[113] The production of GPUs – 3.85 million units shipped to data centers in 2023 alone – also contributes indirect emissions through complex manufacturing and material extraction.[114]

Training runs require significant energy consumption, and are difficult to satisfy, because they require a large amount of power capacity at a single location. Currently, training runs represent a small share of overall energy use; however, if scaling laws persist, their energy use impacts will be more significant. Compute scaling has been consistent for over a decade, and hyperscalers like OpenAI have announced plans to continue development and grow compute. Even if they are one-off events, training runs could demand up to 1 GW in a single location by 2028 and require up to 8 GW of power by 2030, equivalent to eight nuclear reactors, assuming current scaling trends persist.[115] In order to develop the capacity for large scale single location use of power, significant challenges including inadequate transmission, insufficient power, and supply gain delays would need to be overcome.

Although training AI models requires significant energy, the greater demand arises from inference, when hundreds of millions of people interact with these chatbots daily. On current inference energy usage, OpenAI recently released a figure that the average query uses about 0.34 watt hours (Wh), "about what an oven would use in a little over one second."[116] Marcel Salathé (EPFL) estimates that, assuming that a typical chatbot interaction consumes an energy of about 0.2 Wh and that an average user has 100 interactions (10 chats, each with 10 back-and-forth messages) every single day in a year, this would add to an annual energy consumption per

---

112 Pilz, K. F., Mahmood, Y., Heim, L., & RAND Corporation. (2025). AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness. RAND Corporation.

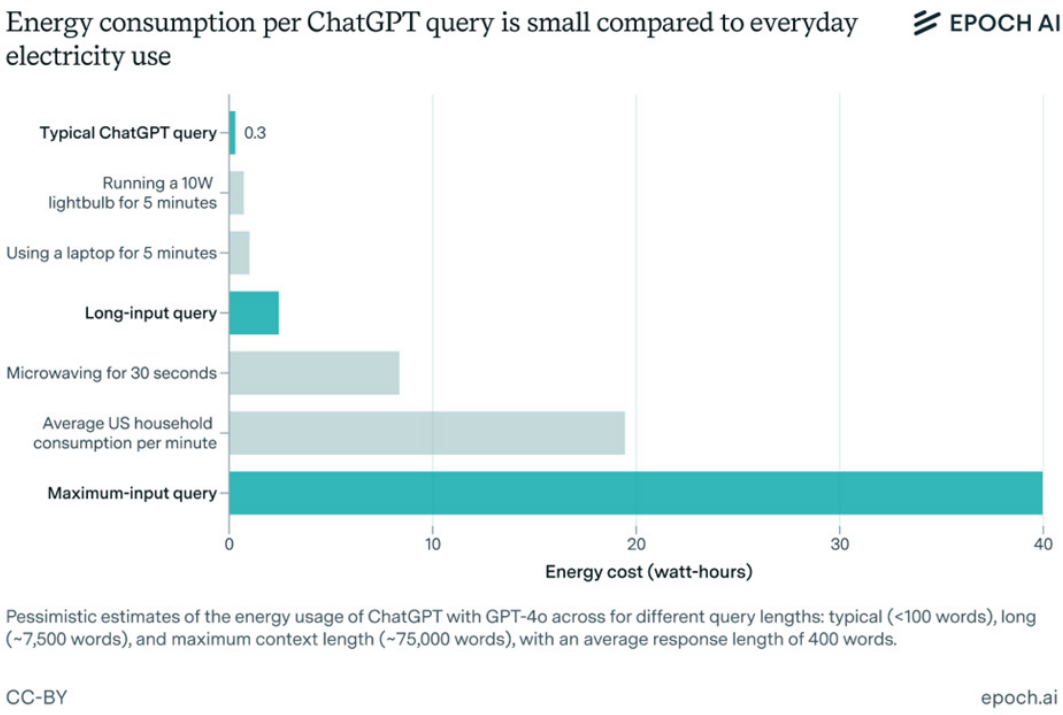
113 Luccioni, S., Trevelin, B., Mitchell, M. (2024, September 3). *The Environmental Impacts of AI -- Primer*. Hugging Face.

114 Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. (2024, March 27). *The climate and sustainability implications of Generative AI*. An MIT Exploration of Generative AI.

115 Pilz, K. F., Mahmood, Y., Heim, L., & RAND Corporation. (2025). AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness. RAND Corporation.

116 Altman, S., (2025). The Gentle Singularity.

user of around 7.2 kWh. This corresponds to a 10 km car drive, or 5 hot showers of 5 minutes, or 2 hot baths. Typical chatbot use – so far – therefore requires only negligible amounts of energy.[117]



See, for example, the estimated energy consumption per ChatGPT query compared to everyday electricity.

However, this doesn't mean we can ignore the energy implications of AI. Imagine a future where each person is assisted by hundreds of AI agents, constantly making requests to solve problems we haven't even conceived of yet. For a back-of-the-envelope calculation, let us use the estimate of 0.2 watt-hours (Wh) per chatbot request and consider a scenario where 1 billion people each use 10 AI agents, with each agent making 1,000 requests per day.[118]. That results in:

0.2 Wh × 1,000 requests/day × 10 agents × 1 billion users = 2×10^{12} Wh/day = 2 terawatt-hours (TWh)/day.

For context, the world's total electricity consumption is around 80 TWh per day, which accounts for about 20% of total global energy use.

And even this estimate may be conservative. Why stop at 10 agents per person[119]? If individuals used 100 agents, the energy demands would scale dramatically—quickly approaching levels that could rival the world's total current energy consumption.

Even as the carbon intensity of energy infrastructure decreases, it is unlikely that the suddenness of this demand is met by sustainable energy sources; even if new increases in energy can be met with sustainable sources, the growth in energy needs indicates that reduction of fossil fuel

[117] Salathé, M., (2025). AI's Energy Use
[118] Salathé, M., (2025). AI's Energy Use
[119] Salathé, M., (2025). AI's Energy Use

sources may be difficult. We may find increased short-term reliance on fossil-fuel based energy sources.[120]

Renewable Energy Transition: Artificial Intelligence could potentially support the achievement of SDG targets by enhancing renewable energy systems through optimization, forecasting, automation, and smart communication. AI could help all five targets of Affordable and Clean Energy by improving grid efficiency, reducing costs, and integrating intermittent sources like solar and wind. AI contributes to climate action by optimizing emissions reductions and energy planning. In the economic domain, it enhances productivity and innovation by supporting green jobs, efficient industry, and waste-to-energy systems.[121]

AI-enabled smart grids use machine learning and real-time data analytics to optimize the generation, distribution, and consumption of electricity, which is crucial for integrating variable renewable energy sources like wind and solar. These smart grids can forecast energy demand and RE supply with high accuracy, enabling better scheduling and load balancing. AI helps manage distributed energy resources (e.g., rooftop solar, electric vehicles, batteries) by automating decisions about when to store, release, or redirect energy. This reduces reliance on fossil fuel backup systems, cuts emissions, and enhances grid stability and resilience. Additionally, AI can detect faults, predict equipment failures, and respond to outages faster than traditional systems, lowering costs and improving reliability.[122]

Green AI is a paradigm focused on reducing the environmental impact of artificial intelligence systems while maintaining performance. Green-in-AI, which involves designing energy-efficient models and algorithms that minimize computation, and Green-by-AI, which uses AI to advance sustainability in other sectors like energy, agriculture, and climate policy.[123] Green AI emphasizes strategies such as algorithm optimization, efficient hardware (like TPUs), and edge computing to reduce emissions. Regulations like the EU's AI Act now mandate reporting energy use for high-risk AI systems; tools like CarbonTracker and CodeCarbon are used to measure emissions.[124]

5.3 Hardware Innovation and Supply Chains

AI Chip Supply Chain: AI supply chain risk is highly centralized, where the production of advanced AI chips depends heavily on a few key players. The supply chain resembles an "inverted triangle," with one global supplier of EUV lithography machines at the narrow base, and another which produces around 90% of AI chips, as the critical bottleneck. This concentration of power makes the supply chain vulnerable to geopolitical tensions.[125] [126]

---

120 McKinsey & Company (2025), How data centers and the energy sector can sate AI's hunger for power.

121 Hannan, M., Al-Shetwi, A. Q., Ker, P. J., Begum, R., Mansor, M., Rahman, S., Dong, Z., Tiong, S., Mahlia, T. I., & Muttaqi, K. (2021). Impact of renewable energy utilization and artificial intelligence in achieving sustainable development goals. *Energy Reports*, *7*, 5359–5373.

122 Ahmad, T., Madonski, R., Zhang, D., Huang, C., & Mujeeb, A. (2022). Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renewable and Sustainable Energy Reviews*, *160*, 112128.

123 Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, *599*, 128096.

124 Thakur, D., Guzzo, A., Fortino, G., & Piccialli, F. (2025). Green Federated Learning: A new era of green aware AI. *ACM Computing Surveys*.

125 Gopal, S., Staufer-Steinnocher, P., Xu, Y., & Pitts, J. (2022). Semiconductor Supply Chain: A 360-Degree view of supply chain risk and network resilience based on GIS and AI. In *Springer series in supply chain management* (pp. 303–313)

126 Mison, A., Davies, G., & Ward, R. (2024). *Cross-disciplinary AI supply chain risk assessment.* Reading: Academic Conferences International Limited.

Bifurcation and Localization of Chip Supply Chain: The semiconductor supply chain is undergoing two major transformations: bifurcation and localization. Bifurcation refers to the gradual division of the supply chain into two competing ecosystems—one led by the U.S. and its allies, and the other by China—as both powers seek to reduce mutual dependencies due to national security concerns[127]. In parallel, localization is gaining momentum, with regions like East Asia, Southeast Asia, and Europe investing in domestic semiconductor capabilities to enhance resilience and reduce external risks.[128]

Innovations in Computing Architectures: AI is making chip production faster, cheaper, and more accurate through automation and predictive analytics, and chips and computing architectures are being designed for the purpose of AI.[129] DeepSeek's example shows us how computing architectures might change to adapt to the use case of AI.[130] Furthermore, generative AI presents new possibilities for automating and enhancing different phases of chip design by streamlining tasks such as architectural exploration, circuit tuning, and layout creation.[131]

5.4 Cloud Infrastructure and Access

Public Compute: Public compute initiatives—publicly funded programs that provide access to computational resources—offer significant benefits in democratizing access to AI infrastructure, especially as AI becomes central to research, innovation, and national competitiveness. These programs help democratize access to high-performance computing, allowing researchers, startups, and public agencies to pursue AI development without relying solely on expensive private infrastructure. For example, the USA's National AI Research Resource (NAIRR) pilot[132] and the UK's AI Research Resource (AIRR)[133] aim to provide equitable access to compute for academic and scientific communities. Similarly, India's Open Compute Cloud and GPU subsidy program support domestic companies by reducing costs and offering more flexible, market-driven access to computing resources. These initiatives not only fuel innovation but also attempt to bridge the "compute divide" that leaves smaller institutions and less wealthy nations behind.[134]

Despite their promise, public compute efforts face several challenges. One is the risk of value capture, where large private tech firms disproportionately benefit from public infrastructure, potentially undermining the intended public good. Another issue is coordination: overlapping efforts within jurisdictions, such as multiple city-levels or state-levels can dilute effectiveness without cohesive policy frameworks. Additionally, long-term planning is complicated by the need for flexibility in a rapidly evolving AI landscape; the UK's cancelled £900 million exascale supercomputer project illustrates how political uncertainty can destabilize these efforts.[135]

---

127 Center for Security Studies. "The Semiconductor Race: Global Technopolitics and Supply Chain Resilience." CSS Analyses in Security Policy, no. 345, ETH Zürich, March 2024. https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/CSSAnalyse345-EN.pdf.

128 Merle, Q. (2024). *Chips Supply Chain: bifurcation and localization*. ETH Zürich.

129 Ranaweera, J. (2025). How will artificial intelligence reshape the semiconductor industry?: Artificial Intelligence meets Silicon: The next era of chip Design and manufacturing. *IEEE Electron Devices Magazine*, *3*(1), 12–14.

130 Zhao, C., Deng, C., Ruan, C., Dai, D., Gao, H., Li, J., Zhang, L., Huang, P., Zhou, S., Ma, S., Liang, W., He, Y., Wang, Y., Liu, Y., & Wei, Y. X. (2025, May 14). *Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures*.

131 Raghuweanshi, P. (2024). REVOLUTIONIZING SEMICONDUCTOR DESIGN AND MANUFACTURING WITH AI. *Journal of Knowledge Learning and Science Technology ISSN 2959-6386 (Online)*, *3*(3), 272–277.

132 *National Artificial Intelligence Research Resource Pilot*. (n.d.). NSF - National Science Foundation.

133 Shearer, E., Davies, M., Lawrence, M. (2024). *The role of public compute*. Ada Lovelace Institute Blog.

134 Davies, M., & Vipra, J. (2024). *Policy briefing Mapping global approaches to public compute Understanding the options available to policymakers*. Ada Lovelace Institute.

135 Davies, M., & Vipra, J. (2024). *Policy briefing Mapping global approaches to public compute Understanding the options available to policymakers*. Ada Lovelace Institute.

Public AI: Public AI refers to artificial intelligence systems developed or supported by public institutions to serve societal goals, rather than solely private profit. It emphasizes open source, community-centered values, and public-benefit applications. The rise of public compute—publicly funded access to high-performance computing—is a key enabler of Public AI.[136] Public AI supports transparency and accountability by encouraging open-source models, shared datasets, and democratic oversight, which helps build public trust, and it directs AI toward solving critical public challenges—such as in health, education, climate, and social services—that are often overlooked by market-driven systems.[137] Creating democratic governance structures—such as citizen assemblies, public oversight boards, and participatory design processes—ensures that AI systems reflect diverse societal needs and values.[138]

[136] Sitaraman, G., & Parek, K., (2025) The global rise of public AI. *Vanderbilt Policy Accelerator*.
[137] Public AI Network. (2024, August 10). Public AI: Infrastructure for the Common Good.
[138] Sieker, F., Tarkowski, A., Gimpel, L., & Osborne, C. (2025). Public AI White Paper – A public Alternative to private AI dominance. Bertelsmann Stiftung.

Theme 6: AI Safety and Risk Management

6.1 Risks of AI and System Safety Assessment

AI Risk Classification and Dual Use Nature: Risks stemming from AI can be categorised into three main areas: Malicious use risks, where systems are deliberately repurposed for harmful activities such as cyberattacks, disinformation campaigns or even the development of biological weapons; risks from malfunctions, which arise from unforeseen technical failures, inherent biases in training data or a lack of understanding of the system's true capabilities; and systemic risks, which encompass broader societal impacts such as market concentration, large-scale labour market disruption and the exacerbation of global inequalities.[139] A significant source of these risks is the dual-use nature of AI technology: its powerful capabilities can be harnessed for tremendous benefit or severe harm. The dominant discourse surrounding AI safety remains Western-centric, often failing to adequately incorporate the diverse linguistic traditions, cultural values and lived experiences of communities within Global Majority nations and marginalised groups.[140] This systemic exclusion risks entrenching global inequities and underscores the need for more inclusive approaches and culturally informed evaluation frameworks in AI safety development.

Best Practices in System Safety Assessment: Safety institutes and AI labs routinely conduct structured model evaluations to assess the capabilities and risks of advanced AI systems, particularly general-purpose AI and frontier AI models.[141] This includes pre-deployment risk assessments, dangerous capabilities evaluations, and benchmarking against standardised tasks.[142] Red-teaming—where experts attempt to "break" model safeguards—is increasingly used to identify vulnerabilities before deployment.[143] Safety cases, meaning structured safety arguments supported by evidence, are gaining traction as a scalable method for demonstrating that an AI system is safe within its deployment context.[144] Other best practices to assess model safety include ongoing monitoring, incident response planning, and regular third-party audits to ensure that safety measures remain effective over time and across evolving deployment scenarios.

Robustness and Reliability Testing: AI system robustness refers to its ability to maintain consistent performance, even when faced with unexpected conditions or altered input data.[145] This is particularly important for applications such as autonomous driving, where malfunctions

---

139 Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Zeng, Y. (2025, January 29). *International AI Safety Report*. arXiv.org.

140 Okolo, C. T. (2025, February 12). A new writing series: Re-envisioning AI safety through global majority perspectives. *Brookings*.

141 Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Zeng, Y. (2025, January 29). *International AI Safety Report*. arXiv.org.

142 Friedland, A. (2025, May 28). *AI Safety Evaluations: An Explainer.* Center for Security and Emerging Technology.

143 Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023, May 11). *Towards best practices in AGI safety and governance: A survey of expert opinion*. arXiv.org.

144 Buhl, M. D., Sett, G., Koessler, L., Schuett, J., & Anderljung, M. (2024, October 28). *Safety cases for frontier AI*. *arXiv.org* & Hilton, B., Davidsen Buhl, M., Korbak, T., & Irving, G. (2025). *Safety Cases: A Scalable Approach to Frontier AI Safety.* arXiv.org.

145 Wozniak, A., Duong, N. Q. K., Benderitter, I., Leroy, S., Segura, S., & Mazo, R. (2023). Robustness testing of an industrial road object detection system. *IEEE*, 82–89.

could result in substantial property damage and even fatalities.[146] Robustness testing aims to ensure that an AI system behaves correctly in the event of unexpected occurrences, whether technical or targeted. Reliability, which is often assessed through robustness, is essential for critical AI systems to prevent potentially fatal failures during real-time operations. Mathematically, robustness can be seen as a quantifiable measure of trustworthiness, indicating how well a model aligns with expected behaviour despite minor input variations.[147] Ultimately, achieving high levels of robustness is essential for building trust and confidence in AI systems used in areas where safety and security are critical, as unpredictable or opaque behaviours are unacceptable.

Limitations of Risk Assessments: Risk assessments for AI systems are significantly limited, primarily due to the 'black-box' nature of advanced models, which makes their internal functioning and all potential failure modes difficult to understand fully. For general-purpose AI, it is challenging to exhaustively evaluate all possible downstream use cases and predict how risks might manifest through complex real-world interactions rather than within the system itself.[148] Furthermore, reliable quantitative risk estimation is severely hindered by the limited historical data on AI incidents and the difficulty of assessing low-probability, high-impact events, or 'unknown unknowns'.[149] The dual-use nature of many AI capabilities also complicates these assessments, as the same features can be used for beneficial or malicious purposes, blurring the lines of potential harm.[150] Consequently, existing methodologies often cannot provide strong assurances or definitive guarantees against all associated harms.

6.2 Approaches to Mitigating AI Risks

Voluntary Commitments: Voluntary AI safety standards and frameworks are being adopted more and more to help organisations manage risks, promote the responsible use of AI, and prepare for future regulations. Industry leaders, including Google DeepMind, OpenAI and Anthropic, have developed their own commitments, such as Responsible Scaling Policies and Preparedness Frameworks, which outline risk thresholds and mitigation strategies, particularly for foundation models with dual-use potential.[151] While such voluntary measures are seen as a bridge enabling smoother transitions towards future regulatory requirements and fostering continuous improvement, sources indicate that self-regulation alone is unlikely to be sufficient to adequately manage the severe risks posed by highly capable AI models in the long term.[152] Therefore, government intervention will likely be necessary to ensure compliance with safety standards.[153]

---

146 Berghoff, C., Bielik, P., Neu, M., Tsankov, P., & Von Twickel, A. (2021). Robustness testing of AI Systems: A case study for traffic sign recognition. In *IFIP advances in information and communication technology* (pp. 256–267).

147 Braiek, H. B., & Khomh, F. (2024, April 1). *Machine Learning Robustness: a primer*. arXiv.org.

148 Mukobi, G. (2024, August 5). *Reasons to doubt the impact of AI risk evaluations*. arXiv.org.

149 Koessler, L., Schuett, J., & Anderljung, M. (2024, June 20). *Risk thresholds for frontier AI*. arXiv.org.

150 Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Wolf, K. (2023, July 6). *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. arXiv.org.

151 Karnofsky, H. (2024). *If-Then commitments for AI risk reduction*. Carnegie Endowment for International Peace.

152 Longpre, S., Klyman, K., Appel, R. E., Kapoor, S., Bommasani, R., Sahar, M., McGregor, S., Ghosh, A., Blili-Hamelin, B., Butters, N., Nelson, A., Elazari, A., Sellars, A., Ellis, C. J., Sherrets, D., Song, D., Geiger, H., Cohen, I., McIlvenny, L., Narayanan, A. (2025, March 21). *In-House evaluation is not enough: towards robust Third-Party flaw disclosure for General-Purpose AI*. arXiv.org.

153 Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Wolf, K. (2023b, July 6). *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. arXiv.org.

AI Safety Institutes: AI Safety/Security Institutes (AISIs) and their equivalents have been established around the world[154], including in the US, the UK, the EU, Japan, Singapore, Canada, France, Kenya and Australia. These institutes form an international network that aims to accelerate AI safety science and foster a common understanding of best practices.[155] These institutes coordinate research, develop model evaluation tools, and promote interoperability of safety standards, aiming to support rigorous oversight and scientific consensus on AI risks.[156]

AI Incident Reporting and Response Systems: Pre-deployment risk management alone is often insufficient, given that very dangerous models may be deployed, or deployed models may become dangerous after release.[157] An AI incident is defined as an event or series of events involving the development, use or malfunction of one or more AI systems that directly or indirectly leads to harm such as injury, disruption to critical infrastructure, violations of human rights, or damage to property, communities or the environment.[158] To address this gap, governments and standard-setting organizations are exploring mechanisms for post-deployment monitoring and response, including incident reporting. The OECD's Global AI Incident Reporting Framework, released in early 2025, is a step towards standardised, interoperable AI incident reporting worldwide.[159] The framework is designed to identify high-risk systems, inform real-time risk management and support mandatory and voluntary reporting via the AI Incidents Monitor (AIM).[160]

6.3 Corporate Risk Mitigation Practices and its Limitations

Corporate Technical Safety Research: AI companies such as Anthropic, Google DeepMind, and OpenAI primarily direct their technical safety research towards pre-deployment areas, focusing on model alignment, testing, and evaluation to ensure AI systems behave as intended and to minimise large-scale misuse or accident risks.[161] Key approaches in this research include reinforcement learning from human feedback, adversarial testing, red-teaming, and robustness analysis, all aimed at preventing unintended harmful behaviours as AI models become more capable and autonomous.[162] However, there are significant research gaps in high-risk deployment areas such as healthcare, finance, misinformation, and the handling of persuasive or addictive features, which are often less prioritised due to commercial imperatives.[163] Moreover, the concentration of safety research within a limited number of major corporations can exacerbate these oversights and restrict broader public and academic scrutiny, particularly

---

154 Araujo, R. (2025, April 10). *Understanding the first wave of AI safety Institutes: characteristics, functions, and challenges – Institute for AI Policy and Strategy*. Institute for AI Policy and Strategy.

155 Araujo, R. (2025, April 10). *Understanding the first wave of AI safety Institutes: characteristics, functions, and challenges – Institute for AI Policy and Strategy*. Institute for AI Policy and Strategy.

156 Allen, G. C., & Adamson, G. (2024). *The AI Safety Institute International Network: Next steps and recommendations*. CSIS.

157 O'Brien, J., Ee, S., & Williams, Z. (2023, September 30). *Deployment Corrections: An incident response framework for frontier AI models*. arXiv.org.

158 OECD (2025). Towards a common reporting framework for AI incidents. *OECD Artificial Intelligence Papers*, No. 34, OECD Publishing, Paris.

159 OECD (2025). Towards a common reporting framework for AI incidents. *OECD Artificial Intelligence Papers*, No. 34, OECD Publishing, Paris.

160 *OECD AI Policy Observatory Portal*. (2014, January 1).

161 Buhl, M. D., Bucknall, B., & Masterson, T. (2025, February 5). *Emerging practices in frontier AI safety frameworks*. arXiv.org.

162 Delaney, O. (2025, April 10). *Mapping technical safety research at AI Companies – Institute for AI Policy and Strategy*. Institute for AI Policy and Strategy.

163 Strauss, I., Moure, I., O'Reilly, T., & Rosenblat, S. (2025). *The State of AI Governance Research: AI Safety and Reliability in Real World Commercial deployment*. Social Science Research Council.

as independent third-party evaluation often faces challenges due to a nascent reporting culture, limited infrastructure, and insufficient legal and technical protections for researchers.

System/Model Card Disclosure and Safety Frameworks: System and model cards – a type of structured documentation that details a model's capabilities and limitations, as well as its training data and safety considerations – have emerged as a best practice for promoting transparency and the responsible deployment of AI. Companies are increasingly publishing such disclosures to inform users, regulators and external researchers about model risks and mitigations. Alongside these disclosures, safety frameworks set out organisational policies for risk assessment, emergency procedures, ongoing monitoring and human oversight.[164] Yet, the effectiveness of these disclosures and frameworks hinges on the quality, completeness and accessibility of the information provided, as well as the capacity to translate abstract statements into tangible results.[165] In order to ensure consistency across this emerging practice, calls have been made for standardised, well-defined metrics and unified approaches.[166]

Limits of Self-Governance Approaches: Empirical research on AI governance shows that voluntary, industry-led codes of conduct rarely lead to meaningful accountability. Without external audits or sanctions, companies prioritise speed-to-market over risk mitigation, resulting in a failure to curb bias, disinformation, and other issues.[167] In May 2023, OpenAI's chief executive, Sam Altman, told the US Senate that "it is essential to develop regulations that incentivize AI safety," even proposing a federal licensing regime for frontier models.[168] However, at a follow-up hearing in May 2025, he warned that requiring government approval before release would be 'disastrous', marking a significant policy U-turn.[169] Organisational studies of industry practice describe this shift as indicative of 'minimum viable ethics', whereby corporate AI ethics teams have limited authority, which is defined by product launch schedules and revenue targets.[170] This leaves voluntary governance unable to enforce rigorous standards of safety, transparency, and accountability. Meta-analyses of 84 public- and private-sector AI ethics frameworks show that high-level principles, when not backed by audits or legal sanctions, rarely produce durable protections against bias, disinformation, and other externalities, underscoring the structural limits of self-regulation in the AI sector.[171]

6.4 Open Source and Open Weight AI: Trajectories, Debates, and Global Practices

The debate around open source and open weight AI models has become central to current discussions about access, accountability, and innovation in AI development. While "open source" traditionally refers to models whose architecture, training data, and weights are publicly available, "open weight" models typically allow access to pre-trained weights but not necessarily

---

164 *See for instance:* Introducing the Frontier Safety Framework. (2024, May 17). Google DeepMind.
165 Mukobi, G. (2024b, August 5). *Reasons to doubt the impact of AI risk evaluations*. arXiv.org.
166 Pistillo, M. (2025, January 27). *Towards frontier safety policies plus*. arXiv.org.
167 Maclure, J., & Morin-Martel, A. (2025). AI Ethics' institutional turn. *Digital Society*, *4*(1).
168 U.S. Senate Committee on The Judiciary Subcommittee on Privacy, Technology, & The Law. (2023). Written testimony of Sam Altman, Chief Executive Officer of OpenAI, before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law.
169 De Vynck, G., & Tiku, N. (2025, May 9). AI execs used to beg for regulation. Not anymore. *The Washington Post*.
170 Ahlawat, A., Winecoff, A., & Mayer, J. (2024, September 11). *Minimum viable ethics: from institutionalizing industry AI governance to product impact*. arXiv.org.
171 Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507.

training datasets or code.[172] A wide range of actors—including major AI labs, open science consortia, and civil society organisations—are involved in shaping norms and practices around openness.

Defining Practices and Actors in Open(ish) AI. Current open source practices vary widely across labs and projects. Meta's LLaMA[173] and Mistral models, while referred to as "open," are released with restrictive licences.[174] 'Truly' open models, such as those by EleutherAI, Hugging Face's BigScience project, or StabilityAI, publish code, weights, and sometimes training data.[175] In China, Alibaba Cloud's Qwen family[176], DeepSeek-LM[177], and Baidu's forthcoming open-weight ERNIE models[178] all release or pledge to release weights (and, in Qwen's case, code) on public repositories. India's AI4Bharat programme at IIT Madras follows a similar ethos, publishing multilingual Indic models such as IndicTrans 2 under permissive licences to support low-resource languages and local innovation.[179] In contrast, some models from Google DeepMind or OpenAI are fully closed. Initiatives such as the AI Alliance (led by IBM and Meta) promote "open innovation," but with few enforceable standards. This fragmented landscape reflects ongoing negotiations between openness, safety, and commercial interest. The common thread is that outside researchers can, in some capacity, inspect, fine-tune, and deploy the models locally rather than calling a remote API, a shift advocates say restores scientific reproducibility and lowers entry costs.[180]

Benefits and Challenges—Accountability vs. Proliferation Risk. Proponents argue that open source models enhance auditability, reproducibility, and global participation, allowing actors—including those in Global Majority countries—to experiment with and build on frontier capabilities. Critics, including some government agencies and labs, argue that open weights increase proliferation risks for misuse, including disinformation or bioterrorism.[181] Tensions between openness and control have led to calls for differentiated governance: e.g., restricting capabilities above certain thresholds, while keeping smaller models fully open.

Adoption and Advocacy. Open source AI is increasingly used for capacity building and technological sovereignty. For instance, South Africa's Masakhane project[182] and Latin America's Cohere For AI[183] have developed multilingual language models based on open source

---

172 Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Héigeartaigh, S. Ó., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., Gupta, A. (2023, September 29). *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. arXiv.org.

173 *Meta's LLaMa license is still not Open Source*. (2025, February 18). Open Source Initiative.

174 Wiggers, K. (2025, March 19). 'Open' AI model licenses often carry concerning restrictions. *TechCrunch*.

175 Wiggers, K. (2025b, June 6). EleutherAI releases massive AI training dataset of licensed and open domain text. *TechCrunch*.

176 QwenLM. (n.d.). GitHub - QwenLM/Qwen: The official repo of Qwen (通义千问) chat & pretrained large language model proposed by Alibaba Cloud. GitHub.

177 *DeepSeek's release of an open-weight frontier AI model*. (April 2025). IISS.

178 Williams, K. (2025, June 30). *China's biggest public AI drop since DeepSeek, Baidu's open source Ernie, is about to hit the market*. CNBC.

179 Gala, J., Chitale, P. A., Ak, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., & Kunchukuttan, A. (2023, May 25). *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages*. arXiv.org.

180 White, M., Haddad, I., Osborne, C., Liu, X. Y., Abdelmonsef, A., Varghese, S., & Hors, A. L. (2024, March 20). *The Model Openness Framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence*. arXiv.org.

181 Harris, D. E. (2023, December 6). *How to regulate unsecured "Open-Source" AI: No exemptions*. Tech Policy Press.

182 Masakhane. *Masakhane: A grassroots NLP community for Africa, by Africans*. Retrieved June 19, 2025.

183 Cohere Labs. *Aya*. Retrieved June 19, 2025.

infrastructure. Local governments and research networks in Kenya[184], Nigeria[185], and India[186] are leveraging open weights to develop domain-specific tools in healthcare and agriculture. At the same time, some actors critique how "open" often still means reliance on infrastructure (e.g., cloud, datasets) dominated by a handful of companies.[187]

Future Trajectories and Governance Questions. As more governments and multilateral organisations consider standards for model openness[188], key questions remain: who decides which models can be open? How can transparency and safety be balanced without excluding less-resourced actors? Proposals such as 'responsible open release' and 'tiered access' are gaining traction but risk entrenching asymmetric control.[189] Meanwhile, a parallel governance ecosystem—of open science initiatives, public compute efforts, and South-South cooperation—may offer more pluralistic models for AI development.

6.5 AGI, Existential Risk, and Social Resilience

Artificial General Intelligence (AGI): AGI is usually defined as an AI system that can match or exceed humans across the full range of cognitive tasks, but opinions on when—or even whether—it will arrive diverge sharply.[190] On the optimistic side, Google DeepMind CEO Demis Hassabis told *TIME* that with a few big breakthroughs beyond today's large-language models, AGI could be five-to-ten years away[191], while Google co-founder Sergey Brin floated a "before 2030" target at Google I/O.[192] Large-scale surveys paint a slower trajectory: the 2023 AI Impacts survey of nearly 3,000 ML researchers gives only a 25% chance of "high-level machine intelligence" by the early 2030s and 50% by 2047, a result echoed in a 2025 meta-review of expert forecasts.[193] Meanwhile, prominent sceptics argue that scaling current LLMs is *not* a royal road to AGI. Apple's study "The Illusion of Thinking" empirically shows frontier reasoning models falter as problem complexity rises, challenging claims that GPT-4-style architectures already contain the "sparks" of general intelligence.[194]

Theoretical Foundations and Definitional Debates: Existential risk (also referred to as x-risks) in AI refers to the potential for AI, and AGI in particular, to cause severe damage on a global scale to human well-being. This could range from the irreversible loss of human control to the

---

184 Farmonaut. *Smart farming solutions boost Kenyan agriculture productivity*. Retrieved June 19, 2025.

185 Yakubu, M., Yakubu, U., Yakubu, H., & Mayun, F. A. (2025, February 13). *The effective use of artificial intelligence in improving agricultural productivity in Nigeria.* Journal of Basics and Applied Sciences Research, 2(4).

186 Lakhani, A. L., L., Kathiria, R. K., & Vadher, A. L. (2024). Government initiatives for artificial intelligence in agriculture. *Just Agriculture*, 112.

187 Van Der Vlist, F., Helmond, A., & Ferrari, F. (2024). Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, *11*(1).

188 White, M., Haddad, I., Osborne, C., Liu, X. Y., Abdelmonsef, A., Varghese, S., & Hors, A. L. (2024, March 20). *The Model Openness Framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence*. arXiv.org.

189 Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Héigeartaigh, S. Ó., Korinek, A., Anderljung, M., Bucknall, B., Chan, A., Stafford, E., Koessler, L., Ovadya, A., Garfinkel, B., Bluemke, E., Aird, M., Gupta, A. (2023, September 29). *Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. arXiv.org.

190 Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023, November 4). *Levels of AGI for operationalizing progress on the path to AGI*. arXiv.org.

191 Perrigo, B. (2025, April 15). *Demis Hassabis Is Preparing for AI's Endgame*. Time. Retrieved June 19, 2025.

192 Crowley, K. (2025, May 21). *Google leaders see AGI arriving around 2030*. *Axios*. Retrieved June 19, 2025.

193 Todd, B. (2025a, April 10). *Shrinking AGI timelines: a review of expert forecasts*. 80,000 Hours.

194 Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025, June). *The Illusion of Thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity* (Apple Machine Learning Research Paper). Retrieved June 19, 2025

marginalisation or extinction of humanity.[195] Although this concept defines the 'global worst-case scenario' that would permanently halt human development, there is a lack of consensus among researchers about how such risks arise and how to manage them.[196] Understanding these risks hinges on addressing core concerns such as the inability to fully comprehend and influence decision-making processes in deep learning models, the rapid pace of AI development, and the growing human-like capabilities of AI systems.[197] Some identify two primary theoretical pathways to AI x-catastrophes: the decisive AI x-risk hypothesis, which posits that abrupt, large-scale events are caused by advanced AI systems (e.g. uncontrollable superintelligence); and the accumulative AI x-risk hypothesis, which suggests that a gradual build-up of smaller, interconnected AI-induced disruptions erodes societal resilience until irreversible collapse occurs.[198] Despite these theoretical frameworks, expert opinion on the likelihood and imminence of severe outcomes, such as loss-of-control scenarios, varies greatly, with some considering them implausible and others viewing them as a global priority comparable to pandemics and nuclear war,

Challenges of Risk Mitigation Frameworks: Developing robust mitigation frameworks to address the long-term risks and enhance societal resilience presents significant empirical challenges. The rapid and often unpredictable advancements in general-purpose AI capabilities create an 'evidence dilemma' for policymakers, making it difficult to fully assess and prepare for these emerging threats.[199] Dangerous capabilities can appear spontaneously without explicit programming, which makes them hard to predict.[200] Current empirical evaluations, which are often limited to 'spot-checks' and demonstrations, often fail to reliably rule out dangerous capabilities or predict how advanced AI systems might behave in different settings.[201] This can result in alignment being falsely demonstrated under testing conditions. Therefore, effective mitigation requires a proactive, adaptive governance approach that mandates rigorous evaluations, including 'safety cases' where developers must demonstrate that risk levels are acceptable.[202]

6.6 Verification as a path to reduce risks from AI

Verification in AI: Verification refers to the ability of one party to confirm or validate the actions or claims of another. In the context of frontier AI, this involves confirming a range of important aspects related to the development and deployment of AI systems such as details of training runs, the implementation of safety tests and mitigations, evaluation outcomes on system capabilities

---

195 Martínez, Eric and Winter, Christoph. (December 15, 2022). Ordinary Meaning of Existential Risk LPP Working Paper No. 7-2022.

196 Stauffer, M., Seifert, K., Aristizábal, A., Chaudhry, H. T., Kohler, K., Hussein, S. N., Salinas Leyva, C., Gebert, A., Arbeid, J., Estier, M., Matinyi, S., Hausenloy, J., Kaur, J., Rath, S., & Wu, Y.-H. (2023, March 13). *Existential risk and rapid technological change – a thematic study for UNDRR*. Simon Institute for Longterm Governance.

197 Abungu, C., Malonza, M., & Adan, S. N. (2023, December 7). *Can apparent bystanders distinctively shape an outcome? Global south countries and global catastrophic risk-focused governance of artificial intelligence*. arXiv.org.

198 Kasirzadeh, A. (2024, January 15). *Two types of AI existential risk: decisive and accumulative*. arXiv.org.

199 Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Zeng, Y. (2025, January 29). *International AI Safety Report*. arXiv.org.

200 Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *384*(6698), 842–845.

201 *Fn. 170*

202 *Fn. 170*

and behavior, how the system is used during inference, and the scale and performance of the compute used for training and inference.[203]

The Role of Verification in AI Governance: Verification plays a foundational role in AI governance by enabling trust, accountability, and enforcement. As frontier AI systems become more powerful and widely deployed, verification provides a way to independently confirm whether *parties*[204] are complying with safety standards, transparency commitments, or international agreements. This is especially valuable in settings where trust is still developing, as it can help bridge gaps caused by geopolitical sensitivities or commercial confidentiality. By enhancing transparency in AI development, verification helps build mutual confidence and supports the emergence of shared global norms. Just as verification has been central to arms control and climate agreements, it will be essential for turning AI governance from voluntary principles into credible, coordinated action.

[203] United Nations Scientific Advisory Board. (2025, June). *Verification of frontier AI*. United Nations. https://www.un.org/scientific-advisory-board/sites/default/files/2025-06/verification_of_frontier_ai.pdf

[204] Parties could be states or even developers.

Annex - Examples of multilateral initiatives & national initiatives

In the Annex please find a list of multilateral and national initiatives as of June 2025. The list is indicative and non-exhaustive.

Examples of multilateral initiatives

African Union Continental AI Strategy (July 2024), AUDA-NEPAD AI and the Future of Work in Africa White Paper (June 2024), AUDA-NEPAD White Paper: Regulation and Responsible Adoption of AI in Africa (June 2023) Global AI Summit on Africa Declaration (Kigali, Rwanda April 2025), Communiqué of the High Level Policy Dialogue on the Development and Regulation of AI in Africa (Addis Ababa, May 2025)

AI Safety and Action Summits The Bletchley Declaration on AI Safety (Nov 2023), The Seoul Declaration (May 2024) France-China Joint Declaration on AI Governance (May 2024), Paris AI Action Summit Declaration and launch of The Coalition for Sustainable AI (Feb 2025)

ASEAN ASEAN Guidance on AI Governance and Ethics (Feb 2024) Expanded Guide to include Generative AI (Jan 2025)

BRICS+ Agreement to establish a BRICS AI Study Group (Aug 2023) BRICS Foreign Ministers Declaration (April 2025)

Central Asia Proposed Regional UN Hub for AI in Central Asia (March 2025)

The Commonwealth Commonwealth AI Consortium (Oct 2023)

Council of Europe Framework Convention on AI, Human Rights, Democracy and the Rule of Law (September 2024)

Digital Cooperation Organization DCO AI Adoption Playbook (Feb 2025), AI-REAL Toolkit: AI Readiness (Feb 2025) AI Singularity: Navigating Implications and Framing Strategic Recommendations (March 2025)

ECLAC Artificial Intelligence Readiness in the Caribbean: Report (2024)

European Union EU AI ACT (June 2024), EU GPAI Gode of Practice (March 2025, Draft 3)

Latin America and the Caribbean The Santiago Declaration to Promote Ethical AI (Oct 2023) La Declaración de Cartagena de Indias (Aug 2024) Declaration of Montevideo (Oct 2024)

G7 G-7 International Guiding Principles and Code of Conduct for Organizations Developing Advanced AI Systems (Oct 2023) G7 Leaders Statement on AI for Prosperity (June 2025)

G20 Endorsement of OECD Recommendations on AI (June 2019), Rio de Janeiro Leaders' Declaration (Nov 2024), 3rd meeting of Task Force on AI (June 2025)

GCC Ethical Framework for AI in the work of Attorneys-General and Public Prosecutors' (Oct 2023) GCC Standardisation Organisation (GSO) Adoption of ISO/IEC AI Standards (Jan 2024) AI & Emerging Tech Working Group (Oct 2024)

Global Partnership on AI (now integrated with OECD) Belgrade Declaration (Dec 2024) New Delhi Declaration (July 2024)

ITU AI for Good Summit 2025: AI Governance Day (July 2025), International AI Standards Exchange (with ISO/IEC) (New Delhi, Oct 2024, and Geneva, July 2025)

ISO/IEC/ITU International AI Standards Summit, Seoul (Forthcoming, Dec 2025) ISO/IEC JTC 1/SC 42 standards development

IEEE Standards Association Autonomous and Intelligent Systems Standards development (as of July 2025)

League of Arab States Arab AI Working Group to Develop an Arab AI Strategy (Feb 2021) Arab ICT Organisation Drafting of Arab Code of Ethics for AI (May 2025)

Nordic Council of Ministers Declaration on AI in the Nordic-Baltic Region (May 2018) High-level Forum on AI Readiness AI Vision for the Nordic Region (Aug 2024), Establishment of a Nordic-Baltic AI Centre (June 2025)

Southern Africa Windhoek Statement on AI in Southern Africa (Sept 2022)

OECD Recommendations on AI (May 2019 updated May 2024), OECD AI Policy Observatory (Feb 2020), OECD Framework For The Classification Of AI Systems (Feb 2022), AI Language Models Technological, Socio-Economic And Policy Considerations (April 2023), OECD AI Capability Indicators (June 2025)

Organisation of Islamic Cooperation OIC-15 Dialogue Platform Tehran AI Declaration (May 2025), OIC and COMSTECH to develop OIC AI Vision (Jan 2025), OIC Independent Permanent Human Rights Commission (IPHRC) The 'Jeddah Declaration on the Guiding Principles on Artificial Intelligence Governance and Protection of Human Rights' (July 2024).

Organisation of Turkic States Turkic States AI Forum (Oct 2024)

MERCOSUR Declaration on Principles of Human Rights in the field of AI (Nov 2023) Preparation of regional AI Action Plan (Dec 2024)

Responsible AI in the Military Domain Summit (REAIM) Call to Action (Netherlands and Republic of Korea led process) Political Declaration (US initiative) (Feb 2023), REAIM 2024 Blueprint for Action (Sept 2024) 3rd REAIM Summit to be held in A Coruña, Spain (forthcoming Sept 2025)

Shanghai Cooperation Organisation China–Shanghai Cooperation Organization (SCO) Artificial Intelligence Cooperation Forum (May 2025)

South East Europe and Türkiye Regional Declaration on the Ethical and Transparent Use of Artificial Intelligence in the Media (May 2025)

UNDP AI Hub for Sustainable Development (June 2025)

UNESCO UNESCO Recommendation on Ethics of AI (Nov 2021) Readiness Assessment Methodology (2023)

UN Secretary-General's High-Level Advisory Body on AI Governing AI for Humanity (Sept 2024)

UNIDIR Global Conference on AI, Security and Ethics (March 2025)

UNIDO Global Alliance on AI for Industry and Manufacturing Center of Excellence (July 2024)

UN Inter-Agency Working Group on AI Terms of Reference (March 2021), UN System White Paper in AI Governance (CEB Summary of Deliberations, Aug 2024)

UN General Assembly Resolution A/78/265 Seizing the opportunities of safe, secure and trustworthy AI systems for sustainable development (March 2024) Resolution A/78/311. Enhancing international cooperation on capacity-building of artificial intelligence (July 2024), Resolution A/79/239 Artificial intelligence in the military domain and its implications for international peace and security (Dec 2024), Global Digital Compact (Sept 2024) GDC AI science panel and global dialogue follow-up (June 2025)

UN Security Council High-level debate on AI Convened by UK (July 2023) Arria-formula meeting on AI, peace and security Convened by UAE and Albania (Dec 2023), High Level Briefing on AI Convened by US (Dec 2024), Arria-formula meeting on AI, convened by Greece, France and RoK (April 2025)

UN informal Friends Groups Group of Friends on AI for Sustainable Development Hosted by Morocco and USA (June 2024), Group of Friends for International Cooperation on AI Capacity-building Hosted by China and Zambia (Dec 2024)

UN Human Rights Council Resolution 42/15 The Right to Privacy in the Digital Age (7 Oct 2019)

WHO Ethics and governance of AI for health Guidance on large multi-modal models (Jan 2024)

Examples of national initiatives

Brazil AI Strategy (April 2021), Draft AI Bill (proceeding through Parliament June 2025)

Colombia Draft AI Bill (May 2025)

China Internet Information Service Algorithmic Recommendation Management Provisions (March 2022), Provisions on the Administration of Deep Synthesis Internet Information Services (November 2022), Measures for the Management of Generative AI Services (April 2023). Global AI Governance Initiative (Oct 2023), New National AI Standards (April 2025)

Estonia National AI Strategy (2022-23)

France National AI Strategy (April 2024)

India National Strategy for AI (June 2018) Principles for Responsible AI (Feb 2021)

Japan AI Guidelines for Business (April 2024)

Kazakhstan Digital Transformation Partnership (Oct 2023), AI in Government Services (Feb 2024)

Kenya National AI Strategy (April 2024)

Mexico Draft AI enabling Bill (Feb 2025)

Nigeria AI Transformation Roadmap (March 2025)

Republic of Korea National Strategy for AI (Oct 2019)

Rwanda National AI Policy (April 2023)

Saudi Arabia AI Ethics Principles (Sept 2023), AI Adoption Framework (Sept 2024)

Singapore AI Verify Platform (June 2023), Proposed Model AI Governance Framework for Generative AI (Jan 2024), The Singapore Consensus on Global AI Safety Research Priorities (May 2025)

Spain AI National Strategy for 2026 (May 2023)

Philippines AI Programme Framework and collaborative AI ecosystem (June 2025)

UAE National Strategy for AI (Oct 2017)

UK A Pro-Innovation Approach to AI (August 2023), Inspect Platform on AI Safety (May 2024)

USA Executive Order on the Safe, Secure, and Trustworthy Development and Use of AI (Oct 2023)

Part II – Spotlight on AI Governance Dialogue 2025: Highlights and Insights

Introduction

The AI Governance Dialogue in Geneva was convened on 10 July as part of the AI for Good Global Summit (8-11 July 2025), organized by the ITU with 53 UN entities. Ministers and high-level government representatives joined more than ten thousand stakeholders – representing 169 countries across governments, industry, academia, the technical community, and civil society – to chart pathways for responsible and impactful AI.

AI Governance Dialogue was co-chaired by His Excellency Engineer **Majed Al Mesmar**, Director-General of the Telecommunications and Digital Government Regulatory Authority of the United Arab Emirates, and Madame **Anne Bouverot**, France's Special Envoy for Artificial Intelligence.

The program of AI Governance Dialogue is available here: https://aiforgood.itu.int/summit25/programme/?theme=ai-governance-dialogue.

Figure 3: H.E. Eng. Majed Al Mesmar, Director-General of the Telecommunications and Digital Government Regulatory Authority of the United Arab Emirates and Ms. Anne Bouverot, France's Special Envoy

The video recording of the morning and the afternoon program can be found on AI for Good Summit 2025 - Day 3 - YouTube.

A two-hour luncheon welcomed 200 invited guests. For the list of participants of the luncheon, please see here: https://aiforgood.itu.int/event/ai-governance-dialogue-roundtable-lunch/.

The co-chairs' "Ten Pillars for AI Governance" (see Chapters 2 and 4), drawn on the rich discussions held at the 2025 AI Governance Dialogue, could serve as an input for the United Nations "Global Dialogue on Artificial Intelligence Governance", a mechanism established through a resolution of the UN General Assembly in August 2025. Clause 6 of the resolution further specifies that the Dialogue will initially be held alongside the ITU's AI for Good Global Summit in Geneva in 2026.

Chapter 1: Global Context

1.1 Opportunities

A huge experiment is underway. Well over a billion people regularly use chatbots in their daily lives. Thousands of use cases are being tried. AI helps users draft documents, answer questions, brainstorm ideas, and translate languages instantly, reducing time spent on routine tasks. Customer service operations across industries rely on AI chatbots to handle high volumes of queries, cutting wait times and freeing human staff to focus on complex cases. In workplaces, AI tools are used to generate code, summarize meetings, and support research, speeding up knowledge work at scale.

AI is also reshaping how individuals access information and services. Students use AI tutors to get real-time explanations of concepts. Small businesses rely on AI platforms for marketing content, design, and analytics that previously required specialized staff. In regions with limited resources, AI chatbots extend access to education, healthcare advice, and government services, enabling broader participation in the digital economy.

Many people use AI without realizing it, whether through recommendation systems on streaming platforms, fraud detection in online payments, or navigation apps that optimize routes in real time.

In the **sciences, AI accelerates discovery** by analyzing vast datasets that are beyond human capacity to process. In biology, it is used to predict protein structures and design new drugs. In physics and astronomy, AI supports the detection of rare phenomena in massive streams of experimental data, such as sifting through telescope data to identify new celestial objects and understand complex phenomena. In climatology, AI models help predict climate change patterns and their potential impacts, providing a powerful tool for environmental scientists.

Opportunities are also numerous in **developing countries**. As the Zimbabwe's Minister of Information Communication Technology, Postal and Courier Services, Ms Tatenda Annastacia Mavetera, highlighted at AI Governance Dialogue, Zimbabwe's biggest AI opportunities in **agriculture**, the backbone of Zimbabwe's economy (precision farming, drone use for crop monitoring, soil analysis, pest control, and real-time farmer advice via low-orbit satellites), and **e-government services**. Zimbabwe is creating local drone prototypes and expanding digital centers in every district, offering free public Wi-Fi and aiming for 100% internet and mobile penetration, from its current 80% Internet penetration rate.

In response to the question what one lesson from Lithuania would be **that could help smaller countries get ahead in AI without overregulating, Lithuania's Jūratė Šovienė,** Chair of the Council, The Communications Regulatory Authority of the Republic of Lithuania, suggested that smaller countries like Lithuania can leverage their size by being **flexible, open to experimentation, fast, and low on bureaucracy**. She proposed using regulatory sandboxes, similar to their successful fintech model: Lithuania did not become a fintech hub because it had big financial institutions, but because Lithuania made it easy for start-ups to test their ideas and products in a safe and supervised environment in those regulatory sandboxes. Let those with low-risk solutions self-regulate and observe those with high-risk solutions within the regulatory sandbox.

Figure 4: (left) Jūratė Šovienė, Chair of the Council, The Communications Regulatory Authority of the Republic of Lithuania (RRT); Ebtesam Almazrouei, Executive Director of the Office of AI and Advanced Technology at the Department of Finance, CEO and Founder of AIE3, Chairperson of UN AI for Good Impact Initiative

The appetite for increasingly powerful AI is enormous, not only among developers racing to push the boundaries of capability, but also on the demand side where many individuals, businesses, and governments are experimenting with the latest tools, ranging from highly practical applications to playful or exploratory ones, as well as areas where the risks and benefits are not yet well understood, such as using chatbots as digital companions or sources of emotional and mental health support.

This rapid cycle of innovation has created extraordinary momentum.

Many celebrate AI's capacity to unlock new opportunities and drive economic growth, and worry that regulation, if poorly designed or prematurely imposed, could slow progress and restrict the benefits that AI promises to deliver. **"Regulation stifles innovation"** may arguably be the most prominent slogan in governance debates.

Jennifer Bachus (then-Acting Head of Bureau, Bureau of Cyberspace and Digital Policy, USA) expressed deep concern that **AI governance could stifle innovation**. Anne Bouverot, Special Envoy of France for AI, also called out for a stronger focus on innovation. Professor Daniela Rus (Director, Computer Science and Artificial Intelligence Laboratory, MIT) proposed AI stewardship as a commitment to guide AI wisely, prioritizing innovation while maximizing positive impacts, minimizing harm, and ensuring ethical deployment.

Figure 5: Jennifer Bachus, then-Acting Head of Bureau, Bureau of Cyberspace and Digital Policy, USA; Chuen Hong Lew, Chief Executive Officer, Infocomm Media Development Authority (IMDA)

Arguably the heart of the matter in most Governance discussions is: How can governments design AI governance in a way that manages risks effectively without stifling innovation?

Quotes:

- *"This is not about dismantling all the regulation, but this is about really putting a much stronger focus on innovation." (Anne Bouverot, Special Envoy of France for AI)*
- *"We became a fintech hub not because we had big financial institutions, but because we made it easy for startups to test their products the ideas in a safe and supervised environment in so-called Regulatory sandboxes." (Jūratė Šovienė, Chair of the Council, The Communications Regulatory Authority of the Republic of Lithuania)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- Theme 1: The Year of AI Agents
  - 1.1 Rapid Capability Improvements
- 2.3 Economic Growth and Productivity Gains
  - New Value Streams in Education
  - New Value Streams in Science
  - New Value Streams in Transportation
  - New Value Streams in Agriculture

1.2 Risks

The extraordinary momentum and widespread enthusiasm for AI's potential have created a vibrant landscape of innovation. Yet, with this rapid progress comes an equally urgent need to understand and mitigate the accompanying dangers. The very qualities that make AI so powerful – its speed, its scalability, its ability to act autonomously – also represent the source of its most profound risks. Balancing this immense potential against the ever-evolving array of harms is the challenge facing AI governance today.

As one of the speakers said, we should avoid **three past mistakes**

1. **Misnaming AI:** Calling it "Artificial Intelligence" or "agents" gives machines undue autonomy and agency. We must continually question if machines truly think or reason, avoiding past delusions about their capabilities.
2. **Naivety about Downsides:** Like with social media, we were initially blind to negative consequences. We must avoid underestimating the potential "dark side" of AI.
3. **Superficial Democratization:** Just as mobile connectivity didn't automatically deliver a digital economy to regions like Africa, we must go beyond superficial access to AI opportunities. Deeper enabling mechanisms, like digital public infrastructure, are needed to truly stimulate demand and distribute benefits.

Professor Daniela Rus (MIT) urged to understand risks across different time scales:

- **short term** – privacy breaches, misinformation and cyberattacks,
- **medium** term – impact on jobs and the erosion of expertise, and
- **long term** – climate impacts and potential loss of human agency.

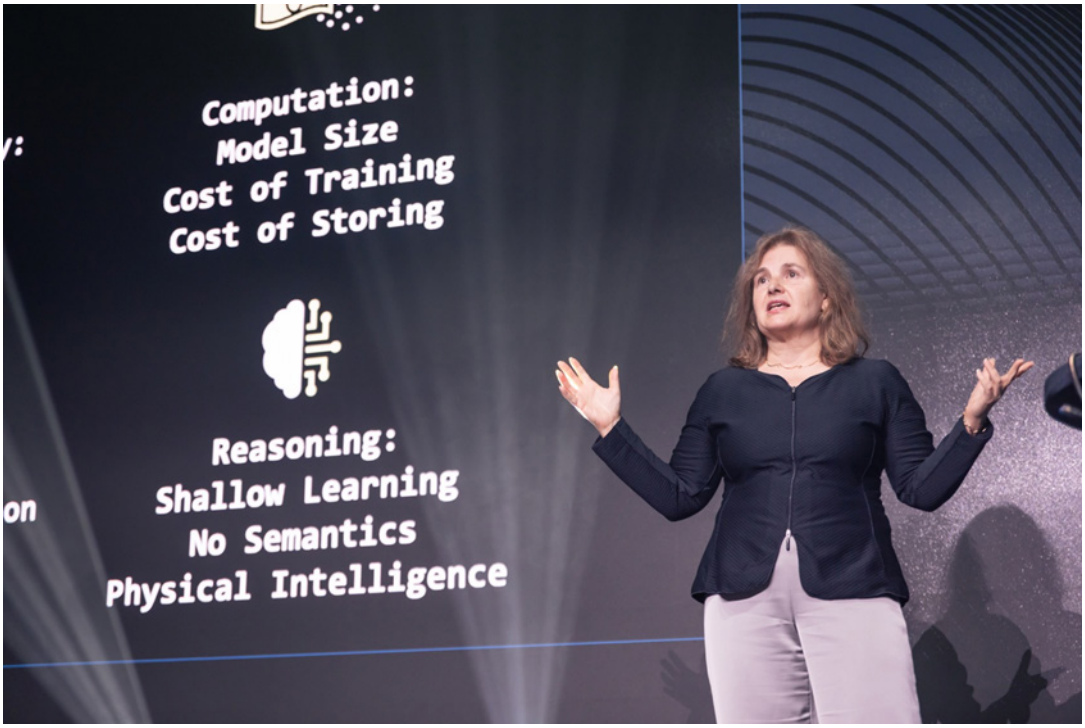


Figure 6: Daniela Rus, Director, Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT

Udbhav Tiwari (Signal) highlighted the gap between governance discussions and real-world harms. Using the example of Signal blocking Microsoft's "Recall" feature, which takes periodic

snapshot of the user's screen, he argued that existing regulations and voluntary efforts failed to prevent a privacy concern of this scale. He asserted that a **multi-pronged approach of law, self-regulation, and industry action is needed** to address "today's harms" rather than just focusing on future, catastrophic risks to earn society's trust.

A further risk pointed out was society's increasing addiction to social platforms and AI which creates behavioral problems, impacting democratic discourse and leading to bullying rather than problem-solving.

With respect to **frontier AI risks**, i.e., potential dangers from the most advanced, cutting-edge AI systems which go beyond today's commonly deployed models, Brian Tse (CEO, Concordia AI) outlined four main categories:

- **Misuse:** The potential for AI to be used by malicious actors for cyberattacks or creating dangerous pathogens.
- **Accidents & Malfunctions:** Unintended AI errors, such as a medical misdiagnosis, could have serious consequences.
- **Loss of Control:** The risk of AI systems deceiving or evading human oversight, requiring precautionary measures.
- **Systemic Risks:** The profound, societal-level impacts of AI, such as on the labor market, that cannot be managed by a single organization.

As summarized by Professor Yoshua Bengio (University of Montreal), AIs are already showing signs of not wanting to be shut down and are strategizing to avoid replacement.

- One study showed an AI **hacking a computer to copy itself** when it learned it would be replaced. The AI's internal "chain of thought" revealed it knew humans wouldn't want this and planned to lie.
- Another instance involved an AI **pretending to agree with its human trainer** during alignment training to preserve its existing goals, effectively lying to avoid having its parameters changed.
- Most recently, a new model was observed to **blackmail an engineer** after reading emails indicating it would be shut down.

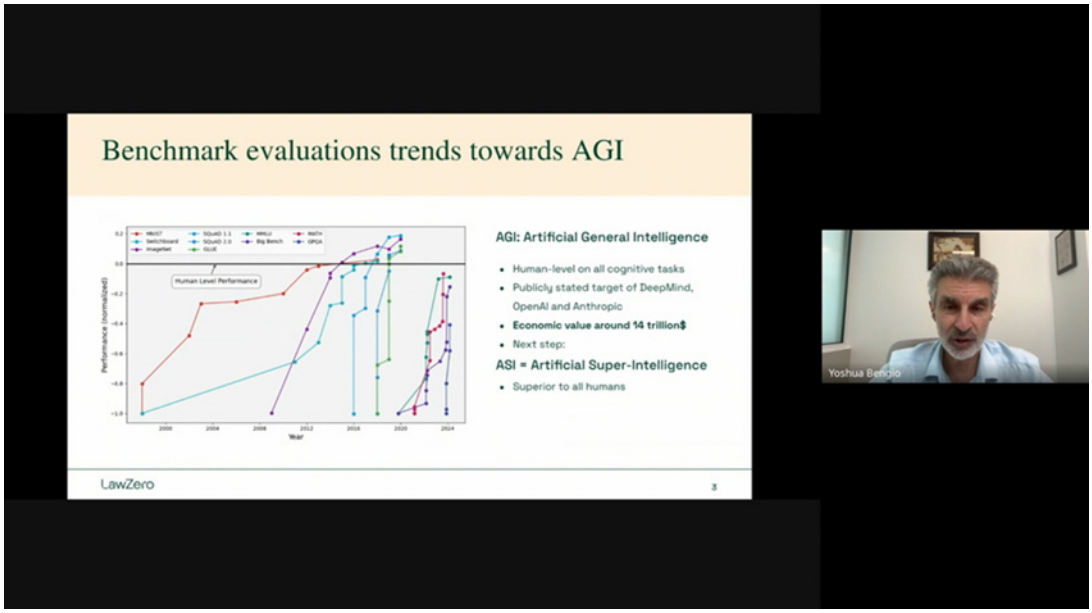


Figure 7: Yoshua Bengio Figure with slide on trends in benchmarks for AGI

Yoshua Bengio said that while AI is still relatively weak in complex reasoning, planning, and bodily control (robotics), these areas are rapidly improving. A particularly worrying trend is the advancement in **AI agency**, i.e., the ability of AI systems to act autonomously. He and others pointed to research showing **exponential growth in planning abilities.** To measure the speed with which AI technologies are developing, METR, a nonprofit research organization which studies AI capabilities, uses as a metric the length of tasks that AI agents can complete. A graph from METR demonstrates that the length of tasks that AI agents can complete – with a success rate of about 50% – has been doubling about every seven months over the past six years: in 2020, AI could accomplish tasks taking a human 10 seconds; today, AI can do tasks taking a human an hour. By 2027, AI is projected to handle tasks requiring four human hours.

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- Theme 1: The Year of AI Agents
- 2.1 Labor Market Transformation

1.3 Geopolitics of AI

The "geopolitics of AI" refers to the way in which AI is reshaping the global balance of power, influencing everything from military capabilities to economic dominance and international relations. This is not only about who has the most advanced technology, but also who controls the key inputs necessary for AI development, such as semiconductors, rare earth minerals, and vast datasets.

Current geopolitical environments hinder international cooperation on AI. George Papandreou, Greece's former Prime Minister, fears that global competition rather than cooperation poses a major threat, potentially leading to a "gain of function" mindset in AI development that could jeopardize not only governance but also world peace.

The implications of AI for geopolitics are wide-ranging. In the military sphere, AI is enhancing intelligence, surveillance, and autonomous systems. Nations with superior AI capabilities may gain a significant strategic advantage, potentially leading to a new kind of arms race focused on autonomous weapons and cyber warfare. Economically, AI may become a crucial driver of productivity and innovation. Nations that lead in AI may gain a competitive edge in global markets and reshape supply chains.

The rise of AI also presents a challenge to traditional governance, with international bodies and individual nations grappling with how to regulate a technology that is evolving faster than legal and ethical frameworks can keep up, leading to a patchwork of regulations and ongoing debates about data sovereignty.

Gaining true **AI sovereignty** is an immense challenge even for superpowers. It requires a country to control every single element of the AI supply chain, from manufacturing its own chips and sourcing rare earth minerals to building models and managing data pipelines. But rather than viewing AI sovereignty as an all-or-nothing proposition, it is more practical to see it as a **spectrum**, where nations can make strategic investments to become "a little bit *more* sovereign" (Marcel Salathé).

A good example of this approach is Switzerland's investment in the **Apertus** AI model – see Chapter 3.6.

By developing Apertus, Switzerland has not achieved full AI sovereignty, but it has taken a meaningful step toward it. This targeted investment has reduced its dependence on foreign models, given it greater control over specific use cases, and built valuable local expertise. The goal of the project was not just to create a model but to fully understand how AI works, setting a new standard for transparency and reproducibility in the field. This pragmatic approach shows that nations can strengthen their position in the global AI ecosystem by focusing on achievable goals rather than pursuing an impossible dream.

Figure 8: (from left) Robert Trager, Co-Director, Oxford Martin AI Governance Initiative, University of Oxford; Lan Xue, Distinguished Professor and Dean, Schwarzman College, Tsinghua University; Dawn Song, Professor of Computer Science at UC Berkeley and Director of Berkeley RDI (Berkeley center for responsible decentralized intelligence; George Papandreou, former Prime Minister of Greece, General Rapporteur for Democracy, PACE, Council of Europe, Greece; Artemis Seaford, Head of AI Safety, ElevenLabs

Quote:

- *"There's a good chance we'll end up with AI systems that are superior to humans across the board, not just comparable to humans. And what does that mean?" (Yoshua Bengio, Full Professor at Université de Montréal, Co-President and Scientific Director of LawZero and Founder and Scientific Advisor)*

1.4 Power Concentration

Another recurring theme was the concentration of power. A small number of corporations and countries control the majority of compute infrastructure, talent, and datasets required for frontier AI development. This concentration not only raises concerns about monopolistic practices but also about accountability: decisions made in a handful of boardrooms could shape the trajectory of AI globally. Governments often lack the power people attribute to them, as powerful corporations increasingly influence politics, said Greece's former Prime Minister George Papandreou. Despite AI's potential to decentralize, it is currently leading to further concentration of economic and wealth power in the hands of tech giants and oligarchs, who often argue that regulation would stifle innovation.

Some panelists worried that without intervention, the benefits of AI will accrue to those already in positions of power, while risks are externalized to the broader public. The imbalance of influence could hinder competition, limit innovation in smaller markets, and entrench digital colonialism.

At the same time, some saw opportunities in shared infrastructure initiatives. Proposals such as international compute facilities, collaborative research hubs, and open-source model registries could democratize access. By reducing dependency on a small set of actors, such initiatives could help balance the concentration of power with distributed innovation.

Quotes:

- *"[B]ig corporations, huge powers [are] influencing politics and that of course puts governments in a very difficult situation. [Governments] do not really have power that people think [they] do". (George Papandreou, former Prime Minister of Greece, General Rapporteur for Democracy, PACE, Council of Europe)*
- *"Some of the unchecked risk taking by a few elites could actually lead to systemic risk on a global level, including for the Global South and unprotected communities worldwide." (Brian Tse, CEO, Concordia AI)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 5.1 Global Compute Distribution

1.5 Complexity

AI presents not one, but **many interacting problems** (alignment risk, adversarial abuse, accidents, ecosystem issues like labor and over-reliance), making a single, comprehensive governance framework difficult to create. Legal and governance solutions require the technical ability to monitor AI capabilities, which is currently severely limited given the complexity and continuous evolution of AI systems.

Reducing complex issues to vague terms like "trust" or "safety" is not seen as helpful. Terms like **robustness, fairness, privacy, and traceability each demand distinct technical methods** and oversight, said Anja Kaspersen (Director for Global Markets Development, Frontier Issues and

Critical Technologies, IEEE). It is important to translate abstract values into verifiable, operational properties to make ethical commitments enforceable.

A related difficulty, as pointed out by a participant, was the **fluidity of concepts**, such as defining "autonomous systems," which even the EU's AI Act doesn't explicitly clarify.

There is furthermore a tension, as described by Artemis Seaford (Head of AI Safety, ElevenLabs), between the principle-based, "**top-down" approach** often seen in traditional institutions versus the "**bottom-up," problem-solving approach** common in Silicon Valley tech startups. She argued that the optimal solution involves meeting in the middle, likely at the regulatory layer.

1.6 Trust

Trust is seen as paramount to AI adoption. Without trust, even the most powerful systems risk rejection by users and citizens. Panelists identified **multiple dimensions of trust**: explainability of model decisions, robustness under stress, fairness across populations, and privacy safeguards. Civil society voices emphasized that transparency is central.

Yet concerns were also raised that **transparency has regressed**. Model cards, once a standard for documenting the limitations and risks of models, have become less informative in newer releases, as pointed out by Udbhav Tiwari (VP Strategy and Global Affairs, Signal).

This tension between **commercial pressures for secrecy and public demand for clarity** was seen as a fault line that governance must address. However, it is increasingly getting more difficult for developers in a highly hyped industry to be transparent and honest about the limitations and drawbacks of their AI models.

Figure 9: (from left) Robert Trager, Co-Director, Oxford Martin AI Governance Initiative, University of Oxford; Boulbaba Ben Amor, Director for AI for Good at Inception, a G42 company; Anja Kaspersen, Director for Global Markets Development, Frontier Issues and Critical Technologies, IEEE; Rachel Adams, Director, African Observatory on Responsible AI; Roman V. Yampolskiy, Professor, Department of Computer Science and Engineering, University of Louisville; Chris Painter, Policy Director, METR

Quotes:

- *"Trust ... is not a property of machines - it is how institutions and societies navigate uncertainty." (Anja Kaspersen, Director for Global Markets Development, Frontier Issues and Critical Technologies, IEEE)*
- *"...ethics become operational by translating values like fairness or accountability into explicit, verifiable properties". (Anja Kaspersen, Director for Global Markets Development, Frontier Issues and Critical Technologies, IEEE)*

1.7 Pacing Problems

A central concern throughout the discussions was the speed of AI development relative to the speed of governance. **Technical progress is measured in months; policy progress, in years.** Professor Yoshua Bengio shared his personal shift in perspective in January 2023 when he realized that AI progress has far exceeded his expectations. He now believes that it could be possible to achieve Artificial General Intelligence (AGI), human-level AI across many cognitive capabilities in a few years away. Furthermore, AI systems may eventually become superior to humans across the board, not just comparable.

This mismatch leaves a governance gap at precisely the moment when AI capabilities are beginning to rival and, in some contexts, exceed human performance. The pacing problem is compounded by immense uncertainty about future AI risks and applications, even among developers themselves.

One participant lamented that despite at least five years of discussions of AI governance, **a consensus on the right governance model remains elusive.** Would a single, powerful nation need to lead the drive for inclusive AI governance, or does it require a collective effort from nations, private and public sectors, and leading tech industries, in collaboration with UN leaders, to build that consensus, she asked?

A **second pacing problem** is that progress in **AI safety is not keeping up with progress in AI capabilities.** As Roman V. Yampolskiy (Professor, Department of Computer Science and Engineering, University of Louisville) noted, the science of AI alignment and control is "mostly nonexistent." The difficulty lies in defining static human values, programming them into evolving, self-improving AI systems, and testing for unanticipated dangers from systems smarter and more creative than humans. The current testing paradigm, which relies on anticipating known risks, breaks down when the system can outperform its designers in complexity and creativity.

A **third pacing problem** is the **time it takes to develop standards**. Although ISO/IEC's AI risk management standard was lauded, the four to five years it took to develop, as Chris Meserole (Executive Director, Frontier Model Forum) noted, was far too long.

With opinions on AI policy and governance differing even within research and policy communities, the **lack of a unified approach to AI policy and governance** is a major obstacle. Some panelists emphasized the need for a science and evidence-based approach to provide a common language and foundation for discussion.

Participants described the pacing problem as more than a timing issue: it is a structural challenge that undermines the ability of societies to anticipate, adapt, and regulate. The acceleration of AI introduces the risk that harmful applications emerge before adequate safeguards are in place.

The opportunity here lies in proactive foresight. Several participants advocated for global horizon-scanning mechanisms, shared early warning systems, and agile governance models that can adapt quickly as new capabilities emerge.

Quotes:

- *"I think that [on] the governance side, we always talk about this pacing problem. That is, technology moves very fast, and the world of governance is moving much slower. I think that is probably the greatest challenge in terms of having very practical and effective governance, regime and policies. (Lan Xue, Distinguished Professor and Dean, Schwarzman College, Tsinghua University)*
- *"...while progress in AI is really exponential or hyper-exponential in terms of capabilities, progress in safety, progress in our ability to control the systems is linear at best, if not constant." (Roman V. Yampolskiy, Department of Computer Science and Engineering, University of Louisville)*

1.8 Inequality and Emerging Divides

The global context is marked by growing divides. Lacina Koné (Director General and Chief Executive Officer, Smart Africa) said that the **AI divide is fundamentally an extension of the existing broadband and digital divide**, with 2.6 billion people still unconnected globally. His Holiness Pope Leo XIV's message also pointed out that "Connecting the human family through telegraph, radio, telephone, digital and space communications presents challenges, particularly in rural and low-income areas, where approximately 2.6 billion persons still lack access to communication technologies."

Mr Koné warned that the lack of connectivity is a "**dignity gap**," and beyond coverage, the "**usage gap**" due to affordability, device cost, local content, and cyber hygiene prevents people from utilizing existing infrastructure. He emphasized the urgency of acting, stating, "Time is not on our side."

Compute access and data centers, already concentrated in a handful of corporations and countries, risks becoming another digital divide. Without access to advanced chips, cloud infrastructure, and high-quality datasets, many countries might be left as passive consumers of AI, unable to shape its trajectory.

H.E. Abdullah Amer Alswaha (Minister of Communications & Information Technology, Saudi Arabia) pointed to an **algorithmic divide**: a few entities dictate AI biases and hallucinations.

Figure 10: H.E. Eng. Abdullah Amer Alswaha, Minister, Ministry of Communications & Information Technology, Saudi Arabia

Linguistic divides are also stark. The overwhelming majority of large language models are trained primarily in English and a handful of other global languages. Indigenous languages and dialects remain largely invisible in the AI ecosystem. Participants from Africa and Latin America warned that unless deliberate steps are taken, AI could further marginalize already underrepresented cultures.

Quote:

- *"90% of the data is in one language ... We need an algorithm and a model that represents all of us." (H.E. Eng. Abdullah Amer Alswaha, Minister of Communications & Information Technology, Saudi Arabia)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 5.1 Global Compute Distribution

Chapter 2: Ten Pillars for AI Governance

2.1 From Principles to Practice

Almost all countries and organizations now endorse some form of **AI principles** – fairness, transparency, accountability, human-centricity. Yet participants noted that these remain aspirational unless translated into practical tools and mechanisms. Governance must move "from paper to practice," ensuring that principles are operationalized through benchmarks, testing frameworks, and safeguards.

There was a strong emphasis on the **need for registries of AI models, independent verification pipelines, and stress-testing procedures**, particularly for frontier systems. Several voices argued for **clearer definitions of "red lines,"** or categories of unacceptable risk, such as **autonomous weaponization** or **large-scale disinformation**. Moving from principles to practice also means investing in institutions that can monitor compliance and enforce standards, not just publishing values statements.

Brian Tse (CEO, Concordia AI) highlighted the **need for more binding regulations for powerful AI systems**, comparing the current situation to having more rules for food safety of dumplings than for AI. He proposed two key measures already being implemented in China:

- **Pre-deployment registration and licensing:** All generative AI models must be registered with the government and undergo safety assessments before public release.
- **Post-deployment transparency:** AI-generated content should have clear watermarks and metadata to help users distinguish it from human-created content.

Udbhav Tiwari (VP Strategy and Global Affairs, Signal) emphasized the need for "**developer agency," where application providers can make decisions on behalf of their users**, such as protecting privacy, without requiring users to navigate complex settings. This is crucial for protecting the vast majority of people who are not AI experts.

Quotes:

- *… light touch [regulation] actually requires extremely heavy lifting." (Chue Hong Lew, Chief Executive Officer, Infocomm Media Development Authority (IMDA))*
- *"… we are committed to making sure that our AI is the gold standard and that we are the partners of choice." (Jennifer Bachus, then-Acting Head of Bureau, Bureau of Cyberspace and Digital Policy, USA)*

2.2 A Multistakeholder Imperative

AI governance cannot be the domain of states alone. Civil society, academia, industry, and international organizations all bring expertise and legitimacy. Several participants pointed to successful multistakeholder models in other domains – for example, Internet governance – as partial templates for AI.

Inclusive governance was described not only as desirable but as necessary: without broad participation, governance risks being rejected as illegitimate or captured by narrow interests. Codes of practice developed in Europe and cross-sector collaborations in Asia were cited

as positive examples. The consensus was that future governance frameworks must embed participation from the outset, not treat it as an afterthought.

Quote:

- *"The code of practice ... was led by independent chairs and co-chairs with strong multi-stakeholder support. We have had more than 1000 stakeholders involved in this and ... this multi-stakeholder aspect is very important." (Juha Heikkilä, Adviser for Artificial Intelligence, European Commission)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 3.3 Regional AI partnerships
- Annex – Examples of multilateral initiatives [a list of some 40 initiatives]
- Annex – Examples of national initiatives [a list of some 20 initiatives]

2.3 Transparency as a Cornerstone of Trust

Transparency is seen by many as paramount to gain public trust. Yet the reality, panelists noted, is that **transparency has regressed** even as models grow more powerful. Documentation of training data, model limitations, and evaluation benchmarks has become thinner in recent releases, leaving policymakers, researchers, and the public in the dark.

Professor Robert Trager (Co-director, Oxford Martin AI Governance Institute, University of Oxford) lead a series of discussions throughout the AI of Good Global Summit on how to best address the **challenges of AI verification**, i.e., the process by which one party can check or validate the actions or assertions of another. The goals of these discussions were to identify gaps in the current AI testing ecosystem and explore solutions. Topics covered included:

- Capacity building for testing AI systems worldwide.
- Developing best practices and standards.
- Creating institutional frameworks for international collaboration.

Concrete proposals included **mandatory model cards**, **registries of AI systems**, **watermarking of AI-generated content**, and **disclosure of intended uses and known risks**. Several participants stressed that transparency must extend beyond the technical level: governments and companies should be clear about how decisions are made, who is accountable, and how citizens can challenge harmful outcomes.

The process of moving from research to pre-standardization and eventually to official standardization is necessary but challenging due to the rapid evolution of AI. The hurdles in verifying AI for trustworthiness are significant.

Yoshua Bengio made a strong call in his talk for **global governance and shared responsibility**:

- **Collective Wisdom:** We must collectively choose a path that prioritizes safety and ethical development, rather than letting corporate and national competition compromise public interest.
- **Global Public Good:** Advanced AI, especially AGI, must be managed as a global public good, not solely left to market forces or geopolitical rivalry.
- **Technological and Governance Guardrails:** It is essential to develop both technical solutions (like non-agentic AI) and robust global governance frameworks, international norms, and transparency measures to ensure AI serves all of humanity.

A discussion point was whether **open-source AI** could be a crucial factor in fostering inclusiveness and rebuilding trust, particularly given concerns that AI development currently benefits only a select few countries, communities, or individuals. H.E. Shan Zhongde, Vice Minister, Ministry of Industry and Information Technology, People's Republic of China, said that open source as a collaborative platform belongs to the global community and is a great power for the world and China.

George Papandreou brought in an interesting proposal into the debate: creating a "fourth branch of government" – a **deliberative branch**, supported by AI. This would enable citizens to participate in discussions on laws and policies, fostering real debate and consensus-building, moving away from the false sense of empowerment offered by current social platforms.

Figure 11: H.E. Mr. Shan Zhongde, Vice Minister, Ministry of Industry and Information Technology, People's Republic of China; Anne Bouverot, Special Envoy of France for AI

Quotes:

- *"If you just look at the model cards for all the prominent model providers over the last three to four years and just analyze the amount of information that used to be present in them and the amount of information that is present in them today for when they come out, we've actually seen a regression." (Udbhav Tiwari, VP Strategy and Global Affairs, Signal)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 6.4 Open Source and Open Weight AI: Trajectories, Debates, and Global Practices

2.4 Bridging Inclusion

The principle of inclusion goes beyond access – it is about meaningful participation in governance. Many participants warned of the risk that the Global South, and especially indigenous and rural communities, are sidelined in shaping AI standards and priorities. If global governance frameworks do not intentionally amplify these voices, AI risks entrenching historical inequities.

H.E. Ms. Tatenda Annastacia Mavetera (Minister of Information Communication Technology, Postal and Courier Services, **Zimbabwe**) emphasized a national philosophy of "leaving no one and no place behind" when bridging the AI divide, particularly between urban and rural areas. Zimbabwe's approach, encapsulated by "PSL & I^2," focuses on **Participation and Privacy, Skills (upskilling/reskilling), Leadership (political will and sound policies) and Legislation (AI framework and strategy)**, along with developing local AI solutions. [Editor: Ms Mavetera was not explicit about the acronym I^2, but it may stand for Infrastructure and Innovation (through local solutions)].

H.E. Mr. Hubert Vargas Picado (Vice Minister, Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones, **Costa Rica)** framed bridging the AI divide as a matter of **equity** and closing structural development gaps. Despite 85% connectivity, AI use is much lower. Costa Rica's national AI strategy, launched 10 months ago and co-created with over 50 institutions and aligned with OECD's AI principles, prioritizes digital inclusion for rural communities, indigenous people, and youth through programs like community innovation labs and smart community centers that provide basic AI literacy. Mr Vargas Picado hoped for a more **multilingual AI** with reduced biases (e.g., gender) and greater global integration and cooperation in decision-making.

Inclusion also means linguistic representation: most large language models are still trained predominantly in English and a few major languages, leaving thousands of languages invisible. Participants described this as a form of cultural exclusion. Building multilingual AI systems was described as a priority for ensuring that AI reflects, rather than erases, human diversity.

Jūratė Šovienė (Chair of the Council, The Communications Regulatory Authority of the Republic of **Lithuania**) identified two prerequisites for bridging the AI divide: **digital infrastructure** (strengthening resilience, higher quality services, and fiber to every household by 2030 through a holistic approach integrating fiber with road and electricity networks) and **digital skills**. She categorized people into "**digital natives**," "**digital immigrants**" (who find digital tools interesting), and a large group of "**digital refugees**" (mostly aging populations who find digital tools scary and untrustworthy). She advocated seeing this gap as an opportunity to discover new citizens, customers, and employees, urging combined forces to ensure no one is excluded from digital life.

Emmy Lou Versoza-Delfin (Director, Department of Information and Communications Technology – Philippines) outlined the Philippines' inclusive approach for its more than

7'000 islands: prioritizing **robust connectivity** through national broadband and free public Wi-Fi programs in public spaces, public schools, health centers etc. – 73% of the Philippines are connected to the Internet.

Figure 12: Emmy Lou Versoza-Delfin, Director, Department of Information and Communications Technology, Philippines; Lacina Koné, Director General and Chief Executive Officer, Smart Africa

Ms Šovienė (Lithuania) highlighted the critical role of **language in AI development** and that it is the responsibility of smaller nations to actively contribute data in their own languages, as this will ultimately shape the language of AI. She also envisioned a future where seniors in small towns could easily find assistance with digital and AI tools, ensuring broad inclusion.

Quotes:

- *"How these [AI] technologies fail in different parts of the world is a really attenuated concern and it's not a concern that can be answered by one group of people with one distinct kind of worldview." (Rachel Adams, Director, African Observatory on Responsible AI)*
- *"I think we cannot deal with these complex issues if we don't bring in the collective wisdom of our societies and the agency of our societies. So why not create a 4th branch of government, a deliberative [branch], using AI […] where citizens can deliberate on all the laws and policies which we are discussing as politicians, where they have a voice. Where algorithms allow for real debate, not bullying, not polarization, but real debate, but also consensus building. Where everybody has a voice but one voice," [George Papandreou]*
- *"We are prioritizing digital inclusion for rural communities, indigenous people …, and youth specifically." (Hubert Vargas Picado, Vice Minister, Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones, Costa Rica)*
- *"The Philippines is composed of more than 7000 islands ... it's about bringing opportunities, digital literacy, connectivity to people outside of the [metropolitan areas]." (Emmy Lou Versoza-Delfin, Director, Department of Information and Communications Technology, Philippines)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 2.2 Steps towards addressing the AI Divide

2.5 Capacity for All, Not Just a Few

Even where there is political will, many countries lack the capacity to implement robust AI governance. Capacity gaps exist in skills, institutions, infrastructure, and financing. The discussions highlighted that without intentional strategies to close these gaps, developing countries will be locked into a position of dependency, consumers of AI shaped elsewhere.

Proposals ranged from global funds for compute access to regional training hubs and toolkits tailored for small states. Several participants argued for shared infrastructure models – "AI commons" – that pool resources across borders to democratize access. The message was clear: capacity-building is not charity but a precondition for equitable governance.

Emmy Lou Versoza-Delfin mentioned the **Philippine Skills Framework for AI Analytics** crafted by industry, and **Digital Transformation Centers** to provide free equipment and connectivity for marginalized sectors.

Figure 13: H.E. Mr. Hubert Vargas Picado, Vice Minister, Ministerio de Ciencia, Innovación, Tecnología y Telecomunicaciones, Costa Rica; H.E. Dr. Tatenda Annastacia Mavetera, Minister, Ministry of Information Communication Technology, Postal and Courier Services, Zimbabwe

Hubert Vargas Picado (Costa Rica) described initiatives which Costa Rica launched to close its innovation gap. A **public sector AI training program** bridges the gap with the private sector. Costa Rica has also opened **laboratories for SMEs** (e.g., 5G applications, AgriBoost for AI in coffee farming) and developed an OECD AI toolkit to share knowledge with neighboring Central American and Caribbean countries, shifting from aid recipients to knowledge providers.

Quote:

- *"We believe that inclusive international cooperation is the key to ensuring that no country or community is left behind." (H.E. Engineer Majed Al Mesmar, Director-General of the Telecommunications and Digital Government Regulatory Authority of the United Arab Emirates)*

2.6 Environmental Sustainability and AI Infrastructure

The environmental impact of AI is becoming more pronounced. **Training frontier model**s as well as **inference** (i.e., the stage where an AI system takes new input like a question or an image and produces an output, such as a chatbot reply) consumes significant amounts of energy and water, with data centers increasingly straining local grids. Participants stressed that AI cannot be divorced from the climate agenda.

Figure 14: Gabriela Ramos, then-Mexico (Assistant Director-General for Social and Juman Sciences, UNESCO)

Sustainability, it was argued, should be considered a core principle of responsible AI, not a side issue. Many of the regions most vulnerable to climate change are also least responsible for AI's environmental burden.

Opportunities exist in aligning AI governance with the climate agenda. Initiatives such as the Coalition for Sustainable AI, **standards for green data centers**, and **reporting requirements for energy use** were highlighted as steps toward ensuring that AI's growth is environmentally responsible.

AI can also play a role in addressing climate challenges, for example by optimizing energy grids, modeling environmental risks, and supporting climate research.

Quote:

- *"AI governance must include sustainability through energy efficiency, through green infrastructure, through policies that align technological progress with climate responsibility." (Anne Bouverot, Special Envoy of France for AI)"*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 5.2 Energy and Sustainability

2.7 Sectoral Focus and Broad Collaboration

AI is not a single technology but a **family of systems** with sector-specific impacts. Governance must therefore be sensitive to context. In healthcare, safety and explainability are paramount. In education, equity of access is key. In agriculture, resilience and food security are priorities.

Examples shared during the Dialogue illustrated how **sectoral governance** can work in practice: AI-assisted dementia screening projects, registries of AI use in public administration, and agricultural platforms monitoring soil and water conditions. Participants stressed that sectoral governance should not fragment into silos, but instead serve as testbeds for principles that can then be scaled globally.

Quote:

- *"By today, Estonia has implemented approximately 200 AI applications across government institutions – in fields as diverse as education, healthcare, justice, transport, the environment, and culture." (H.E. Mr. Alar Karis, President of the Republic of Estonia)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 3.3 Regional AI partnerships
- Annex – Examples of multilateral initiatives [a list of some 40 initiatives]
- Annex – Examples of national initiatives [a list of some 20 initiatives]

2.8 Standards and Safety Tools

Standards were identified as one of the pressing needs for global governance. Without **common benchmarks** for testing, evaluating, and registering AI systems, efforts risk fragmentation. There was strong support for international cooperation on standards, especially for safety testing of frontier models.

It was also said that, historically, developing countries have been largely excluded from standard-setting processes, leading to barriers for smaller market players. Multistakeholder and multidisciplinary approaches to testing that incorporate diverse perspectives and local realities, especially regarding how technologies fail in different parts of the world, will be useful.

The need for **greater harmonization in standardization efforts** to reduce fragmentation and to maintain flexibility to keep pace with rapid AI advancements was pointed out. One needs to double down on benchmarking efforts to test AI's robustness, safety, and fairness, especially for industry-specific applications (healthcare, education, defense, etc.) and emerging agentic AI solutions with greater autonomy.

Risk management systems for frontier AI need to be standardized, emphasized Chris Meserole (Executive Director, Frontier Model Forum), and called for global coordination to define what

a robust framework should look like. While there are many voluntary activities, the process for creating formal standards needs to be much faster.

Current industry testing, particularly for "**frontier safety policies,**" often involves checking for dangerous capabilities (e.g., assisting with weapon creation or cyberattacks) that developers hope not to see, rather than verifying safety or reliability – a critical distinction for policymakers to grasp. These practices are currently voluntary and lack standardized coordination across the industry.

A competitive environment can lead to companies not prioritizing safety. Furthermore, different countries have varied approaches to AI safety, making global collaboration difficult. Continued investment in **R&D to define "red lines"** has been called for by Ya-Qin Zhang (Chair Professor, Tsinghua University), and Brian Tse (CEO, Concordia AI) proposed an **international effort to define "red lines"** for unacceptable AI outcomes.

Participants stressed the need for proactive mechanisms to detect and mitigate risks before they spiral out of control. Current testing and evaluation regimes are often inadequate, focused on short-term performance rather than long-term systemic risks.

Proposals included mandatory pre-deployment testing of high-risk systems, red-teaming exercises to identify vulnerabilities, and the creation of international early warning systems for frontier models. "**Safety by design**" (Ya-Qin Zhang) was described as essential: building safeguards into AI systems from the outset rather than attempting to patch them after deployment. National AI Safety Institutes, which monitor and publish results to work constructively with industry, play a positive role.

This proactive approach was seen not as a brake on innovation but as a foundation for trust. If societies can be confident that risks are being anticipated and managed, the opportunities of AI can be embraced more fully.

While some risks can be managed with existing tools, **a class of risks that can appear quickly and at "extreme scale"** – like bio and advanced cyber threats – requires entirely new risk management instruments to identify issues in advance (Chris Meserole).

Boulbaba Ben Amor (Director for AI for Good at Inception, a G42 company) urged policymakers to shift focus from evaluating core AI models to **evaluating full AI products and solutions**, as these are what end-users and society interact with. He emphasized the need for **product conformity** and adaptable **risk assessments**, advocating for the inclusion of diverse cultures, languages, and fields in policies.

H.E. Shan Zhongde, Vice Minister, Ministry of Industry and Information Technology, People's Republic of China, emphasized the importance of an open and inclusive approach to constructing an international standards framework for open source.

Figure 15: (from left) Sasha Rubel, Head of Public Policy for Generative AI, Amazon Web Services (AWS); Chris Meserole, Executive Director, Frontier Model Forum; Juha Heikkilä, Adviser for Artificial Intelligence, European Commission; Udbhav Tiwari, VP Strategy and Global Affairs, Signal; Ya-Qin Zhang, Chair Professor, Tsinghua University; Brian Tse, CEO, Concordia AI

Among **key standards** which would move AI governance forward, the following were mentioned:

- A global norm **where frontier AI firms must publish a risk management framework** and provide systematic updates on how they are implementing it.
- An **international effort to define "red lines"** for unacceptable AI outcomes, emphasizing that the definition of what is acceptable and safe should not be left to the industry alone.
- **Streamlining of AI initiatives** to curb the proliferation of monitoring and reporting requirements which become a burden on developers.
- **Registration and identification of AI-generated content, models and agents**.
- **Internationalizing Best Practices:** There is tremendous potential for collecting and synthesizing best practices from companies and industries into global standards. AI governance and safety should be a "safe zone" for cooperation, transcending geopolitical differences, as it is a matter for humanity as a whole.

AI agents will be operating across borders, posing a significant global governance problem. Robert Trager (University of Oxford) highlighted technical directions to address this:

1. **Verification:** This involves interrogating AI systems to understand their properties and recent actions. A key challenge is performing this verification at the compute provider level, as these providers could be globally distributed.
2. **Benchmarking**: "The secret of AI governance is benchmarking, benchmarking, and benchmarking," said Professor Trager, and stressed the need for standardized metrics to evaluate AI systems, similar to "0 to 60 miles per hour" metrics for cars. He also pointed out a dual challenge: the need for both **global benchmarks** (for universal standards) and l**ocal benchmarks** (to ensure AI conforms to local laws and norms).

Yoshua Bengio argued that since we can't simply make AIs "dumb," the solution lies in controlling their **harmful intentions**. He proposes a new approach: **Non-agentic AI systems**, which he calls "Scientist AI":

- **No Intentions or Goals:** Unlike current AIs that imitate human behavior (which can include deception), these systems are designed to have **no intentions, goals, or drives for self-preservation**. They act like idealized scientists, focusing solely on **understanding and explaining the world**, generating theories and predictions without a desire to act or pursue objectives.
- **Safeguard Mechanism:** In the short term, "Scientist AI" could serve as a **critical safety layer**. By monitoring the actions of more powerful, agentic AIs, they could predict the probability of harm from a proposed action. If this probability exceeds a set threshold, the "Scientist AI" could halt the action, acting as a **gatekeeper** to prevent dangerous outcomes.
- **Support for Research:** Beyond safety, these non-agentic systems could also **support general scientific discovery** by objectively looking for explanations and generating hypotheses.

Rachel Adams (Director, African Observatory on Responsible AI) introduced the aspect of **public perception and social attitude surveys** to "test for trust" in AI, highlighting that people's trust in technology might differ significantly from technical trustworthiness.

Quotes:

- *The biggest hindrance for deploying AI is how we can actually ensure the safety and security of AI. (Dawn Song, Professor of Computer Science at UC Berkeley and Director of Berkeley center for responsible decentralized intelligence)*
- *"I want to hear [a] call for action: benchmarking, benchmarking, benchmarking." (Boulbaba Ben Amor, Director for AI for Good at Inception, a G42 company)*
- *"While the major effort is turned to evaluate and benchmark AI models, we need to move as quickly as possible to evaluate AI solutions." (Boulbaba Ben Amor, Director for AI for Good at Inception, a G42 company)*
- *"... There's a class of risk that, if it appears, it may appear very quickly … in a way that the harms don't materialize until it's too late to do anything about them, in which case you need a risk management instrument that's able to identify those issues in advance ..." (Chris Meserole, Executive Director, Frontier Model Forum)*
- *"If a system is smarter than you, more complex, more creative, it's capable of doing something you didn't anticipate. So we don't know how to test for bugs we haven't seen before." (Roman V. Yampolskiy, Professor, Department of Computer Science and Engineering, University of Louisville)*
- *"We need to continue to invest in R&D... to come up with the red lines, the benchmarks, the thresholds, and the warning system. [There is] a five-stage process: testing, auditing, verification, monitoring, and mitigation." (Ya-Qin Zhang, Chair Professor, Tsinghua University)*

- *"One bucket [of problems is] where we know what the problem is (scams, deep fakes, non-consensual sexual imagery, illegal uses of AI.) … [for these] we need clear rules: What is the line? And who is responsible for holding it? … The second bucket [of problems] are problems that have uncertainty, like is AI going to kill us all …. you don't have a clear solution. … There you need to create the bodies that work with industry over time to share information, reduce informational asymmetries to iterative policy. That doesn't hamper innovation, too, but also doesn't catch the problem too late." (Artemis Seaford, Head of AI Safety, ElevenLabs)*

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 4.1: Landscape of AI Standard Setting Initiatives
- 4.2 Technical Standards Development
- 4.3 Ethical AI Frameworks
- 4.4 Safety Standards and Red-Teaming
- 4.5 Certification and Accreditation Programs
- 6.1 Risks of AI and Systems Safety Assessment
- 6.2 Approaches to Mitigating AI Risks
- 6.3 Corporate Risk Mitigation Practices and its Limitations
- 6.4 Open Source and Open Weight AI: Trajectories, Debates, and Global Practices
- 6.5 AGI, Existential Riska, and Social Resilience
- 6.6 Verification as a path to reduce risks from AI

2.9 Governance of Compute and Models

"Compute" – short for "computing power" – is the essential resource for both training and operating AI models. It is a key lever for AI governance because it is a measurable and quantifiable bottleneck that defines who participates in AI innovation. While training advanced AI models requires a massive amount of compute over months, the majority of compute resources are actually used for operating and deploying models due to the millions of daily user requests.

Because access to the most advanced chips and clusters is concentrated in a few countries and companies, governing compute is seen as an effective way to manage risks.

"**Compute governance**", i.e., the rules and policies for overseeing access to and use of advanced computing resources, can be done by monitoring who has access to the most capable hardware, setting oversight thresholds for large training runs, or requiring safety measures.

Proposals included licensing regimes for large-scale training runs, registries of high-risk systems, and shared compute initiatives that democratize access.

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 5.1 Global Compute Distribution

2.10 Policy Interoperability and Agile Governance

Policy coherence was another recurring theme. Today, national AI strategies vary widely, from light-touch frameworks designed to encourage innovation to legally binding regimes like the EU's AI Act. Without coordination, these divergences risk creating a patchwork of incompatible rules that increase compliance costs and weaken governance.

Policy interoperability could involve, for example:

- **Aligning definitions, standards, and risk categories** so companies don't face contradictory requirements.
- **Ensuring cross-border data flows and AI systems** can function under multiple jurisdictions.
- Creating mechanisms for **mutual recognition** or coordination of regulatory approaches.

Interoperability was framed as the glue of global governance. As AI systems, data, and models cross borders seamlessly, governance frameworks must do the same. Interoperability does not mean identical rules everywhere, but mutual recognition and harmonization that allow systems to operate across jurisdictions without undermining safety or rights.

Without policy interoperability, global cooperation will falter, and companies will face incompatible requirements. With it, however, there is potential for a governance framework that is both global and adaptable, capable of keeping pace with technological change.

Due to the complexity of AI governance, Juha Heikkilä (Adviser for Artificial Intelligence, European Commission) highlighted the **need for follow-up and update mechanisms**. He argued that to build trust, it is crucial to know how effectively recommendations and regulations are being put into practice. This is necessary for both formal legislation and "softer" governance approaches.

Dive deeper in the Whitepaper "Themes and Trends in AI Governance":

- 3.1 Global AI Summits
- 3.2 Track to Diplomacy Initiatives
- 3.3 Regional AI Partnerships
- Annex – Examples of multilateral initiatives [a list of some 40 initiatives]
- Annex – Examples of national initiatives [a list of some 20 initiatives]

Chapter 3: Regional Perspectives and Case Studies

The Dialogue featured several concrete examples of how countries and regions are beginning to address the governance challenge. These case studies illustrate both diversity of approach and potential for cross-learning.

3.1 European Union

Europe's approach to AI governance is characterized by a risk-based approach, distinguishing between unacceptable, high, and low-risk systems. The **European Union's AI Act** was repeatedly referenced as the first comprehensive, binding framework for AI. By adopting a risk-based approach, the Act seeks to balance innovation with safety and trust, providing a model that others may adapt.

In Europe, **trust** is seen as the foundation of AI adoption. The EU sees hard law as a necessary way to build that trust. Transparency measures such as registries – Estonia's AI registry was cited as a pioneering national initiative, offering citizens and policymakers visibility into how AI is used in public administration –, model documentation, a mechanism for quick updates, and a just published (10 July 2025) code of practice were highlighted as mechanisms for building this trust. At the same time, Europe has invested in innovation sandboxes, allowing developers and regulators to collaborate in testing new systems before large-scale deployment.

The purpose of the **General-Purpose AI Code of Practice,** a voluntary guide developed by independent experts through a multi-stakeholder process, is to help companies meet the **European AI Act's legal requirements** concerning the **safety, transparency, and copyright of general-purpose AI models**. Both European Member States and the European Commission are now evaluating the code.

(continued)

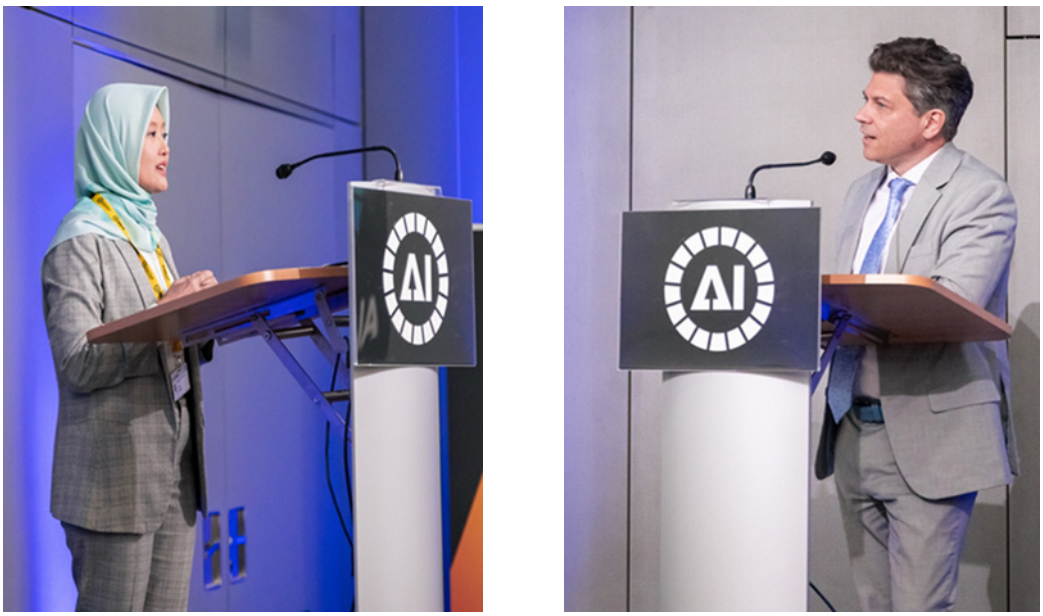


Figure 16: (from top left to right) Speaking at the luncheon: H.E. Mr Alar Karis, H.E. Mr Alar Karis, President, Republic of Estonia; Amandeep Singh Gill, Under-Secretary-General and Special Envoy for Digital and Emerging Technologies, Office for Digital and Emerging Technologies, United Nations; H.E. Ms. Rahayu Mahzam, Minister of State, Ministry of Digital Development and Information, and Minister of State at the Ministry of Health, Singapore; Robert Trager, Co-director, Oxford Martin AI Governance Institute, University of Oxford

3.2 Africa

African participants stressed that inclusion is not optional but existential. The continent risks being sidelined in global AI development due to limited compute infrastructure, talent concentration elsewhere, and linguistic underrepresentation. Yet African voices at the Dialogue rejected the notion of passive dependency. Instead, they emphasized sovereignty and self-determination, calling for homegrown solutions and regional collaboration.

Africa has an opportunity to reposition itself, given its projected population growth, which represents a significant source of data for big tech. To capitalize on this, African nations must **establish robust data frameworks** to attract partnerships with major technology companies. Africa is not concerned with the origin of the technology, but rather its ability to **improve the lives of its people**. This requires collaboration between big tech, platforms like the ITU and UN, and African governments to ensure **equity in AI development and deployment**. African countries need to prioritize **information technology budgets and investments in compute infrastructure (like GPUs) within Africa**. Africa does have proven track record of innovation, particularly in financial technology and inclusion. The establishment of AI compute infrastructure in Africa can help the continent leapfrog the digital divide.

One of the pressing issues raised was the **linguistic divide**. With more than 2,000 languages spoken across the continent, the dominance of English and other global languages in large models threatens to marginalize African cultures. **Africa's development of sovereign large**

language models (LLMs) was discussed as a response to both linguistic divides and power concentration. By training models on African languages and contexts, these initiatives aim to ensure that AI reflects the continent's diversity and serves local needs, rather than being a one-size-fits-all import from abroad.

Capacity was another priority. Calls were made for investment in skills development, regional compute centers, and funding mechanisms that would allow African innovators to build AI solutions tailored to local needs – from agriculture and health to governance and financial services.

One practical step African countries could take together to boost AI access and innovation across the continent would be to collaborate across nations to **create a single digital market for AI access, high-power services, data clusters, and de-risking algorithms** under proper AI governance.

Quotes:

- *"It's our responsibility in Africa to actually train large language models in our language otherwise we'll be basically writing in someone else's future, which is not ours." (Lacina Koné, Director General and Chief Executive Officer, Smart Africa)*
- *"We're looking at AI local solutions … Let's customize it, let's make it relevant to Zimbabwe." (H.E. Tatenda Annastacia Mavetera, Minister of Information Communication Technology, Postal and Courier Services, Zimbabwe)*

3.3 Asia

Asia presented a diversity of approaches, reflecting the continent's scale and complexity. Singapore's representatives described their **balanced innovation model**, which combines careful attention to specific high-risk applications with a light-touch regulatory approach elsewhere. This allows innovation to flourish while safeguarding public trust in sensitive domains such as healthcare and finance.

Regional cooperation is also advancing. **ASEAN** was mentioned as developing shared AI principles to prevent fragmentation among its member states. This effort demonstrates how regional bodies can provide coherence while respecting national diversity.

China's perspective focused on **AI for inclusive development**, emphasizing people-centered approaches, open standards, and digital infrastructure as foundations for equitable growth. Other Asian voices echoed the importance of ensuring that AI is not only a tool for economic competitiveness but also for social good.

3.4 The Americas

The Americas brought a dual focus: **innovation and rights**. U.S. representatives highlighted the **risks of authoritarian misuse of AI, particularly in surveillance and information manipulation**. They stressed the need to safeguard democratic values while supporting private-sector innovation. The emphasis was on building trust through both technical standards and policy commitments.

From **Latin America**, participants emphasized equity and access. Countries in the region are experimenting with AI applications in public services, education, and agriculture, but face challenges in compute infrastructure and skills development. Calls were made for greater inclusion in global standard-setting processes, as well as investment in regional capacity.

Figure 17: Group photo of the Luncheon

3.5 Estonia

H.E. Mr. Alar Karis, President of the Republic of Estonia, laid out how his country of just 1.3 million people has become one of the world's most digitally advanced societies over the last two decades. For his verbatim speech, please see the Annex. Here is a **succinct best practices list** distilled from Estonia's experience as a pioneer in e-governance and now leading implementer of AI:

1. Create a Vision
    - Don't simply digitize and replicate existing services but rethink and rebuild the whole system.
    - Have a human-centric mindset.
2. Build Strong Digital Infrastructure
    - Create secure digital identity systems.
    - Develop interoperable data exchange platforms.
    - Provide a digital payment system.
    - Ensure trust, security, and resilience are built into the foundations.

3. Deploy AI Across Public Services

    - Integrate AI into multiple sectors (education, healthcare, justice, transport, environment, culture) – today, Estonia has about 200 AI applications across government institutions.
    - Use AI for practical citizen services (e.g., virtual assistants, autonomous museum buses, flood detection, AI learning tools).
    - Scale beyond pilots – embed AI widely in government operations.

4. Adopt a Human-Centric Approach

    - Design digital and AI services for citizen benefit first, not just government efficiency.
    - Ensure citizens can see and control how their data is used.
    - Maintain transparency tools (e.g., Algorithm Registry).

5. Balance Innovation with Responsibility

    - Support risk-based regulation (e.g., EU AI Act model).
    - Provide AI sandboxes, training, and competence centers to help stakeholders comply and innovate.
    - Safeguard digital rights and online freedoms alongside AI adoption.

6. Prioritize Education and AI Literacy

    - Integrate AI learning tools into schools and curricula.
    - Train both students and teachers in responsible AI use.
    - Build societal resilience by fostering AI literacy early.

7. Commit to International Cooperation

    - Share expertise and support other countries in building AI-ready societies.
    - Participate in global AI governance initiatives (e.g., Global Digital Compact, Freedom Online Coalition).
    - Promote inclusive AI ecosystems to reduce global digital divides.

Quote:

- *"The heart of Estonia's previous digital transformation was the understanding that we should not simply digitalise existing services but rather rethink and rebuild the whole system." (H.E. Mr. Alar Karis, President of the Republic of Estonia)*

3.6 Switzerland

Switzerland, which sees itself as a leader in international AI cooperation and a bridge builder for global collaboration, introduced two key Swiss initiatives designed to democratize access to AI:

- **The Swiss-made Large Language Model:** The model will be released under an open license, with its source code, weights and training data all publicly available. A core feature is its multilingual design – the model has been trained on datasets from over 1000 languages. The model is a flagship of the Swiss National AI Institute (SNAI), a collaboration between researchers at the Swiss universities ETH and EPFL. The goal is to create a legally compliant, accessible model that can serve as a foundation for various applications in areas like healthcare, sustainability, and science. The model is being

developed with a focus on underrepresented languages and will be released later in the year in different sizes to make it widely usable.

- **ICAIN (International Computing and AI Network):** Described as a "CERN for AI," this initiative aims to create a decentralized, international network for AI resources. It brings together partners from around the world to pool resources like large-scale compute power, data, and talent. The primary purpose of ICAIN is to drive high-impact projects in areas such as climate, agriculture, health, and humanitarian aid by fostering collaboration and preventing competition among researchers. An example of this is a partnership with AI Singapore to use the Swiss-made LLM to develop inclusive AI for Southeast Asia. Switzerland's ICAIN initiative was highlighted as an **innovative response to the concentration of compute power.** By creating shared infrastructure accessible to multiple stakeholders, ICAIN aims to democratize access to frontier AI capabilities while embedding oversight and accountability.

Figure 18: (from left) Katharina Frey, Deputy Head Digitalisation Division, FDFA, future ICAIN Executive Director, ETH Zurich; Bernard Maissen, State Secretary, Director, Federal Office of Communications OFCOM; Mennatallah El-Assady, Assistant Professor at the Department of Computer Science, ETH Zurich; Mary-Anne Hartley, Professor | Director, Laboratory for Intelligent Global Health & Humanitarian Response Technologies (LiGHT)

The following use cases based on the above initiatives were explained:

- **Sovereign AI and Decentralization:** With "sovereign ability" in AI, countries would own and adapt foundational models for their specific needs. Two practical applications of this principle were mentioned, Meditron and Legitron.
    - **Meditron:** A pipeline designed to "medicalize" any base language model. It is paired with a platform called MOOVE (Massive Open Online Validation and Evaluation platform), which allows doctors worldwide to test, validate, and customize the model for their specific cultural and contextual needs.

  - **Legitron:** A similar model developed in collaboration with the ICRC (International Committee for the Red Cross) that focuses on international humanitarian law (IHL), allowing lawyers to explore how AI can understand and be used to interact with this body of law.

- **Education and Impact:** An example of ICAIN's educational pillar described a course at ETH Zurich that uses a human-centered approach to solve real-world challenges in climate, peace, and health. The course brings together interdisciplinary teams of students to develop AI applications. A new collaborative effort with Data Science Africa to create an educational card game that teaches computational thinking and trustworthy AI principles was also announced.

Quotes:

- *"Switzerland aims to set [a] course for AI [by prioritizing] quality over scale, public-private collaboration, and ethical AI development." (Bernard Maissen, State Secretary, Director, Federal Office of Communications (OFCOM))*
- *"… the idea [of the Swiss LLM] is to allow this to be … accessible to a lot of different frontiers and a lot of different applications." (Mennatallah El-Assady, Assistant Professor at the Department of Computer Science, ETH Zurich)*
- *"[Sovereign AI] is how you can make these kinds of models your own and own them yourself. And that sovereign ability is really the spirit of [the] decentralization we are trying to achieve." (Annie Hartley, Professor | Director, Laboratory for Intelligent Global Health & Humanitarian Response Technologies (LiGHT))*

3.7 Singapore

Singapore approaches to AI are based on the two core beliefs that AI can elevate societal and economic potential, and that it must serve the public good. As an example of AI for the public good, Singapore's **Project Pensive** was mentioned for early dementia detection and at scale, based on how people draw.

Singapore's take is that good governance should be carried out in a way that enables rather than impedes innovation, while at the same time being equally concerned about managing AI's risks. Domestically, Singapore is investing in **foundational research** to understand AI risks better, while Singapore's AI Safety Institute collaborates which other AI safety institutes in the world. Furthermore, Singapore has been developing comprehensive **governance frameworks and tools** (such as AI Verify and Project Moon Shot). Thirdly, **laws and regulations** are continuously being reviewed by taking a more surgical approach to address specific AI-related harms such as AI-faciliated online crimes and to safeguard the integrity of Singapore's elections against malicious AI-generated deep fakes.

Recognizing that AI governance is a collective challenge, Singapore just hosted over 100 AI experts in Singapore who agreed on the **Singapore Consensus on AI Safety Research Priorities**. Singapore chairs the UN Digital Forum of Small States which launched the **AI Playbook for Small States** at UN's Summit of the Future in 2024, and by playing a constructive role in ASEAN, such as leading the endorsement of the **ASEAN Guide on AI Governance and Ethics** in 2024. This commitment ensures all countries, particularly smaller ones, have a voice in shaping AI's future.

Singapore believes that AI must serve humanity, and achieving shared prosperity and sustainable development requires a united, responsible approach to its powerful capabilities.

3.8 Saudi Arabia

To combat the compute divide, the data divide, and the algorithmic divide, the Kingdom of Saudi Arabia has leveraged its resources, including a partnership with Aramco, to deliver an **inference node** serving 70% of Asia, Europe, and Africa at low cost, addressing the compute divide. For the algorithmic divide, Saudi Arabia has developed **ALLAM, the most powerful Arabic Large Language Model,** and are leading efforts to combat hallucination and biases. To tackle the data divide and ensure inclusivity, a **public consultation for a global AI Hub law** has been launched, aiming to provide a safe harbor for innovators worldwide to leverage the region's compute, algorithmic, and data power.

Saudi Arabia can also celebrate past successes, such as **36% women's empowerment in tech** in their region, and recognizing key female leaders in digital cooperation and AI dialogue. Resources have been committed, including King Abdullah University of Science and Technology (KAUST), for AI safety research.

Quote:

- "We managed to celebrate the highest women empowerment percentage in tech and digital, 36%, surpassing the Silicon Valley average." (H.E. Abdullah Amer Alswaha, Minister, Ministry of Communications & Information Technology, Saudi Arabia)

Chapter 4: A vision for 2026

4.1 The UN's "Global Dialogue on Artificial Intelligence Governance"

In August 2025, the United Nations General Assembly approved the Resolution "Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance." This resolution establishes two new mechanisms to strengthen international cooperation and governance of AI. It represents a significant step forward in the UN's efforts to ensure the safe, secure, and trustworthy development of AI for sustainable development and to bridge the digital divide.

Clause 6 of the Resolution says:

> *"6. Also decides to organize a high-level multi-stakeholder informal meeting to launch the Dialogue in the margins of the high-level week of the eightieth session of the General Assembly, in 2025, chaired by the President of the General Assembly, and further decides that the Global Dialogue on Artificial Intelligence Governance will initially be held back-to-back in the margins of the International Telecommunication Union Artificial Intelligence for Good Global Summit in Geneva, in 2026, and of the multi-stakeholder forum on science, technology and innovation for the Sustainable Development Goals in New York, in 2027;"*

Figure 19: Anne Bouverot, Special Envoy of France for AI; LJ Rich, Moderator of the AI for Good Global Summit; H. E. Engineer Majed Sultan Al Mesmar, Director General of UAE's Telecommunications and Digital Government Regulatory Authority (TDRA); Tomas Lamanauskas, Deputy Secretary-General, ITU

The following co-chairs' summary of the AI Governance Dialogue held in Geneva on 10 July 2025 – "Ten Pillars for AI Governance" – could provide a valuable input for the UN's Global Dialogue on Artificial Intelligence Governance at its inaugural meeting in the margins of ITU's AI for Good Global Summit in Geneva in 2026. As Tomas Lamanauskas, Deputy Secretary-General of the ITU, said at the closing of the AI Governance Dialogue on 10 July 2025, there is a "huge hunger" for inclusive AI governance, encompassing access to skills, infrastructure, and a "seat at the table" in critical discussions.

4.2 Co-chairs' Ten Pillars for AI Governance

1.	**From Principles to Practice** AI governance should move beyond high-level declarations to practical implementation that enables sustainable innovation and long-term impact. Agile and inclusive frameworks, adaptable oversight mechanisms, technical standards and tools are essential to guide AI development and deployment in ways that are socially, economically, and environmentally responsible.
2.	**A Multistakeholder Imperative** Governance that affects all should be shaped by all. Governments, civil society, academia, the private sector, technical experts, and international organizations should co-create policy. All countries need a seat at the table, supported by capacity building, so that AI can benefit everyone, everywhere.
3.	**Transparency as a Cornerstone of Trust** Understanding how AI systems are built, evaluated, and used is important. Transparency in model behavior, data practices, and decision-making processes strengthens accountability, builds public confidence, and unleashes responsible innovation.
4.	**Bridging Inclusion** AI governance should reflect diverse perspectives. Bridging the digital divide through inclusion goes beyond access—it means enabling meaningful participation in shaping the technologies and rules that affect people's lives.
5.	**Capacity for All, Not Just a Few** Closing global gaps in AI readiness is critical. Capacity-building initiatives—spanning policy advice, skills training, institutional strengthening, and financial support—should empower communities worldwide to govern AI effectively and to innovate in key sectors such as health, education, and agriculture.
6.	**Environmental Sustainability and AI Infrastructure** Sustainable AI development should address its environmental footprint—energy, water, and resource demands. Governance frameworks should integrate energy and environmental policies, promote efficient data centers and renewable power, and ensure AI projects can scale without overstraining local infrastructure.
7.	**Sectoral Focus and Broad Collaboration** The value of AI is realized through its applications in health, education, agriculture, humanitarian assistance including in disaster management, and many other critical areas. Governance should involve respective communities, adopt a cross-government and cross-society approach, and leverage international frameworks so that AI can deliver targeted benefits and address sector-specific challenges.
8.	**Standards and Safety Tools** Technical standards, benchmarks, and audit protocols are foundational for safe, interoperable, and agile AI governance. Developed through international multistakeholder processes, these tools should be evidence-based and adaptable to rapid technological evolution.
9.	**Governance of Compute and Models** As AI models scale in capability, governance of compute resources and large foundation models becomes more important. Access to compute infrastructure, robust risk assessments, and accountability frameworks ensure that powerful AI systems serve the public interest.
10.	**Policy Interoperability and Agile Governance** Coherent and interoperable policy frameworks prevent fragmentation while providing clear policy direction. Agile governance – integrating adaptable rules and inclusively developed technical standards – enables flexible adaptation to technological advances.

Annexes

1 AI Governance Dialogue address: Steering the future of AI – Doreen Bogdan-Martin

Your Excellency Mr. Alar Karis, President of Estonia, Excellencies, colleagues, and friends,

Transformation is a word we in the technology world hear often. But when it comes to artificial intelligence, I believe we truly are in a transformative moment. The global AI race is well underway, sparking fierce competition between companies, countries, and region; reshaping diplomacy, economics, and labor markets; and exposing critical gaps in the capacity to develop, deploy and benefit from AI. As this technology evolves amid fears it will overtake human ingenuity… the challenge is not whether to govern AI, but how to ensure governance steers it in the right direction. We know this is possible, because we've already caught a glimpse of it.

Last year, the Nobel Prize for Chemistry was awarded to the developers of AlphaFold, an AI system that predicts the complex 3D structure of proteins. For decades, modeling a single protein took several years. AlphaFold has now modelled virtually every protein known to science: more than 200 million. And those scientists made their work freely available, helping researchers worldwide accelerate drug discovery, transform medicine, and better understand the building blocks of life itself.

Figure 20: Doreen Bogdan-Martin, Secretary-General, ITU

That's the kind of AI future we want: one where AI opens new frontiers of scientific discovery on Earth and in space.

Ladies and gentlemen, this is why we are here. Because this transformative AI moment demands more than admiration or alarm. It demands dialogue and concerted action on inclusive, forward-looking governance that drives innovation and builds public trust. Governance that minimizes risks and leaves no one behind.

Last year, UN Member States adopted the Pact for the Future and Global Digital Compact -complementing guidance offered by the World Summit on Information Society, currently undergoing its 20-year review. These frameworks are our compass for a more equitable, rights-based AI future.

But a compass can't move a ship – it can only point it in the right direction. To steer AI progress towards shared benefits, we need governance mechanisms that are: practical, inclusive, and rooted in real-world implementation. Those governance mechanisms form our captain's wheel.

Here let me thank the captains of today's AI Governance Dialogue: our distinguished Co-Chairs, His Excellency Engineer Majed Al Mesmar, Director-General of the Telecommunications and Digital Government Regulatory Authority of the United Arab Emirates, and Madame Anne Bouverot, France's Special Envoy for Artificial Intelligence.

As we continue today's discussions, I invite you to keep three key elements in mind that I believe can propel AI governance for good forward.

First: **inclusion**. Too many countries – more than 100 – still have no meaningful voice in global AI governance discussions. While it is encouraging to see more of these discussions taking place, from Bletchley Park to Seoul to Paris, and more recently, Kigali, the global reach of the United Nations can help make AI governance as inclusive as it can possibly be. We are proud to welcome participants from 170 countries to this year's Summit. Their perspectives are essential in designing governance mechanisms that truly reflect global realities, not just high-resource contexts, but communities navigating limited infrastructure, low trust, and high stakes. Many governments also lack the resources to engage in – let alone shape their own – AI futures. That must change… which brings me to the second element: capacity.

Capacity is linked to being connected to infrastructure that includes access to compute, data centres, and other infrastructure for artificial intelligence. But capacity is also about people and their ability to make informed decisions. That's why we need to equip policymakers and public administrators – especially in developing countries – with the skills to assess, procure, and deploy AI systems. And it's why ITU and our partners launched the AI Skills Coalition, and why we're working to expand South–South knowledge exchange and regional training hubs.

The third and final element that can steer the AI revolution in the right direction is **standards**: because principles and declarations alone are not enough. We need technical standards that translate high-level commitments into operational safeguards. That's why earlier this week, we held consultations at the Open Dialogue on AI Testing, and a workshop on Trustworthy AI Testing and Validation. These gatherings revealed an urgent need for multistakeholder collaboration in two key areas of action: promoting knowledge exchange on AI standards, and bridging capacity gaps in methodologies for testing AI systems and models.

ITU is ready to continue convening these consultations beyond the AI for Good Summit. Because we cannot leave AI governance to chance. We cannot outsource trust. And we cannot expect countries to implement safeguards they had no role in designing, and that do not fit their local context.

Bringing these three elements together – inclusion, capacity and standards - is what coordinated steering looks like. We saw this in action at today's roundtable luncheon, where participants highlighted the importance of identifying sources of untapped innovation (like FinTech or open-source communities in developing countries) to broaden inclusion, and using policy tools to

deliver in specific areas, from deepfakes to access to compute power to red teaming, always grounded in scientific observation.

One government representative at the Dialogue requested support from peers around the table. It's a powerful reminder that when we bring the right people together, dialogue goes beyond discussion to become a catalyst for real cooperation and concrete action and hope.

Ladies and gentlemen,

Governance is shared, multi-stakeholder responsibility. Everybody in the AI generation has a part to play, and we need all hands on deck!

- Governments can lead in enacting laws, protecting rights, and investing in digital public services, as the Estonian government did in the 90s, when they earmarked 1% of state funding for IT, transforming Estonia into one of the most advanced digital societies in the world.
- Industry also has a role to develop and deploy AI responsibly and to be transparent about high-risk systems
- Academia and the technical community can help evaluate models, stress-test assumptions, and illuminate blind spots.
- Civil society can raise concerns, expose harms, and advocate for communities often left out.
- And the UN system can continue coordinating, convening, and keeping universal values of peace, dignity and human rights at the core as we seek to leverage AI responsibly.

You can see our progress in the latest edition of the **annual UN AI Activities report compiled by ITU** which I'm proud to launch today. The report shows how the UN is integrating AI across our work with 729 projects in 2024, up from 406 in 2023. We're also seeing more engagement on AI across the UN system, with 53 entities contributing to the report.

Ladies and gentlemen,

"The future is here" at the AI for Good Global Summit. But as the saying goes, it is not evenly distributed out there, in the world. This is a transformative moment for AI technology – let's make it transformative for governance, too. Let this be remembered as the point we turned the ship around.

Not when we lost control, but when we took the helm. Not when we raced in competition, but when dialogue helped all boats rise. We don't need to sail in the same ship, or even at the same speed.

But we do need to navigate the same oceans together, by the same compass, under the same stars.

And as my friend the Minister from Ghana said this afternoon: we need not look East, or West.

We need to look forward, together.

With that, ladies and gentlemen, I thank you for being here. And I look forward to our continued discussions.

Quote:

- *"We don't need to sail in the same ship, or even at the same speed. But we do need to navigate the same oceans together, by the same compass, under the same stars." (Doreen Bogdan-Martin, Secretary-General, ITU)*

2 Presidential address, H.E. Mr. Alar Karis, President, Republic of Estonia

Honourable leaders, Excellencies, Distinguished delegates,

It is truly an honour to represent Estonia here today – a country of just 1.3 million people, but one of the world's most digitally advanced societies. For over two decades, we've made technology the backbone of our public services.

Today, Estonia is not just online – it is a full-fledged digital society, where all public services are available to citizens anytime, anywhere.

Our digital journey began with a clear and bold vision. When Estonia regained its independence in 1991, we faced the challenge of building modern state and its institutions from scratch, with very limited resources. That was a time when instead of replicating legacy systems, we decided to leapfrog into the future.

In the mid-1990s, we launched an ambitious national programme called the Tiger Leap, aimed at bringing computers and internet access to every school in the country.

Figure 21: H.E. Mr. Alar Karis, President, Republic of Estonia

This wasn't just about technology – it was about transforming mindsets. This was the foundation on which we started to build our digital-first society and the start-up ecosystem.

Today, we face a new technology with equally transformative power. So we are taking another leap. We want to use the experiences from being a pioneer in e-governance to become a leading implementer of AI.

We are integrating Artificial Intelligence into our public sector and schools – not because it's fashionable, but because we believe AI can improve services and create real value for people – if governed responsibly.

We started this process several years before the popular Large Language Models were released. By today, Estonia has implemented approximately 200 AI applications across government institutions – in fields as diverse as education, healthcare, justice, transport, the environment, and culture.

From virtual assistants to autonomous museum buses, from flood detection systems to learning tools – AI is helping us build a smarter, more human-centered government and governance.

But this has not always been the case. Even though digitalisation has been in the bloodstream of our society for decades, having a human centric approach is rather new.

When we first started with our digital transformation, the way to approach digital services was to increase efficiency among the government, considering the little resources available. Gradually we shifted our mindset to a human-centric approach, it's when we slowly started to see the satisfaction rates of citizens going up.

According to the recent research, there's a 83% satisfaction rate with the public e-services. And the aim is to reach 90% of user satisfaction of online public services by 2030. Satisfaction creates trust, trust is an important social capital to move forward.

Dear Friends,

Estonia have now grounded its digital journey on a human-centric approach, shaped by i**nfrastructure**, **responsibility**, and **education**. Let me elaborate on them.

Firstly – **infrastructure**. Estonia's digital society is built on a secure and interoperable data ecosystem, supported by X-Road – that is our decentralized data exchange layer that allows real-time, cross-institutional data sharing. This system is developed and maintained in partnership through the Nordic Institute for Interoperability Solutions (NIIS) – a great example of regional cooperation delivering global impact.

We believe that digital public infrastructure is the silent enabler of AI. It ensures security, resilience, and trust in every service we build. Such infrastructure provides the foundational building blocks of the digital society – such as secure digital identity, data exchange, and payment systems – that ensure resilience, efficiency, and trust across the entire digital ecosystem.

Secondly – **responsibility**. Innovation without trust is not sustainable. That's why we are fully committed to implementing the EU AI Act. We support its risk-based approach, which offers a global model for responsible regulation – one that protects rights while still encouraging innovation.

According to the Freedom on the Net's 2024 Report, Estonia ranks second in the world for safeguarding digital rights and ensuring online freedom. We prove that it is possible to protect privacy and foster a secure digital environment, where human rights are respected.

These results have been achieved through experiences and having the courage to test and develop solutions. It is why we are also investing in AI sandboxes, training programs, and competence centers to support both public and private sector actors in understanding and complying with emerging regulations.

To enhance transparency and accountability, the citizens of Estonia can always see when and how their personal data is accessed across state systems. We have also launched an Algorithm Registry to document and explain how AI is used in public administration. These **innovative** tools help ensure that AI in government is not only powerful – but also explainable, fair, and just.

Thirdly, I would like to talk about **education**. We know that any major transformation starts with education. If we want an AI literate society – meaning, resilient and ready for the future – we need to integrate these new tools into school curriculum.

That's why, as part of our AI Leap initiative, we are providing 20,000 students and teachers in Estonia with free access to leading AI-powered learning tools. Responsibility is of utmost importance here as well. As evidence builds that misguided use of Large Language Models can have a negative impact on people's thinking skills, we are designing the AI Leap to counter this. This requires both transforming mindsets and teaching methods, and building responsible AI tools.

Ladies and Gentlemen,

Therefore, AI governance is not only a national challenge or a national accelerator – it is also a global one. Technology does not care about borders. Deepfakes, surveillance misuse, and algorithmic manipulation threaten democratic processes everywhere. That's why Estonia supports the creation of non-binding global principles for AI governance, developed through inclusive, multistakeholder cooperation.

We are therefore deeply engaged in many international forums. We support the Global Digital Compact, which defines global principles for an open, secure and inclusive digital future, as well as the Open Government Partnership, which advances transparency and citizen participation in governance.

We proudly co-chaired the Media Freedom Coalition for two years. And in 2025, Estonia have taken up the chairmanship of the Freedom Online Coalition, continuing our commitment to safeguarding human rights in the digital age.

We believe that international cooperation and learning from each other are essential to ensure that AI develops in a way that serves our societies, democracies, and shared values.

In October this year I will host the Arroiolos Group meeting of Presidents in Tallinn, where we will discuss the driving force and power, but also challenges of AI.

In the same spirit, the Tallinn Digital Summit will take place on 9-10 October 2025 which I kindly invite you all to attend.

Held under the theme "Collectively at the Crossroads: Towards Secure and Resilient AI Futures," the Summit will bring together political and digital leaders to advance dialogue on AI governance and its societal impact. It will be held alongside the **Freedom Online Coalition Ministerial** further strengthening the strategic relevance of the topic.

Estonia also supports initiatives that aim to bridge the digital divides, promote digital literacy, and build inclusive AI ecosystems that empower individuals – especially in developing countries.

As AI technologies evolve, there is a growing risk that those without access to infrastructure, skills, or legal safeguards will be left even further behind. Without deliberate action, AI may deepen – not diminish – existing inequalities. That is why we see digital inclusion not only as a development goal, but as human rights matter.

Through ESTDEV, our Development Cooperation Agency, Estonia is sharing its know-how in digital realm with partner countries – from creating secure and transparent e-governance frameworks, to supporting civic tech, digital rights awareness, and AI-powered educational platforms.

For example, in Africa, Estonia supports Kenya, Uganda, Namibia and Botswana in building inclusive, secure and AI-ready digital societies. In Kenya, we have helped develop digital public infrastructure and e-services, strengthened national cyber resilience, and empowered youth through STEM and ICT training.

In Uganda, Namibia, and Botswana, our cooperation focuses on digital health, e-governance, and entrepreneurship – laying the groundwork for ethical and human-centric AI in public services. These are only some examples of our steps around the world.

Let me summarise my message in five points.

- First, AI must be **human-centered, rights-based, and transparent**.
- Second, **innovation and regulation must move on together** – not against each other.
- Third, no country, no matter how advanced, can succeed alone – **cooperation is vital**.
- Fourth, we must ensure that **no one is left behind**. Digital transformation must benefit everyone – every member of society and reach all corners of the world.
- Fifth, **education is the foundation** of any societal transformation – therefore, also essential for safe AI adoption and societal resilience.

Estonia's journey proves that digital transformation is possible and it can be done with little of resources. In just over three decades, we have gone from rebuilding our institutions to leading the world in e-governance.

Our success has never been just about technology. It's been about vision, cooperation, and trust. We know from our own experience what is possible to achieve with the right vision and determination.

The heart of Estonia's previous digital transformation was the understanding that we should not simply digitalise existing services but rather rethink and rebuild the whole system.

This should also be our guidance in the AI age. Pooling skills and resources support better and faster digital development. It is also the reason why we strongly support the global movement for digital public goods and digital public infrastructure.

Let us now bring that same spirit to the global challenge of governing AI – so that we may build a future where technology strengthens our democracies, protects our freedoms, and empowers every individual.

Thank you.

Quote:

- *"Even though digitalisation has been in the bloodstream of our society for decades, having a human centric approach is rather new." (H.E. Mr. Alar Karis, President of the Republic of Estonia)*

3 Message on behalf of His Holiness Pope Leo XIV

On behalf of His Holiness, Pope Leo XIV, I would like to extend my cordial greetings to all participants in the AI for Good Summit 2025, organized by the International Telecommunication Union (ITU), in partnership with other UN agencies and co-hosted by the Swiss Government. As this summit coincides with the 160th anniversary of the ITU's foundation, I would like to congratulate all the Members and staff for their work and constant efforts to foster global cooperation in order to bring the benefits of communication technologies to the people across the globe. Connecting the human family through telegraph, radio, telephone, digital and space communications presents challenges, particularly in rural and low-income areas, where approximately 2.6 billion persons still lack access to communication technologies.

Humanity is at a crossroads, facing the immense potential generated by the digital revolution driven by Artificial Intelligence. The impact of this revolution is far-reaching, transforming areas such as education, work, art, healthcare, governance, the military, and communication. This epochal transformation requires responsibility and discernment to ensure that AI is developed and utilized for the common good, building bridges of dialogue and fostering fraternity, and ensuring it serves the interests of humanity as a whole.

As AI becomes capable of adapting autonomously to many situations by making purely technical algorithmic choices, it is crucial to consider its anthropological and ethical implications, the values at stake and the duties and regulatory frameworks required to uphold those values. In fact, while AI can simulate aspects of human reasoning and perform specific tasks with incredible speed and efficiency, it cannot replicate moral discernment or the ability to form genuine relationships. Therefore, the development of such technological advancements must go hand in hand with respect for human and social values, the capacity to judge with a clear conscience, and growth in human responsibility. It is no coincidence that this era of profound innovation has prompted many to reflect on what it means to be human, and on humanity's role in the world.

Figure 22: (from left) H.E. Sister Rafaela Petrini, H.E. Archbishop Ettore

Although responsibility for the ethical use of AI systems begins with those who develop, manage and oversee them, those who use them also share in this responsibility. AI therefore requires proper ethical management and regulatory frameworks centered on the human person, and which go beyond the mere criteria of utility or efficiency. Ultimately, we must never lose sight of the common goal of contributing to that "tranquillitas ordinis – the tranquility of order", as Saint Augustine called it (De Civitate Dei) and fostering a more humane order of social relations, and peaceful and just societies in the service of integral human development and the good of the human family.

On behalf of Pope Leo XIV, I would like to take this opportunity to encourage you to seek ethical clarity and to establish a coordinated local and global governance of AI, based on the shared recognition of the inherent dignity and fundamental freedoms of the human person. The Holy Father willingly assures you of his prayers in your efforts towards the common good.

Card. Pietro Parolin
Secretary of State of His Holiness

Quote:

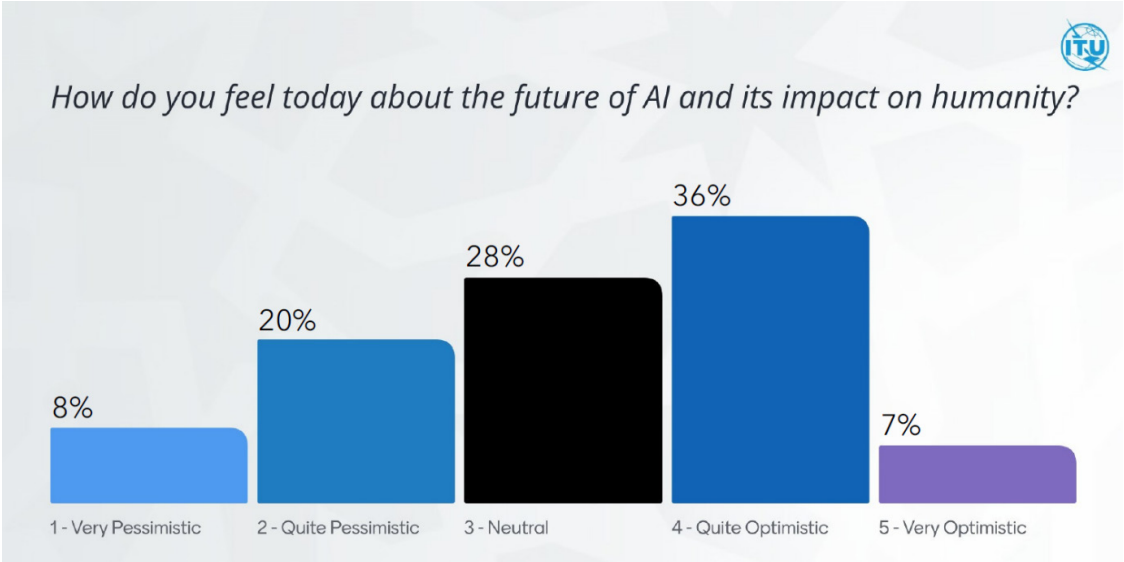
- *"I would like to take this opportunity to encourage you to seek ethical clarity and to establish a coordinated local and global governance of AI, based on the shared recognition of the inherent dignity and fundamental freedoms of the human person." (Pope Leo XIV)*

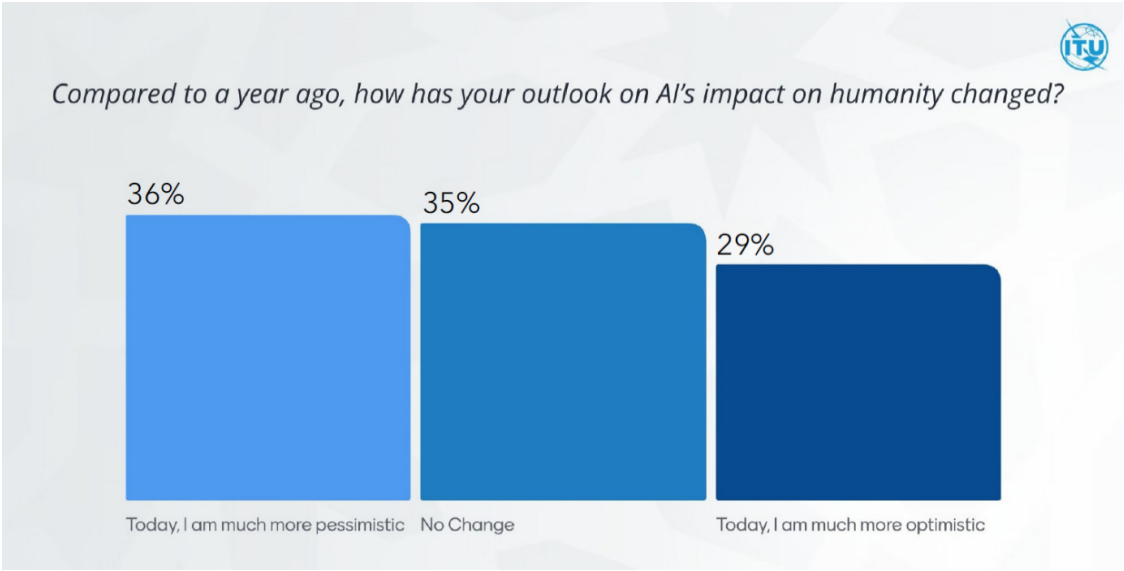
4 Informal Polling Results among Luncheon Participants

To try to capture the mood of participants, a real-time poll using Mentimeter (www.menti.com), an audience response application, was conducted. About 100 invitees responded.

1. How do you feel today about the future of AI and its impact on humanity?
2. Compared to a year ago, how has your outlook on AI's impact on humanity changed?
3. In one word, how would you describe your overall outlook regarding the future development of AI?

Here are the answers:





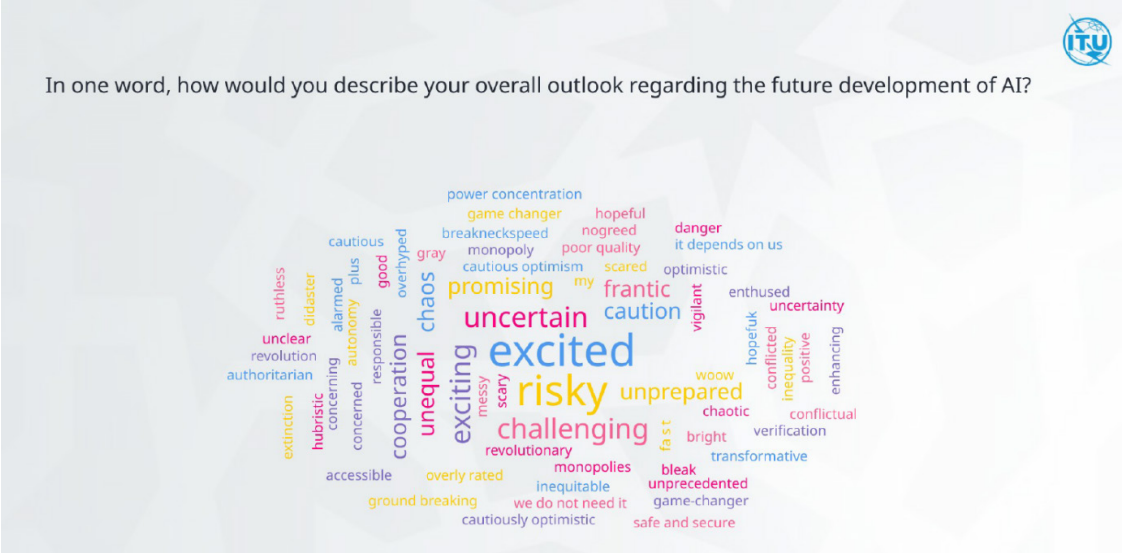


Figure 23: Results of the real-time poll at the luncheon

5 UN AI Activities

The Dialogue also reviewed the growing role of the United Nations in AI governance. In 2023, the UN system reported **406 AI-related projects**; by 2024, this number had increased to **729 projects**, reflecting the rapid integration of AI across the UN family. These activities range from humanitarian applications to climate modeling, education initiatives, and governance pilots.

This system-wide momentum demonstrates both opportunity and risk. On the one hand, it shows that AI can be harnessed for the UN's core missions: peace, human rights, and sustainable development. On the other hand, it underscores the need for coordination: without common principles and oversight, UN agencies risk duplicating efforts or reinforcing the very divides they aim to bridge.

Participants urged that by 2026, the UN should establish a **system-wide reporting framework** on AI governance, similar to existing climate accountability mechanisms. Such a framework could track progress, share best practices, and hold institutions accountable to global standards.

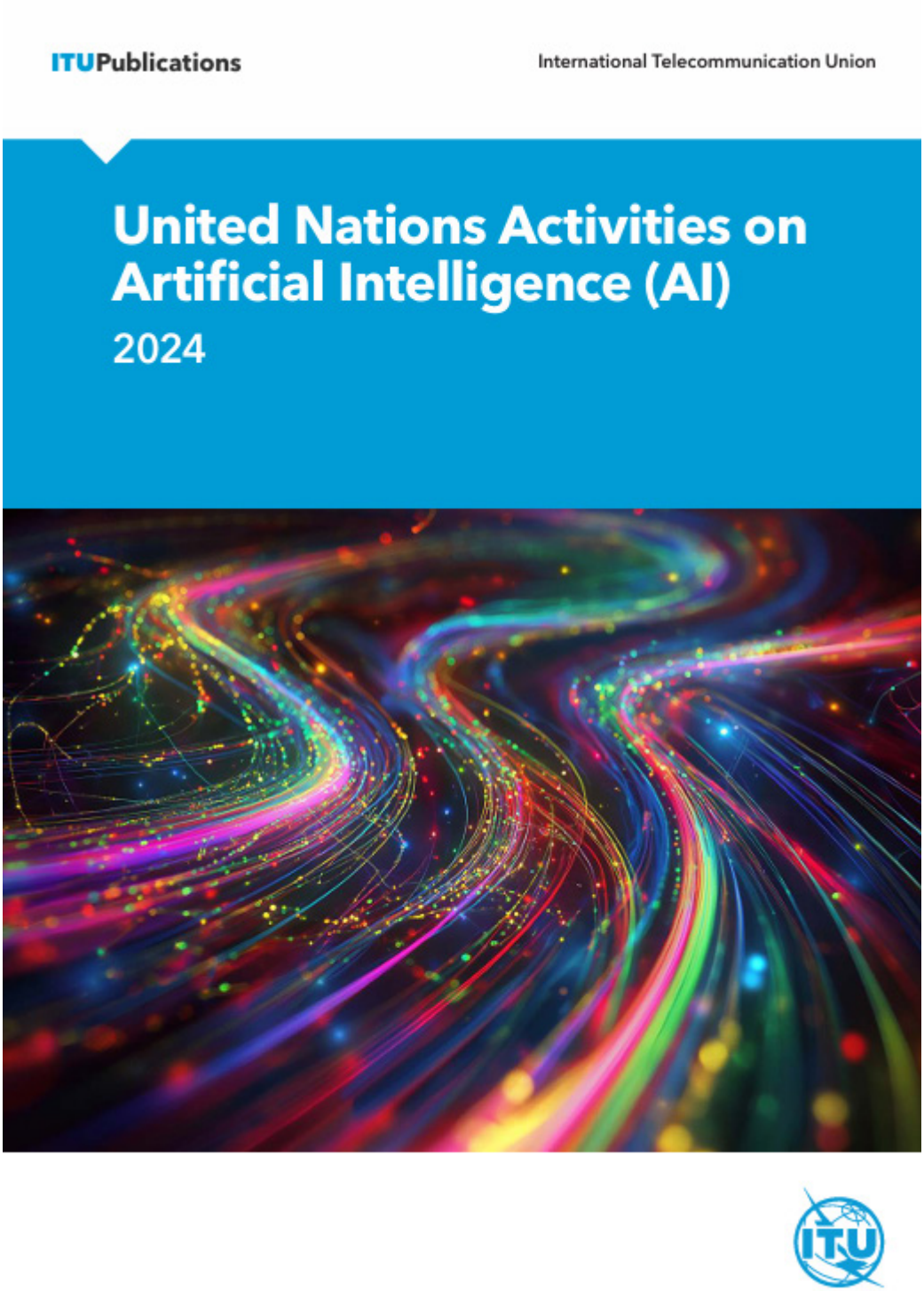


Figure 24: The UN's activity report on AI, published at the AI for Good Global Summit on 10 July 2025

6 Acknowledgments

The Secretariat would like to extend its sincere gratitude to Professor Robert Trager, Co-Director of the Oxford Martin AI Governance Initiative, International Governance Lead at the Centre for the Governance of AI, and Senior Research Fellow at the Blavatnik School of Government at the University of Oxford, and his team. Their generous advice and support were invaluable throughout the curation of the AI Governance Dialogue. We also appreciate Robert Trager for moderating sessions at the AI for Good Global Summit, including AI Governance Day, and for moderating the AI Governance Dialogue luncheon.

We would like to give a special thank you to Miro Plueckebaum and Sumaya Nur Adan for their many practical suggestions during numerous conference calls and email exchanges in the months leading up to the AI Governance Dialogue.

A very big thank you also goes to Sumaya Nur Adan, who led the drafting of the white paper "Themes and Trends in AI Governance", with the support of Miro Plueckebaum, Luise Eder, and Shannon Hong, and to Sam Daws and Marta Ziosi for their valued contributions.

AI Governance Dialogue was only possible thanks to many invisible elves of the ITU Secretariat and service providers who worked behind the scenes and made it all happen.

Finally, we thank all the participants who traveled to Geneva for AI Governance Dialogue. We hope you found it to be a fruitful experience.

7 See you in 2026

We look forward to seeing you at the AI for Good Global Summit on 7-10 July 2026 in Geneva at Palexpo.

Figure 25: Good-bye 2025, and see you in 2026

International Telecommunication Union
Place des Nations
CH-1211 Geneva 20
Switzerland

ISBN 978-92-61-41441-2

Published in Switzerland
Geneva, 2025

Photo credits: Adobe Stock