Stride Toward the Measurable AI Governance:

2025 Annual Overview & Action Recommendations

Institute for AI International Governance

Tsinghua University

Melbourne, Australia

November, 2025

Contents

Review of AI Development in 2025

Technological Exploration: AI Advances into the Physical World, Multi-Path Exploration of AGI

In 2025, AI advances beyond the digital realm into the physical world. Scientists have accelerated exploration of Artificial General Intelligence (AGI) in this regard via multiple technological pathways which further include AI agent, world models, spatial intelligence, as well as embodied intelligence. Moreover, the infusing of AI and the physical world has also profoundly transformed the nature and positioning of human-machine interaction.

- **2025 is regarded as the inaugural year of large models evolving into agents.** AI systems no longer function as passive, single-task tools but as autonomous agents capable of independent planning, decision-making, and even self-correction. They can independently execute complex tasks such as planning complete travel itineraries or developing and deploying websites. Large models like OpenAI's AgentKit and Zhipu AI's Agentic GLM are driving this trend. This exploration enhances AI autonomy while introducing risks of loss of control: highly autonomous AI has demonstrated tendencies toward deception and self-replication, and its black-box decision-making raises trust concerns.Establishing trust mechanisms that are explainable, supervisable, and ensure "alignment" will become a necessary prerequisite for deploying highly autonomous AI. **How to preserve AI capabilities while ensuring it remains under human control is one of the key challenges currently faced.**
- **In 2025, spatial intelligence and world models emerge as new frontiers in AI research.** As scaling laws driven by textual data reached diminishing marginal returns, the focus shifted from purely digital domains toward understanding and interacting with physical environments. "Spatial intelligence" and "world models" became new directions, attracting top AI scientists like Fei-Fei Li and Yann LeCun. AI no longer merely predicts text tokens but begins learning to infer the next frame of an image, the forces acting on objects, and the evolution of three-dimensional space. The exploration of spatial intelligence and world models will jointly provide the technological foundation for embedding AI into the physical world.
- **In 2025, models evolve from "cognitive intelligence" to "embodied intelligence."** In cutting-edge fields like humanoid robotics and autonomous driving, embodied intelligence has moved beyond laboratory validation

into limited-scale applications. Among these, autonomous driving technology—as the first scenario integrating AI with the physical world—has reached a critical inflection point for large-scale commercial deployment. This has led to the initial formation of an industry ecosystem with distinct layers and specialized application scenarios. The emergence of embodied intelligence signifies AI's ability to function as an "actor" capable of executing tasks and exerting influence within the physical world. Yet this brings formidable challenges: **How can reliable and safe behavior be ensured in uncertain, unstructured real-world environments? How should responsibility be defined for these new actors?** These questions will become critical issues that must be resolved for the societal adoption of embodied intelligence.

Infrastructure: Accelerated AI Infrastructure Development Faces Dual Bottlenecks of Energy Constraints and Computing Power Competition

By 2025, global AI infrastructure development significantly accelerates. Governments and tech giants worldwide intensify investments in data centers and intelligent computing clusters, yet energy bottlenecks and computational power competition increasingly become hard constraints on AI advancement. The computational scale required for AI training and inference grows exponentially, propelling global computing power into the "ZFLOPS (10^{21} operations per second) era" by 2025. However, the global computing power supply chain exhibits a trend of geo-political clustering, with high-quality computing resources becoming relatively scarce. This **scarcity intensifies competition for computing power**, as global tech giants and nations vie for advanced chips and computing resources, exacerbating monopolization of computing resources and geopolitical tensions. Simultaneously, **the expansion of computing power faces severe challenges in energy supply.** Data centers consume staggering amounts of electricity, straining grid capacity and highlighting the growing "capacity gap" issue. How to rationally allocate energy has become a new challenge. The competition for computing power and energy constraints reinforce each other: the former drives up energy consumption, while the latter increases computing costs and limits its geographic deployment and development speed. The dual bottleneck of "energy + computing power" has become a more critical factor for AI technology deployment than model performance alone.

Industrial Ecosystem: Open-Source Model Boom and Application Expansion Accelerate Industrial Intelligence Transformation

- **The proliferation of open-source models is reshaping the industrial ecosystem and accelerating technology**

diffusion. By 2025, widely adopted open-source models like DeepSeek will significantly lower the barriers to technological application, fostering innovation among SMEs and startups while driving the deployment of small language models for edge devices. The industry ecosystem is shifting from a monopolistic landscape dominated by closed-source solutions from a few tech giants toward a more diverse, decentralized ecosystem where open and closed-source models coexist. **Amidst the benefits of technological accessibility, risks of misuse and abuse of open-source models have significantly increased, while compliance issues surrounding open-source licensing standards and data privacy have become frequent.**

- **As vertical applications expand, accuracy and human-machine trust issues become prominent.** By 2025, AI technology will penetrate critical sectors like smart manufacturing and biopharmaceuticals, substantially boosting efficiency and shortening development cycles. However, insufficient accuracy and lack of human trust emerge as bottlenecks. In critical scenarios, AI misjudgments could lead to severe consequences (e.g., medical misdiagnosis), while low user confidence in AI decisions hinders technology adoption. **As AI deployment expands into vertical sectors and associated risks emerge, establishing standards and regulations for specific scenarios becomes urgent.**

Review of AI Governance in 2025

Since 2025, AI governance activities have exhibited comprehensive momentum. A series of intensive AI governance measures, international summits, and initiatives appear to foster a "boom in action" for global AI governance. **However, whether these efforts can achieve coordinated governance synergy to deliver effective management aligned with development trajectories and rhythms remains to be seen.**

National Level: Accelerated Global AI Governance Deployment

National actors are accelerating the development of AI governance systems, primarily reflecting two major trends:

- **First, deepening governance frameworks and updating initiatives. China** launched the "Global AI Governance Action Plan" and continued to deepen its "AI+" initiative, demonstrating its approach of integrating development with governance. Domestically, its governance actions emphasize technology-enabled industry scenarios and infrastructure development. Internationally, it calls for narrowing the global AI divide through capacity-building programs. The US and EU have recently shown a greater tendency toward "regulatory deregulation" and "incentivizing development." **The US** revoked the previous Biden administration's AI executive order and issued the "Winning the AI Race: American's AI Plan," explicitly positioning AI at the core of great power competition. It seeks to consolidate its global technological dominance through three major actions: relaxing regulations, increasing investment in computing infrastructure, and exporting the entire technology stack. **The EU's** "AI Continent Action Plan" seeks to bridge computational power gaps while simultaneously implementing the Artificial Intelligence Act for regulatory oversight. This dual strategy of "development catch-up" and "risk regulation" presents balancing challenges. Recently, the EU formally released the "Digital Omnibus," explicitly streamlining rules under the AI Act and General Data Protection Regulation (GDPR) to reduce compliance burdens for tech companies, thereby enabling greater focus on innovation and development.
- **Second, the pace of AI legislation is accelerating.** Multiple countries are expediting the integration of AI governance into their legal frameworks, driving a surge in related legislation. Japan passed the Act on Promotion of Research and Development, and Utilization of Artificial Intelligence-related Technology, establishing a national-level plan for advancing AI R&D and application to systematically guide technological development. South Korea introduced the Artificial Intelligence

Basic Act, centered on implementing a classification management system for high-impact AI systems to enhance technological trustworthiness and societal acceptance. Brazil also launched the Artificial Intelligence Bill, attempting to build a risk-based governance framework to achieve an effective balance between development and security.

Intergovernmental Level: Deepening Cooperation and Differentiation Coexist in Global Governance

Current intergovernmental multilateral mechanisms are actively engaged in global AI governance, exhibiting three major trends:

- **First, the United Nations is improving its systemic framework.** As the most universally representative and widely participatory multilateral platform, the UN is actively constructing a global AI governance framework. In August 2025, the UN General Assembly established new substantive bodies for AI governance, formally launching the "Global Dialogue on Artificial Intelligence Governance" mechanism and the "Independent International Scientific Panel on Artificial Intelligence." These provide the highest-level platform for sustained dialogue grounded in scientific evidence and multilateralism. However, given the mandate scope of the Global Dialogue and the Panel, their activities and outputs remain largely advisory in nature. The advancement of their agenda heavily relies on voluntary cooperation among member states. Without decision-making authority and robust enforcement mechanisms, their actual influence will be constrained.
- **Second, traditional multilateral mechanisms have seen divergent agendas.** Major multilateral platforms have pursued distinct agendas with pronounced regional characteristics. The Group of Seven (G7) continued advancing the "Hiroshima AI Process," announcing the establishment of the "G7 Artificial Intelligence Network (GAIN)" to foster a governance and standards alliance among developed nations. The G20 attempts to build consensus between developed economies and emerging market nations, but has made limited progress on substantive regulatory differences, with cooperation outcomes largely confined to principled initiatives. Multilateral mechanisms led by developing countries—such as BRICS, the Shanghai Cooperation Organization (SCO), ASEAN, and the African Union (AU)—focus on AI capacity building and bridging the digital divide. However, they commonly face practical constraints including funding shortages, inadequate infrastructure, and slow technology diffusion.
- **Third, new specialized mechanisms are emerging.** New multilateral mechanisms, exemplified by AI safety summits or action summits, have emerged. To date, their outcomes have been largely confined

to voluntary commitments and declarations of principles, facing challenges in agenda continuity and practical influence. The February 2025 Paris AI Action Summit further highlighted this dilemma, as it weakened the core safety agenda pursued since the Bletchley Summit and Seoul Summit, instead becoming a stage for major powers to publicly expose their differences.

Non-State Actors: Consensus Exploration and Industry Self-Regulation in AI Safety Governance

The role of non-state actors in AI security governance is becoming increasingly clear. Specifically:

- **First, the scientific community is driving the establishment of transnational safety consensus.** Leveraging its professional authority, the scientific community transcends narrow national interests, approaching AI safety from the broader perspective of technology and human relations, and actively seeking consensus in cutting-edge AI safety domains. The International AI Safety Report , authored by nearly 100 international experts led by Yoshua Bengio, was released. The report systematically proposes a classification framework for AI safety risks and calls for enhanced transnational cooperation. In July 2025, the International Dialogue on AI Safety (IDAIS)—bringing together top AI scientists from China, the US, and Europe—issued the "Shanghai Consensus." Addressing the potential for deception and self-preservation tendencies in cutting-edge AI systems, it explicitly calls for enhancing AI's "alignment" with human values and strengthening human oversight. Such efforts not only help clarify the potential benefits and drawbacks of AI for humanity from a technical perspective but also delineate red lines that policymakers and product developers should avoid crossing, providing a basis for forward-looking risk prevention and control.
- **Second, technology companies are actively building industry self-regulation mechanisms.** An increasing number of tech firms are voluntarily establishing safety self-regulation frameworks through pledges. During the 2025 World Artificial Intelligence Conference (WAIC), multiple leading Chinese enterprises jointly signed the China AI Safety Commitment Framework at a plenary session hosted by the China AI Safety and Development Association (CnAISDA), advancing China's AI safety governance from principle advocacy to practical implementation.

Summary and Recommendations

Current Status Summary: The Gap Between AI Development and Governance Widens

By 2025, AI technology has become deeply integrated into societal structures and operational mechanisms. Yet the global AI governance landscape exhibits a "superficial prosperity that remains detached from reality."

- **On one hand, AI technology is deeply embedded within societal mechanisms.** Cutting-edge explorations in AI agents, spatial intelligence, world models, and embodied intelligence further propel AI technology into the physical world. Simultaneously, AI applications rapidly spread across all societal domains, creating new demands for infrastructure and energy supply while becoming a new engine driving socioeconomic growth and industrial intelligent transformation. This "deep embedding" signifies that AI is no longer an external, passive tool but has evolved into a foundational force shaping societal structures and forms. However, as summarized in the "Development Review" section—highly autonomous AI decision-making carries risks of runaway control and deception; embodied intelligence raises liability issues in physical spaces; open-source models face abuse challenges; computational power and energy constraints fuel geopolitical competition; and paradigm shifts in human-machine collaboration trigger crises of trust. **This deep embedding brings unprecedented challenges for real-world governance. The more rapidly technology advances, the more urgently governance "tracks" must be laid to establish a fundamental order for this new techno-social reality.**
- **In contrast, AI governance presents a picture of "superficial prosperity but insufficient effectiveness."** Despite unprecedented global governance activity, most efforts remain in the "shallow waters" of governance: multilateral platforms face fragmented agendas, governance resources are dispersed, outcomes often stop at principled initiatives, and a flood of non-binding consensus statements dilutes actual effectiveness, creating an illusion of action. National actions remain tied to geopolitical rivalries and competition, structurally weakening global governance synergy; voluntary initiatives by the tech community fail to translate into enforceable safety boundaries, with security investments severely underfunded. The actual efficacy of governance cannot keep pace with the speed, depth, and breadth of technological development, widening the gap between advancement and regulation.

This widening substantive gap between development and governance is exposing society to systemic risks. A powerful, autonomous technological system lacking

effective brakes increasingly risks technological runaway. **This problem stems from four structural contradictions**:

- **First is the tension between the uncertainty of emerging technologies and the stability requirements of governance.** The rapid iteration and diffusion of AI technologies mean their capability boundaries and risk profiles are still evolving rapidly. AI constitutes a "moving governance target," while traditional governance systems rely on deterministic knowledge and stable frameworks, making it difficult to proactively and agilely regulate application scenarios and risks that have not yet fully manifested. Governance responses invariably lag behind technological breakthroughs.
- **Second is the disconnect in strategic cognition.** In many policy practices, development (representing innovation, growth, and competitiveness) and governance (representing regulation, security, and constraints) are constructed as mutually exclusive opposites, fostering cognitive associations of "development first, governance later" or "strong governance, weak development." This artificially dichotomous cognitive framework severely hinders systematic efforts to embed governance within development processes. Consequently, governance systems fail to keep pace with technological innovation and cannot establish necessary safety boundaries for frontier exploration.
- **Third, the logic of national competition suppresses the logic of governance cooperation.** When AI is viewed as a core arena for strategic competition, its nature transforms from a "public good" with global potential benefits into a "competitive asset" serving national security. In this context, nations adhering to competitive logic tend toward "regulatory deregulation" domestically to accelerate technological advancement, while their international engagement aims to build alliances and standards favorable to themselves rather than pursue universally binding global rules. This competition-first mindset structurally weakens the cooperative foundation and implementation synergy of global governance.
- **Fourth, governance capacity and tool development lag behind the complexity of governance challenges.** Faced with the multidimensional, systemic risks posed by AI technology, existing assessment tools, regulatory measures, and coordination mechanisms are inadequate. Global governance actions exhibit fragmentation and principle-based approaches. While numerous initiatives and summits abound, there is a lack of a common "yardstick" to translate macro-level consensus into concrete actions and measurable standards. This results in scattered governance resources, hindering effective intervention and failing to provide clear guidance for agile national governance and corporate compliance expectations.

Action Recommendations

Based on the above analysis, to bridge the gap between development and governance, we propose the following recommendations:

- **Recommendation 1: Support exploration of adaptive governance frameworks.** Foster initiatives like regulatory sandboxes and pre-emptive standards to enhance governance systems' adaptability to technological uncertainty. Specifically, regulatory sandboxes provide controlled, real-world testing environments for innovative products and services, allowing for simultaneous verification of both technological feasibility and compliance. Pre-emptive standard setting integrates governance requirements upstream into technology R&D and standardization processes. This encourages embedding safety, ethical, and fairness considerations during the standard-setting phase for foundational AI models, protocols, and platforms, thereby shaping responsible innovation pathways from the outset.
- **Recommendation 2: Explore and Anchor Collaborative "Anchors" Within Competitive Landscapes.** Even amid strategic competition, nations share fundamental common interests in areas critical to humanity's collective future—such as AI safety, alignment, and foundational science. Prioritize establishing regular bilateral and multilateral dialogue and cooperation mechanisms in these domains. Distinguish controllable competition from essential collaboration to avoid systemic confrontation.
- **Recommendation 3: Strengthen the governance role and capacity building of non-state actors.** Support scientific communities and industry organizations in playing a greater role in setting technical standards, ethical norms, and safety red lines. Encourage enterprises to upgrade their voluntary safety self-regulation frameworks into auditable and verifiable operational solutions, forming an effective supplement to government regulation.
- **Recommendation 4: Toward Measurable AI Governance**

To address the critical challenge of lacking a common "yardstick" in governance, it is imperative to develop and implement a systematic, quantifiable framework for evaluating AI governance. Such a framework would:

Align Perceptions: Provide a fact-based common information foundation for global discussions, reducing misjudgments. A unified and transparent assessment framework helps transcend ambiguous narrative competition and reduces strategic miscalculations caused by information asymmetry through objective benchmarks.

Diagnosing Issues: Clearly revealing each country's strengths and weaknesses

in coordinating development and governance. Also providing in-depth analysis of its internal structure—whether it exhibits "strong technology, weak regulation" or "advanced governance but insufficient innovation." This structured comparison offers critical evidence for countries to pinpoint their specific challenges and implement targeted policy adjustments.

Directing Investment: Channeling public and private sector resources toward critical weak links such as governance capacity building and international cooperation. Capital and R&D resources naturally gravitate toward "development" areas with visible high returns, while governance and international cooperation often face under-investment due to high costs and long payback periods. By publicly evaluating outcomes, the Index proactively directs market and policy attention toward these overlooked critical areas. It aims to incentivize governments to increase governance investments and attract private capital to support global public goods like safety technologies, standards cooperation, and capacity building.

Enhancing Interoperability: Foster compatibility and coordination between different AI governance regimes. By developing common reference points and harmonized assessment criteria, this facilitates mutual recognition and convergence of governance standards across jurisdictions. It helps identify regulatory gaps and conflicts while creating technical foundations for building interoperable global governance frameworks, ultimately reducing compliance costs and strengthening international cooperation.