GLOBAL AI INNOVATIVE GOVERNANCE

2026 / 1 (Vol. 1, Issue 1)

CGAIG

CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

CONTENTS

Global AI Innovative Governance
2026/1

Inaugural Message

Advancing Global AI Innovative Governance Research in the Intelligent Age

Artificial intelligence is integrating into economic development, social governance, scientific research, international relations, and human life at an unprecedented pace. It serves not only as a key driver of the latest round of technological revolution and industrial transformation, but also as a pivotal variable shaping the evolution of global order, national governance capacity, social equity and justice, and the trajectory of human civilization. The advancement of AI has not only raised practical demands for technological innovation and industrial upgrading, but also triggered a host of major governance issues concerning rule-building, risk prevention and control, ethical constraints, international cooperation, and the protection of public interests.

Currently, global AI governance is at a critical stage where accelerating rule-making, intensifying institutional competition, proliferating technological risks, and reshaping international consensus are unfolding in parallel and interwoven with one another. On the one hand, generative AI, large-scale models, computing infrastructure, cross-border data flows, and algorithmic applications continue to expand the frontiers of AI capabilities. On the other hand, issues such as technological misuse, algorithmic bias, data security, employment disruption, digital divides, value conflicts, and governance fragmentation are becoming increasingly prominent. AI governance has far exceeded the scope of technology supervision; it is now deeply embedded in the comprehensive agenda of national development, global governance, international order, and the common future of humanity.

Against this backdrop, Global AI Innovation Governance, a serial publication jointly launched by the Center for Global AI Innovation Governance and Shanghai People's Publishing House, is dedicated to building a research platform that features academic depth, policy relevance, international vision, and interdisciplinary character. It is our aspiration that this platform will continuously bring together the intellectual contributions of scholars, policy researchers, technical experts, and industry practitioners from both China and abroad, fostering in-depth discussions on cutting-edge issues, institutional logics, and practical pathways in AI governance,

and thereby advancing a Chinese approach to AI governance that is both explanatory and constructive, and capable of engaging in international dialogue.

Fudan University has long placed great emphasis on serving national strategies, promoting interdisciplinary research, and participating in global knowledge production. The Center for Global AI Innovative Governance, building on Fudan's research strengths in the humanities and social sciences, including AI global governance, international relations, law, economics, public administration, journalism and communication, and related interdisciplinary fields—and with strong support from the Shanghai Municipal Government, is committed to building a collaborative platform connecting Shanghai, China, and the world in the field of AI governance. The Center's Secretariat, housed at the Fudan Development Institute, aims to contribute to the construction of broadly consensual governance frameworks, standards, and guidelines for AI through academic research, policy advisory, international dialogue, talent cultivation, and industry-academia-research collaboration.

The launch of *Global AI Innovative Governance* is the academic, institutionalized, and sustained expression of this mission. We hope that this publication will serve not only as a platform for publishing papers, but also as a forum for problem identification, knowledge accumulation, rule exploration, and intellectual exchange. Research on AI governance should not be confined to conceptual description or policy tracking; it must go further to address a series of deeper questions: How is AI reshaping global power structures? How do governance models in different countries and regions interact, compete, and shape each other? How do data, computing power, models, standards, and platforms become new governance resources? How can the Global South enhance its participatory capacity and agenda-setting power in AI governance? How does AI affect human subjectivity, social trust, public ethics, and mutual learning among civilizations? These questions require not only technological understanding but also institutional analysis; not only Chinese experience but also international comparison; not only practical concern but also theoretical innovation.

It is with these considerations that this publication will adhere to the following four fundamental orientations.

First, the integration of global vision with Chinese subjectivity.

AI governance is a global issue, but countries differ significantly in their stages of development, institutional structures, technological capabilities, and value traditions. China is both a major participant in AI development and an important contributor to global AI governance. This publication aims to examine Chinese issues against the broad backdrop of global governance transformation, and to respond to global challenges on the basis of China's practices and experiences, thereby fostering a more open, rational, explainable, and dialogue-oriented Chinese approach to AI governance.

Second, the integration of theoretical innovation with policy contribution. Research on AI governance should neither merely chase hot topics nor stop at general appeals. Truly valuable research should be able to reveal institutional logics, explain real-world changes, identify governance risks, and provide a solid and translatable knowledge base for policy design, rule-building, and international cooperation. This publication encourages research that demonstrates clear problem awareness, rigorous reasoning, original viewpoints, and practical explanatory power.

Third, the integration of interdisciplinary integration with problem-orientation. AI governance naturally spans multiple disciplines, including political science, international relations, law, economics, sociology, journalism and communication, ethics, public administration, and computer science, and no single discipline can fully grasp its complexity. This publication aims to promote deep dialogue and knowledge integration across disciplines around real-world problems, rather than simple conceptual juxtaposition or self-referential disciplinary cycles.

Fourth, the integration of open exchange with academic community-building. The complexity of AI governance means that relevant research cannot be conducted within closed systems. In the face of rule divergence, technological competition, and value differences, this publication is committed to building a sustained, open, and constructive platform for exchange, bringing together scholars, policy researchers, technical experts, and industry practitioners from China and abroad to jointly advance the formation of an internationally influential research community on global AI governance.

The inaugural issue is only the beginning. In the face of the rapidly evolving AI era, no governance research can achieve a once-and-for-all solution. Rules must be tested in practice, theories

must be updated in response to change, and international consensus must be accumulated through dialogue. This publication will maintain a long-term commitment to tracking this epochal issue, documenting the major questions in the evolution of AI governance, promoting high-quality knowledge production, and facilitating deeper, more equitable, and more constructive exchanges between Chinese academia and the international community on AI governance.

We sincerely thank the experts, scholars, and colleagues who have contributed to the writing, review, editing, and publication of the inaugural issue, and we look forward to welcoming more researchers who care about the future of AI, global governance transformation, and humanity's common well-being to join this platform. May this publication, with its rigorous scholarship, open-mindedness, and constructive spirit, contribute intellectual strength to global governance in the intelligent age.

Jin Li

President, Fudan University

June 2026

Structures and Frameworks

Global Governance of Artificial Intelligence: Current Landscape, Inherent Paradoxes, and Future Pathways

*Li Yan**

Abstract Against the backdrop of an accelerating evolution of intelligent society intertwined with profound changes unseen in a century, artificial intelligence (AI) has risen from a technological application to a comprehensive strategic engine that drives great power competition, reshapes security rules, changes the nature of warfare, and even restructures international power dynamics. Driven by the United Nations framework, multilateral mechanisms, regional platforms, and multi-actor coordinated efforts, the current global governance process of AI is progressing. However, due to factors such as the technological characteristics of AI, the law of technology governance, and the complex effects of geopolitics, there remains a significant structural gap between governance supply and demand. Paradoxes encountered in the governance process—including security control versus innovation vitality, transnational coordination versus divergence of interests, national security versus public interest, and ethical consensus versus cultural diversity—pose deep constraints on advancing global governance. Closing these gaps requires new thoughts and strategies. On the premise of maintaining the overall strategic direction of AI for good, practical stage-based goals should be clearly defined, and initiative pragmatic strategies focused on the largest governance commons and leveraging "small cooperation" to drive "large mechanisms," innovative agile governance models embedded in technology, scientifically defined risk thresholds and safety "red lines," and the promotion of a diverse governance system that is resilient, inclusive, and flexible. This will guide AI development along a path that remains directionally correct and provides a solid foundation for sustainable AI development and

* Li Yan, Director and Research Professor of the Institute of Science, Technology and Cyber Security at the China Institutes of Contemporary International Relations (CICIR).

international strategic stability.

Keywords Artificial Intelligence; Technological Evolution; Social Applications; Geopolitics; Global Governance

Introduction

Artificial intelligence, as the core engine of the latest round of technological revolution and industrial transformation, has permeated various fields including politics, economy, military, culture, society, and diplomacy, with its distinctive attributes of autonomous learning, cross-domain application, and self-iteration, profoundly reshaping human production and lifestyle, social operational mechanisms, and the international order. Unlike previous technological revolutions, such as those driven by the steam engine, electricity, or information technology, AI possesses the capabilities of autonomous decision-making, autonomous generation, and autonomous evolution[1]. It breaks through the boundaries of instrumental rationality and has become a disruptive technology capable of reshaping security paradigms, reconfiguring power distribution, and challenging ethical bottom lines. The rapid proliferation and deep application of AI, while driving leapfrog productivity gains and creating unprecedented development opportunities, have also given rise to a series of complex challenges, including the escalation of traditional security risks and the emergence of novel security threats. The call for strengthened global governance of AI has thus grown increasingly urgent within the international community. Existing academic researches on AI governance largely lacks systematic, panoramic, and case-grounded scholarly interpretation of the global supply-demand imbalance, the generative mechanisms of inherent paradoxes, and the logic of pathway transformation. Some studies overly focus on ideal model designs, detached from geopolitical realities and the laws of technological development; others concentrate too narrowly on specific risks, failing to grasp the overall structure of the governance system. In response, this article draws on major governance practices and cases in recent years and follows an analytical logic of "Reality–Paradox–Pathway" to systematically address three core questions: First, what is the current landscape and core supply-demand gap of global AI governance? Second, what are the inherent paradoxes of global AI governance, and how are they generated and how do they affect the governance process? Third, what are the pragmatic and feasible pathways for advancing the global governance of AI in the future?

---

[1] Xue Lan and Zhao Jing, "Artificial Intelligence International Governance: An Analysis Based on Technological Characteristics and Issue Attributes," *International Economic Review,* No. 3, 2024, p. 56.

I. The Current Landscape of Global AI Governance: An Initial Stage Marked by a Significant Supply-Demand Gap

From the perspective of technological evolution, AI originated in the 1940s-1950s and has since experienced several cycles of boom and bust. However, from a governance standpoint, previous waves of AI development did not lead to widespread social application and thus did not give rise to corresponding governance mechanisms. After the second decade of the 21st century, with advances in algorithms, the enhancement of computing power, and the accumulation of massive data, the technological bottlenecks constraining AI were finally overcome, propelling the field into an unprecedented period of rapid development. Frontiers such as generative AI, artificial general intelligence, and multimodal agents have continued to achieve breakthroughs. Technologies have moved swiftly from laboratories to real-world deployment, with commercial, military, and social applications unfolding across the board, formally propelling the global governance process of AI forward.

The year 2017 marked a critical juncture for global AI governance. The International Telecommunication Union (ITU) convened the inaugural AI for Good Global Summit, and the UN Secretary-General officially placed AI on the global governance agenda. In the following years, the United Nations established the High-level Panel on Digital Cooperation, and UNESCO initiated the development of the world's first *Recommendation on the Ethics of Artificial Intelligence*, laying the initial ethical and institutional groundwork for global AI governance. During 2025-2026, generative AI and large model technologies matured further, and social applications expanded comprehensively. More importantly, AI was applied in a practical and systematic manner in geopolitical conflicts such as the Russia-Ukraine war and the U.S.-Israel-Iran conflicts, leading to the concentrated outbreak of various risks, including deepfakes, AI-powered automated attacks, algorithmic discrimination, and the widening of the AI divide. Yet, in contrast to the accelerating pace of AI technological application, global governance mechanisms still largely follow a gradualist, consensus-based, and incremental logic of evolution. The inherent time lag and capability gap between these two dynamics constitute the foundational predicament of global AI governance, and also predetermine its long-term complexity and difficulty.

1. Overall, the Initial Stage of Institutional Building and Principle-Oriented Deliberation

In just a few years, global AI governance has achieved a leap from agenda-setting to institutional building, and from principle advocacy to practical advancement. A multi-layered and pluralistic governance architecture has taken shape, with the United Nations playing a leading role, multilateral and regional mechanisms coordinating with one another, sovereign states providing direction, and a wide array of actors, including technology companies and research institutions, participating extensively. On the supply side, governance efforts have exhibited a momentum of rapid takeoff and parallel progress across multiple fronts.

Within the UN framework, AI governance is endowed with core functions of value guidance, agenda-setting, rule coordination, and multi-party facilitation, positioning it as the most universal, representative, and authoritative platform. Since AI was formally placed on the UN governance agenda in 2017, institutional development has advanced steadily. The UN Secretary-General has repeatedly emphasized, on occasions such as Security Council and General Assembly meetings, the ethical principles of AI governance and the imperative of global cooperation, and has pushed for thematic debates on AI governance in the Security Council, thereby directly linking AI to issues of international peace and security. UNESCO took the lead in developing and securing the adoption of the world's first *Recommendation on the Ethics of Artificial Intelligence*, establishing the first broadly representative ethical framework centered on human-centricity, fairness, non-discrimination, safety and controllability, transparency, and explainability. Other institutions, including the ITU, have contributed to the UN system's governance capacity from the perspectives of technical standards, industrial applications, and sustainable development. In 2025, the UN General Assembly adopted a resolution to establish an independent International Scientific Panel on AI and a Global Dialogue on AI Governance, pushing governance from ethical advocacy toward a more institutionalized phase of scientific assessment, risk monitoring, and multi-stakeholder consultation, which further consolidates the central role of the UN framework.

Multilateral mechanisms and regional organizations have become important pillars of global AI governance. The Group of Seven (G7), the Group of Twenty (G20), the Shanghai Cooperation Organization (SCO), Brazil, Russia, India, China and South Africa (BRICS), and the Association of Southeast Asian Nations (ASEAN) have all placed AI governance on their core agendas, promoting consensus-building and rule alignment through declarations, action plans, and technical guidelines. The G7 was relatively early in issuing AI codes of conduct, focusing on military applications, security risks, transparency, and accountability, and has established dedicated working networks to advance implementation. The G20 has integrated AI into digital economy governance and global development agendas, emphasizing

inclusiveness, risk prevention, and international cooperation. The SCO and BRICS have placed greater emphasis on the shared concerns of developing countries, opposing technological hegemony and bloc confrontation, and advocating a governance orientation of balance, non-discrimination, and mutual benefit. At the regional level, the European Union has established a relatively systematic tiered regulatory system through its AI Act, while Asia-Pacific, African, and Latin American regions have also advanced rule coordination and capacity building through regional cooperation mechanisms, creating a pattern of mutual reinforcement and complementarity between global and regional governance.

New governance platforms have continued to emerge, driven by states and corporate actors alike. The United Kingdom, the Republic of Korea, France, India, and other countries have successively hosted high-level international summits on AI, issuing governance documents on security, controllability, sustainability, and ethical boundaries, in an effort to build new international platforms specifically focused on advancing global AI governance. China has also continuously convened the World Artificial Intelligence Conference (WAIC) and put forward the Global AI Governance Initiative, emphasizing the equal importance of development and security, the balance between innovation and regulation, respect for sovereignty, and inclusiveness and mutual learning, while advocating for a more just and equitable governance architecture. At the same time, global technology companies are rapidly transforming from mere entities of technological R&D and commercial operations into important participants in governance, practitioners of rules, and mitigators of risks, with some even wielding power in governance domains that were once the exclusive preserve of nation-states[1]. Moreover, in the realm of technical security governance, the role of the technical community has become increasingly significant; as a collective, it not only grasps the logic of technology but also transcends national and parochial interests[2]. Currently, there are already more than 50 AI governance-related initiatives, pledges, and self-regulatory norms worldwide. The safety commitments, model testing standards, and ethical review mechanisms of non-state actors, including corporations and the technical community, have been steadily improved, rendering them an indispensable component of the global governance system.

---

[1] Ian Bremmer and Mustafa Suleyman, "The AI Power Paradox: Can States Learn to Govern Artificial Intelligence—Before It's Too Late?" *Foreign Affairs*, August 16, 2023, https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox.

[2] See Lu Chuanying, *Cyberspace Governance and Multi-Stakeholder Theory* (Beijing: Current Affairs Press, 2016), p. 34.

Overall, the institutional framework for global AI governance has been preliminarily established, marking a significant transition from dispersed and spontaneous efforts to systematic advancement, and from technical-circle discussions to high-level political agendas. The progress in institutional building has yielded non-negligible interim results, characterized by a transition and shift from multi-stakeholder processes to multilateral-led processes[1], thereby laying a practical foundation for deepening cooperation, refining rules, and preventing and controlling risks in the future.

2. The Governance Demands of Development and Security Are Growing Simultaneously at a Rapid Pace

With the rapid and comprehensive proliferation and penetration of AI technologies and applications, the demands of development and security have become intertwined and mutually reinforcing. Risks are spilling across borders, competition is intensifying, and divides are widening, all of which place enormous pressure on the governance system. Existing governance capacity, rules, and efficiency fall far short of meeting the actual needs on the ground.

First, from the development perspective, the imperative to bridge the AI divide is urgent. As a general-purpose technology, AI possesses immense potential for inclusive growth and development. It can provide robust support for global public goods in areas such as agriculture, industry, healthcare, education, poverty reduction, and climate change response, making it a key instrument for advancing the UN Sustainable Development Goals. However, current global AI development is characterized by pronounced imbalances, inadequacies, and unsustainability. The gaps between developed and developing countries, between tech giants and small and medium-sized enterprises, and between technology-leading groups and marginalized communities are widening, giving rise to an increasingly severe AI divide. As some experts have pointed out, in the long run, those most deeply affected by automation may not be workers in the United States and other developed countries, but rather developing countries that rely on low-cost labor as their competitive advantage[2]. This divide is reflected not only in the ability to access core enablers such as technology, algorithms, computing power, and data, but also, and more fundamentally, in standard-setting power, rule-making

---

[1] Jia Kai, Yu Hanzhi, and Xue Lan, "Characteristics, Deficits, and Reform Directions of the New Phase of Global AI Governance," *International Forum*, No. 3, 2024, p. 68.

[2] Erik Brynjolfsson and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, trans. Jiang Yongjun (Beijing: CITIC Press, 2016), p. 252.

discourse, industrial dominance, and value-chain distribution. In some cases, it may even give rise to a new North-South problem. Accordingly, the development demands placed on global governance by the international community are concentrated on several key priorities: ensuring the equal participation and development rights of developing countries, facilitating technology and capacity transfer, promoting the equitable allocation of resources, narrowing the intelligence gap, and ensuring that AI genuinely benefits all of humanity. The UN *Global Digital Compact* explicitly identifies AI as a priority area for international cooperation, with the goal of bridging the digital divide and promoting shared development, which reflects the broadest development expectations worldwide.

Second, from the security perspective, various types of risks have already become manifest. The security risks posed by AI are cross-domain, systemic, and difficult to control, spanning both traditional and non-traditional security domains. Politically, AI has become the centerpiece of great-power strategic competition. AI technologies will exert a potentially decisive influence on military power, strategic competition, and global political transformation[1], directly intensifying the weaponization of technology and the politicization of rules, widening trust deficits among states, and undermining international strategic stability. Economically, AI is disrupting employment structures, exacerbating market monopolies, and distorting competition, thereby posing systemic risks to economic systems and industrial chains. Socially, issues such as algorithmic discrimination, deepfakes, privacy breaches, information cocoons, and public opinion manipulation are becoming endemic, eroding social trust, fracturing social consensus, and threatening fundamental rights. Militarily, AI is enabling unmanned operations, cyberattacks, cognitive warfare, and precision strikes, drastically shortening decision-making cycles, lowering the threshold for conflict, and increasing the risk of miscalculation, and make it highly likely that small-scale skirmishes could escalate into large-scale confrontations. Technically, inherent flaws such as AI hallucinations, model escape, adversarial attacks, and emergent risks are difficult to eliminate entirely, carrying with them the potential for uncontrollable, unpredictable, and irreversible consequences at the foundational level. The cross-domain proliferation and rapid evolution of these security risks have made security governance an urgent core task for global governance.

Development demands and security demands are mutually reinforcing and mutually constraining, forcing global AI governance to contend with dual objectives, dual pressures, and dual constraints simultaneously. It must

[1] Fu Ying, "A Preliminary Analysis of the Impact of Artificial Intelligence on International Relations," *International Political Science*, No. 1, 2019, pp. 11–12.

both sustain innovation vitality and development momentum while holding the line on security and ethical boundaries; it must both promote global coordination and rule harmonization while respecting sovereignty differences and varying stages of development; it must both address current manifest risks while anticipating and guarding against long-term latent risks. This highly complex set of demands has further magnified the difficulty and complexity of governance, placing the still-nascent governance system under immense strain and revealing a significant structural gap.

II. The Challenge of Closing the Structural Gap: Governance Paradoxes under the Influence of Multiple Factors

The so-called governance "gap" refers to the inability of existing governance mechanisms to keep pace with, or adequately respond to, rapidly growing governance demands. Concretely, this gap manifests in four dimensions. First, there is a governance mechanism gap: current global AI governance relies predominantly on initiatives, principles, declarations, and self-regulatory norms, while mandatory rules, enforcement mechanisms, oversight systems, and accountability frameworks remain insufficient. Much of the consensus exists only in rhetoric and on paper, lacking binding force and operational effectiveness, and thus failing to effectively constrain high-risk behaviors, cross-border risks, and malicious actions. Second, there is a governance capacity gap: existing capabilities in monitoring, early warning, traceability, response, and remediation are severely inadequate. Faced with emerging risks such as AI-powered automated attacks, deepfakes, malicious fine-tuning of models, and agent escape, the world lacks a unified, efficient, and coordinated response system, and often remains in a state of passive response and ex-post remediation. Third, there is a governance efficiency gap: technological iteration accelerates on the scale of months, weeks, or even days, whereas governance negotiations, rule-making, and institutional development proceed at the pace of years. Governance speed has consistently lagged behind the evolution of risks, generating a persistent governance deficit. Fourth, there is a governance equity gap: developed countries and tech giants occupy a dominant position, while the demands of developing countries are marginalized. Rules are designed to favor the interests of a few actors, making it difficult to reflect the common interests of the global community. The representativeness, fairness, and inclusiveness of governance remain inadequate.

These gaps are not merely the result of insufficient willingness to

cooperate or imperfect institutions. Rather, they stem from the interplay and mutual reinforcement of three deep-seated factors. First, the inherent properties of AI technologies and applications are unprecedented, presenting governance challenges for which there are no precedents, no prior experience, and no ready-made answers. Second, technology governance is inherently subject to the objective laws of lag and irreversibility: innovation speed has consistently outpaced the speed of regulation, relevant measures are difficult to reverse, and the cost of "trial and error" is extremely high. Third, geopolitical competition has fully permeated the governance process: great-power rivalry has intensified, trust deficits have widened, and interest divergences have become more pronounced, steadily eroding the foundation for cooperation. Under the complex interplay of these three factors, global AI governance is confronted with a series of core paradoxes that are mutually contradictory, mutually constraining, and difficult to reconcile. These paradoxes constitute deep-seated constraints that cannot be avoided in the governance process, rendering the supply-demand "gap" in global AI governance not only persistent but also structural.

1. The Multiple Factors Affecting the Governance Process

First, global AI governance is constrained by the inherent attributes of the technology and its applications. AI is the first technological form in human history to possess the capacities for autonomous learning, autonomous reasoning, autonomous generation, and self-iteration. It fundamentally breaks through the instrumental attribute of traditional technologies, which merely execute human commands passively, and thus poses entirely new governance challenges. At the same time, although AI technology is still in its early stages of development, whose evolutionary pathways, capability boundaries, and potential risks impossible to fully predict or control, it has already been rapidly embedded in almost all domains, including the economy, society, national defense, security, and diplomacy. Risks are transmitted across domains, geographies, and actors, with strong chain reactions, wide-ranging impacts, and profound consequences, placing extremely high demands on the agility, coordination, and precision of governance. More importantly, from a governance perspective, new technologies represented by generative AI have a disruptive impact on society, giving rise to a host of issues concerning employment, ethical boundaries, power distribution, responsibility attribution, and security paradigms. Human history offers no ready-made experience, mature model, or fixed answer that can be directly applied; the only way forward is to learn through trial and error, and to refine through practice.

Second, global AI governance is constrained by the objective laws of technology governance. Technology governance generally follows the basic pattern that innovation outpaces regulation and practice precedes rule-making. In the field of AI, this pattern is magnified to an extreme, manifesting in two key characteristics: lag and irreversibility. Lag is reflected in the exponential growth of technological innovation, while governance mechanisms, legal norms, and policy tools can only be updated at a linear pace. Between the two lies an unbridgeable temporal gap, which refers to the classic Collingridge Dilemma: when a technology has not yet been fully applied, humanity cannot accurately foresee all its consequences; when the consequences become fully apparent and risks erupt on a large scale, the technology has already become widely diffused and society has formed path dependencies, making governance intervention extremely costly and difficult. Irreversibility means that once the balance between development and security is broken, it is difficult to correct retroactively; once technological application is fully rolled out, the chains of risk diffusion, patterns of social adaptation, and configurations of interest distribution assume irreversible characteristics. Governance trial and error may trigger large-scale, long-term, and systemic negative effects, making countries more cautious, more hesitant, and less able to reach consensus in regulatory decision-making.

Third, global AI governance is constrained by the complexities of geopolitics. AI has become the central arena for competition in technological strength and comprehensive national power among major powers. Geopolitical competition has fully permeated the governance process, severely undermining the foundation for global cooperation. AI is defined as a strategic high ground and a decisive factor in competition. Major powers are engaged in all-encompassing, high-intensity competition around technology, talent, computing power, data, standards, and supply chains. Governance issues have become highly politicized, camp-based, and even "weaponized," with space for negotiation continuously compressed and trust deficits further exacerbated. Countries are suspicious of and defensive toward one another, making it difficult to build effective mutual trust in key areas such as data sharing, technological cooperation, military AI control, and joint risk assessment. Polarized positions, rule confrontation, and diverging actions have become the norm. Against the backdrop of great-power competition and regional conflicts, countries are more focused on their own security advantages and competitive interests, contributing insufficiently to global public goods, with diminishing motivation for cooperation, and governance has fallen into a predicament. Particularly against the background of certain countries emphasizing first-mover advantages, standard-setting leadership, and the preservation of dominance, the core concerns of the vast number of developing countries

regarding the AI divide have not been effectively addressed, further undermining the representativeness, fairness, and balance of global AI governance.

2. The Four Major Paradoxes That Give Rise to Governance Dilemmas

Under the cumulative effect of the above deep-seated factors, global AI governance faces four major core paradoxes that are highly challenging and directly constrain the effective advancement of the governance process.

The first is the paradox of security control versus innovation vitality. This is the most direct and universal contradiction in global AI governance. AI is fundamentally driven by sustained technological innovation: the faster the pace of innovation, the broader the scope of application, and the stronger the industrial vitality, the more significant the gains in social productivity, but at the same time, the more prominent, complex, and intractable the derived security risks become. Whether viewed from the perspective of domestic governance practices or global governance, all actors face a dilemma. If security control is overemphasized, with strict restrictions, prior approval requirements, and comprehensive prohibitions, innovation vitality may be directly suppressed, industrial development hindered, and technological dividends missed, placing countries at a disadvantage in global competition. If the pursuit of innovation speed and commercial interests is overemphasized, with risk prevention relaxed, ethical constraints weakened, and safety thresholds lowered, the result may be loss of risk control, endangerment of public safety, disruption of social order, undermining of strategic stability, and even existential risks. During 2025-2026, global large models accelerated their iteration, and productivity effects continued to manifest—but at the same time, issues such as AI-powered automated cyberattacks, cross-border deepfake fraud, algorithmic discrimination, and agent escape erupted intensively, fully demonstrating the real-world pressure of this paradox. The EU AI Act adopts a tiered and stringent regulatory approach, which, while enhancing safety and controllability, has also been criticized for potentially suppressing innovation and industrial competitiveness among small and medium-sized enterprises. The United States, for a time, adopted a relatively loose federal regulatory model to encourage rapid technological breakthroughs, but has also faced widespread criticism for risk laissez-faire and the accumulation of hidden dangers. Finding a dynamic balance between holding the security bottom line and maintaining innovation vitality has become the primary challenge that global governance must address.

The second is the paradox of transnational coordination versus

divergence of interests. This is the core and most intractable contradiction in global AI governance. From the perspective of technological attributes and risk characteristics, AI industrial chains, supply chains, data flows, and risk transmission are inherently global in nature. No single country, regardless of its technologically advanced and capacity, can accomplish effective governance on its own; transnational coordination, rule alignment, information sharing, and joint prevention and control are indispensable. However, from the perspective of real-world interest configurations, countries differ enormously in their stages of development, industrial foundations, security concerns, value systems, and strategic objectives. Interest demands are highly fragmented, making it difficult to form unified, efficient, and robust global norms. Developed countries are more focused on market expansion and dominance in standard-setting and rule-making. Developing countries are more concerned with development rights, technology acquisition, capacity building, fair treatment, and bridging the AI divide. Tech giants focus on commercial interests, innovation space, compliance costs, and global market access. Civil society focuses on privacy protection, social equity, ethical bottom lines, and the public interest. In recent years, multiple high-level global summits on AI have revealed pronounced differences in positions, with final outcomes largely consisting of voluntary guidelines and declaratory statements, lacking mandatory binding force and thus unable to effectively address cross-border, cross-domain, and systemic risks. The global public goods nature of AI stands in sharp conflict with the realist logic of the international system, in which national interests take precedence and competition is prioritized, and this constitutes a fundamental obstacle to transnational collaborative governance.

The third is the paradox of national security versus the global public interest. This is the most sensitive and conflict-prone contradiction in global AI governance. AI has extremely strong national security implications, involving critical infrastructure security, military security, data security, strategic deterrence, and the self-reliance and controllability of supply chains. All countries place national security at the top of their governance priorities, and this position has both rationality and legitimacy. However, when national security demands are excessively emphasized, broadly interpreted, and politically instrumentalized, they can easily trigger techno-nationalism, leading to supply chain fragmentation, data barriers, standard confrontation, and camp polarization, ultimately undermining the global innovation ecosystem, common security, and the public interest. Conversely, if national security red lines are completely abandoned, and technology is allowed to diffuse, be

misused, and militarized without restriction, the consequences may include cognitive manipulation, mass surveillance, asymmetric attacks, and strategic imbalances, endangering international peace and stability. In 2026, the U.S. Department of War (Department of Defense) placed the AI company Anthropic on a restricted list on national security grounds. The reason was that the company refused to make its models available for autonomous lethal weapons and mass surveillance systems. This incident highlights the sharp conflict between national security demands and corporate autonomy, technological ethics, and the global public interest. How to achieve a balance between safeguarding legitimate national security interests and promoting the global public interest is a major challenge that global AI governance must confront.

The fourth is the paradox of ethical consensus versus cultural diversity. This is the deepest and most long-term value contradiction in global AI governance. AI ethical governance calls for universal principles such as human-centricity, AI for good, fairness and justice, safety and controllability, and non-discrimination. The international community has already reached a certain level of consensus at the macro level, as exemplified by the broad support for UNESCO's *Recommendation on the Ethics of Artificial Intelligence*. However, the specific interpretation of ethical principles, their prioritization, institutional translation, and implementation standards are deeply influenced by history, culture, political systems, social structures, and value systems. Significant differences exist among countries with different civilizations, institutions, and stages of development, making it difficult to form unified, enforceable, and monitorable global ethical standards. Western value systems place greater emphasis on individual rights, privacy primacy, algorithmic transparency, and autonomous decision-making. Eastern cultures place greater emphasis on social well-being, the public interest, overall security, and harmony and stability. On sensitive issues such as autonomous weapons systems, deepfake applications, social credit evaluation, and the boundaries of data use, perceptions of ethical limits differ markedly across cultures and institutional contexts. The tension between this universal ethical aspiration and the reality of cultural diversity means that global ethical governance is prone to a situation in which principled consensus is relatively easy to reach, but the understanding and interpretation of key terms diverge, constituting a value-level constraint that has long constrained the governance process.

III. Pathways Forward for Global AI Governance: Pragmatic Choices under Realistic Expectations

In the face of severe real-world gaps and deep-seated internal paradoxes, global AI governance must abandon unrealistic expectations of idealization, one-step solutions, and comprehensive coverage. It must adhere to the overall direction of pragmatic rationality, gradual progression, and focused prioritization, seeking the optimal feasible path under the current geopolitical circumstances. Broadly speaking, in the period ahead, efforts should focus on three major tasks: building consensus around the "largest common denominator" to establish a foundation of trust; empowering the entire governance process through technological and conceptual innovation, while continuously improving institutional development; and holding the bottom line with safety "red lines" to avoid extreme scenarios.

1. Moderating Governance Expectations and Seeking the "Largest Common Denominator"

The issues in global AI governance are highly diverse, interests are highly fragmented, and contradictions are highly complex. Attempting to advance all issues simultaneously, reach comprehensive rules at one stroke, and achieve perfect governance immediately is neither realistically possible nor practically feasible. Therefore, global governance should, on the premise of keeping the overall direction unchanged, articulate and establish more realistic and attainable stage-by-stage goals, accumulating long-term trust through incremental achievements and achieving continuous improvement through dynamic adjustment.

First, focus on the largest "common denominator," that is, prioritize issues on which there is the most consensus, the greatest urgency, and the greatest technical feasibility, avoiding the quagmire of highly contentious issues that are difficult to resolve in the short term. Priority areas for cooperation should concentrate on: risk classification and grading standards for AI, along with globally shared early warning mechanisms; explainability, transparency, and safety assessment frameworks for large models; data security, privacy protection, and rules for responsible data use; technical cooperation on deepfake identification, traceability, and countermeasures; prevention and control of AI-enabled cyberattacks and coordinated emergency response; and human oversight of autonomous weapons systems and minimum ethical bottom lines. These issues combine practical urgency, a basis for international consensus, and technical operability, making them suitable as early-stage breakthrough areas for cooperation that can quickly produce visible results and strengthen international confidence in the

governance process.

Second, leverage "small cooperation" to drive "large mechanisms." Adopt a pragmatic strategy of tackling the easy before the difficult, technology before rules, and pilot projects before broader rollout. AI technological development and application are inherently global in nature, and the problems it faces are also largely universal. Practices in which AI benefits society should be actively and rapidly shared with the international community[1]. Bilateral, minilateral, technical, and non-sensitive small-scale cooperation should be encouraged, forming demonstration projects around model testing, risk assessment, standard alignment, and information sharing, and creating replicable, scalable, and interoperable best-practice sharing. Flexible mechanisms such as digital partnerships and joint research centers should be supported to lower the threshold for cooperation and expand participation. The role of non-state actors, including technology companies, research institutions, and industry associations, should be fully leveraged to carry out non-governmental governance cooperation, generating bottom-up momentum for governance and providing experiential support and social foundations for official institutional development.

Third, harness the "radiating" effect of a core governance circle. Priority should be given to forming a relatively stable core governance circle among countries with a high degree of consensus, strong foundations for cooperation, and complementary interests. This circle should pioneer information sharing, standard alignment, joint testing, and emergency response, establishing effective collaboration models and promoting best practices. Through practical cooperation results, mutual trust can be enhanced, experience accumulated, and influence expanded, gradually attracting more countries and actors to participate and continuously broadening the representativeness and coverage of governance. Governance objectives should remain stage-based, dynamic, and adjustable, not aiming for one-step perfection, but continuously optimized in response to technological evolution, risk changes, and adjustments in the international landscape, achieving a spiral ascent of the governance system.

2. Strengthening Technology Empowerment: A New Model of Full-Lifecycle Agile Governance

[1] Nicolas Miailhe and Cyrus Hodes, "Making the AI revolution work for everyone," *The Future Society*, March 13, 2017, https:// thefuturesociety.org/wp-content/uploads/2019/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf.

To overcome the challenges of lag, rigidity, and low efficiency in AI governance, it is necessary to move beyond the single path of traditional administrative regulation and legal prescription, fully leverage the technological advantages of AI itself, and innovate a technology-embedded agile governance model–governing technology with technology, responding to risks with intelligence, and driving the governance system to resonate with technological iteration through full-lifecycle governance.

First, highlight the enabling role of technology in governance. Use AI technologies and applications themselves to resolve the information dilemmas, prediction dilemmas, and traceability dilemmas inherent in traditional governance. Build a global AI risk intelligent monitoring network that leverages large models, multimodal recognition, and anomaly detection technologies to identify in real time hidden and high-speed risks such as malicious fine-tuning, model escape, deepfakes, automated cyberattacks, and large-scale public opinion manipulation. Establish a global risk prediction and scenario simulation platform to conduct forward-looking assessments of emergent risks, systemic risks, and cross-border interconnected risks, improving early warning precision and response timeliness. Improve intelligent traceability and forensic systems to achieve end-to-end traceability, evidential admissibility, and accountability for AI-generated content, cyberattacks, data breaches, and algorithmic harms, providing technical support for responsibility attribution. Through technology empowerment, governance can be elevated from passive response and ex-post remediation to proactive sensing, early warning, and precision response.

Second, apply agile governance processes. The core challenge is to overcome the lag, rigidity, and slow response of traditional governance. Drawing on agile development concepts, build governance mechanisms for rapid sensing, flexible decision-making, dynamic adjustment, and continuous iteration, shortening the cycle of rule updates to adapt to rapid technological evolution. Promote flexible regulatory tools such as sandbox regulation, pilot-first approaches, gray testing, and interim guidelines, allowing new technologies and applications to be tested and verified within controllable boundaries, balancing innovation and security. Simplify multi-level decision-making processes and establish high-risk rapid response channels to issue timely guidance and transitional norms for new risks and new forms of misuse. Establish mechanisms for dynamic evaluation and optimization of rules, regularly adjusting regulatory requirements in light of practical outcomes, technological progress, and risk changes, avoiding institutional rigidity and regulatory failure.

Third, establish the concept of full-lifecycle embedding. Emphasize

embedding safety, compliance, and ethical requirements from the source and throughout the entire process. Drive the pre-embedding of governance requirements into the full lifecycle of AI, including R&D, data collection, model training, testing and verification, deployment, operation and maintenance, and decommissioning, truly achieving design for safety, development for compliance, deployment for control, and application for oversight. Implement ex-ante systems such as algorithm registration, model safety testing, ethical review, and risk assessment. Develop standardized compliance tools, detection tools, and protection tools to reduce compliance costs for small and medium-sized enterprises. Establish full-process recording, auditing, and documentation mechanisms to ensure that risks are traceable, manageable, and trackable, achieving source prevention, full-process control, and systematic risk mitigation.

3. Drawing Safety Red Lines: Scientifically Defining Risk Thresholds

In the face of the catastrophic, systemic, and existential risks that frontier AI may bring, the international community cannot remain at the level of voluntary initiatives and principled declarations. It must, on the basis of scientific assessment, clearly define intolerable risk thresholds, draw “rigid” safety red lines, and establish trigger-based, mandatory control mechanisms to safeguard humanity’s common security bottom line.

First, continuously reaffirm the reality and urgency of extreme risks and accelerate the formation of a minimum global consensus. During 2025-2026, hundreds of top AI scientists, corporate executives, and authoritative scholars issued repeated public appeals to prioritize the mitigation of AI extinction risks as a global priority, calling for binding international agreements on the control of high-risk AI. The international community should discard wishful thinking and myopic competition, acknowledge that breakthroughs in frontier model capabilities may bring uncontrollable risks, and establish a shared understanding that risk prevention must precede harm and that red lines are non-negotiable, thereby laying the political and social foundations for drawing safety red lines.

Second, scientifically define risk thresholds and tiered control standards. Building on documents such as the Frontier AI Safety Commitments, and based on risk severity, threat scope, controllability, and potential consequences, delineate red lines and yellow lines and implement differentiated controls. Red lines correspond to intolerable risks, those with credible pathways to mass casualties, systemic societal collapse, paralysis of critical infrastructure, or existential threats. Once triggered, R&D and

deployment must be immediately suspended, third-party verification initiated, the highest level of control implemented, and the chain of risk diffusion blocked. Yellow lines correspond to high-alert risks—models exhibiting anomalous capabilities, boundary-crossing behaviors, or risk tendencies, triggering comprehensive assessment, enhanced monitoring, upgraded mitigation measures, and restricted application scenarios to prevent further escalation.

Third, develop clear prohibition and restriction lists for high-risk scenarios to implement safety red lines in practice. The international community should reach corresponding international norms and international law on the issue of the AI arms race as soon as possible, avoid excessive development of AI military weapons, control the safety and proliferation of AI weapons, raise the threshold for their use, and prevent the abuse of AI weapons from challenging the international security system[1]. High-risk scenarios subject to priority control include: autonomous weapons systems must retain ultimate human decision-making authority, and autonomous lethal attacks without human supervision must be prohibited; the use of AI to autonomously carry out large-scale destructive cyberattacks must be prohibited; AI must be prohibited from being used to design, synthesize, or proliferate high-risk biological agents or technologies related to weapons of mass destruction; AI must be prohibited from being used for large-scale political deception, identity forgery, public opinion manipulation, and social division; and AI must be prohibited from engaging in self-replication, self-iteration, or cross-system autonomous privilege escalation without human approval. Through clear, specific, and enforceable prohibitive provisions, insurmountable boundaries should be drawn for the most dangerous AI applications.

IV. Conclusion: Historical Rationality and the Responsibility of the Era

Exploring governance paradoxes and real-world predicaments, and proposing pragmatic stage-by-stage goals and strategic choices, does not imply any abandonment of the goal of orderly and effective global governance. On the contrary, it is precisely through the philosophy that "a journey of a thousand miles begins with a single step" via steadily achieving small milestones that global AI governance can be progressively refined toward a pluralistic, open, inclusive, fair, balanced, and collaboratively efficient ecosystem. The ultimate objective is to build a new type of global

[1] Heather Roff, *Meaningful Human Control or Appropriate Human Judgment? The Necessary Limits on Autonomous Weapons* (Geneva: Global Security Initiative, Arizona State University, 2016), p. 3.

governance system that is effectively adapted to the data-intelligence era, reconciles the interests of all parties, and safeguards humanity's common well-being. In this process:

First, it is essential to uphold historical rationality and acknowledge the long-term, gradual, and trial-and-error nature of governance. History demonstrates that the global governance of major emerging technologies, such as the steam engine, electricity, and the internet, has invariably gone through complex processes of prolonged exploration, divergent bargaining, partial cooperation, and institutional refinement. These are stage-by-stage, uneven processes of development. AI governance is even more sensitive, more complex, and more global in scope; it cannot be achieved overnight. It demands the patience of long-termism, a constructive approach to differences, pragmatic action to advance cooperation, and continuous improvement of rules and mechanisms through practice.

Second, it is necessary to improve multi-layered and pluralistic governance mechanisms and promote concerted efforts between formal and informal arrangements. Strengthen the agenda-setting of global AI governance and raise the level of international attention to AI issues[1]. Consolidate the universality, authority, and leading role of the UN framework, and strengthen the coordination and synergy among platforms such as the G20, UNESCO, and the ITU. Support diverse explorations including regional mechanisms, minilateral arrangements, and industry self-regulation, forming a governance architecture that combines top-down and bottom-up approaches, and complements official and non-official efforts. Avoid the monopolization of governance discourse by any single actor, encourage the equal participation and complementary strengths of diverse actors, and enhance overall governance effectiveness.

Third, it is necessary to enhance inclusiveness and representativeness, and effectively safeguard the rights of developing countries to equal participation and benefit-sharing. Focus on addressing the prominent issues of governance inequity, rule hegemony, and the AI divide. Safeguard the voice and representation of developing countries in agenda-setting, rule-making, and oversight and implementation. Increase technical assistance, capacity building, and investment in computing and data infrastructure for developing countries, helping them improve their governance capacity and development capabilities. Respect differences in culture, institutions, and stages of development, encourage mutual learning among diverse ethical and governance practices, and promote the construction of a more just and

[1] International Telecommunication Union, *AI for Good Global Summit Report* (Geneva: ITU, 2017), p. 7.

equitable governance order.

Fourth, it is necessary to deepen multi-stakeholder coordination and build a co-governance ecosystem in which governments, enterprises, academia, civil society, and international organizations all pull in the same direction. Clarify the roles and responsibilities of each party: governments should fulfill the functions of top-level design, rule-making, public interest protection, and national security maintenance; technology companies should assume primary responsibility and embed safety and ethics throughout the entire technology and product lifecycle; research institutions should provide independent scientific assessment, risk early warning, and technical support; international organizations should undertake the functions of coordination, dialogue, oversight, implementation, and the provision of global public goods; and civil society should participate in ethical oversight, public education, and social constraints. Through multi-party coordination, clear rights and responsibilities, and positive interaction, a powerful governance synergy can be formed.

Fifth, it is essential to stay committed to the direction of AI for good and harness broad consensus to guide collective global action. Despite differences and competition, the international community has already formed an irreversible broad consensus that AI should benefit humanity, be safe and controllable, fair and inclusive, and ethically sound. The 2023 report *Governing AI for Humanity*[1] issued by the UN High-level Advisory Body on Artificial Intelligence, together with the "human-centric" concept and the "AI for good" principle advocated by China, are both aimed at ensuring that AI remains under human control, and at developing AI technologies that are auditable, overseen, traceable, and trustworthy. In the future, building on this consensus, it is necessary to strengthen the alignment of initiatives, coordination of rules, and synergy of actions among all parties, gradually forming a globally accepted, inclusive, and enforceable governance framework. This will enable AI to better serve global sustainable development, international peace and security, and humanity's common well-being, truly realizing the vision of AI for good, co-governance and shared benefits.

[1] United Nations High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Interim Report* (United Nations, 2023), p. 30.

The Dual Tracks of International AI Governance Debates under the UN Framework: The Cases of LAWS Arms Control Negotiations and the Global Digital Compact

*Xu Peixi**

Abstract The processes involved in AI international governance are complex and interwoven. There are many related aspects with the existing Internet governance, digital governance, and cyber disarmament negotiation processes. The AI governance dialogue within the UN framework has not been limited to a single track but follows military and non-military threads. This article takes the LAWS disarmament process and the Global Digital Compact negotiations as examples. The article observes that the LAWS disarmament negotiations have in recent years attracted heavy attention and achieved many results, but the views of all countries are seriously polarized on issues such as whether to sign a treaty, how to apply international laws such as International Humanitarian Law, whether to expand the scope of negotiations, and whether to start a new negotiation track. In the area of non-military AI governance, the United Nations has established legitimacy by negotiating and agreeing on the Global Digital Compact, which requires the United Nations to play an important role in AI governance. China's many propositions in the field of international governance of non-military AI governance are inextricably linked to texts such as the Global Digital Compact under the United Nations framework. The future rule-making processes need to take into consideration AI capabilities of different countries, distinguish between military and non-military dimensions, analyze the similarities and differences

* Xu Peixi, Director and Professor of the Center for Global Cyberspace Governance, Communication University of China.

between Internet governance and AI governance, coordinate domestic governance and international governance, clarify the roles of various stakeholders of governments, businesses and civil society, and formulate governance strategies in a cross-ministry and multistakeholder manner. This article records the LAWS disarmament process and the Global Digital Compact negotiations as two specific scenarios about military and non-military AI governance under the United Nations framework, and how China articulates international positions linked to these processes.

Keywords UN Framework; AI International Governance; LAWS Disarmament; Global Digital Compact

Introduction

In recent years, artificial intelligence has accelerated its permeation into military, intelligence, disinformation, healthcare, automotive, finance, education, and other domains, becoming another complex and intertwined issue since the advent of the Internet. It possesses an unprecedented degree of complexity, interconnectedness, comprehensiveness, and cross-cutting relevance, touching extensively upon national security, social cognition, and economic competitiveness.

In military and intelligence fields, AI-driven Lethal Autonomous Weapon Systems (LAWS) have emerged as transformative national security technologies comparable to nuclear weapons, aircraft, computers and biotechnology, which are reshaping the correlations between military strength, population scale and economic prowess, while disruptive swarming technologies bring fatal risks to high-end combat platforms including aircraft carriers. “Suppose a low-cost drone possessed the capabilities of a Canada goose, able to travel 1,500 miles non-stop at 60 miles per hour for 24 hours. When millions of those drones assemble into swarms advance toward aircraft carrier battle group to carry out a kamikaze-style strikes, how should countermeasures be deployed?”[1] What this scenario reveals is not merely the lethality of a new type of weapons platform, but the possibility that AI could transform traditional military superiority through low-cost, distributed, swarming, and autonomous means.

AI is changing the means of lethality on the battlefield and the ways in

[1] Belfer Center for Science and International Affairs, “*Artificial Intelligence and National Security*,” July 2017, https://www. belfercenter.org/publication/artificial-intelligence-and-national-security.

which states identify and control risks. In May 2026, eight AI companies, including SpaceX, OpenAI, Google, NVIDIA, Microsoft, Amazon, and Oracle, signed contracts with the U.S. Department of War to deploy AI capabilities onto the U.S. military's highest-level classified networks. The widespread application of AI technology has greatly enhanced the surveillance capabilities of national security agencies. In the traditional era, the East German Ministry for State Security employed 102,000 intelligence officers to monitor 17 million people, with an average of 166 people per officer. In the data-intelligence era, the U.S. National Security Agency can monitor the digital activities of billions of people worldwide with only a few thousand employees.

In disinformation terms, intelligent technologies have been used to mass-produce and disseminate fake news and fake comments, and have greatly transcended previous racial, cultural, and linguistic barriers in image, audio, and video production, increasing the efficiency of information dissemination at an exponential rate. As early as 2017, propaganda researchers had already coined the term "computational propaganda" to define this new phenomenon. This concept emphasizes the technological attributes of propaganda, analyzing large-scale databases to target audiences and using bots and algorithms to disseminate propaganda information. A research team from the Oxford Internet Institute (OII) at the University of Oxford published a report analyzing politically relevant social media posts from China, the United States, Russia, Germany, Canada, Brazil, Poland, and Ukraine during 2015-2017, concluding that the role of bots in influencing public opinion and shaping political situations could no longer be underestimated[1].

In economic terms, AI is empowering a vast array of industries, becoming deeply embedded in key sectors such as manufacturing, agriculture, healthcare, finance, education, and transportation, reshaping industrial structures and employment patterns. Combined with previously accumulated advantages in manufacturing, green and low-carbon industries, and digital industries, AI has given rise to a diverse range of new industries and business forms, significantly enhancing and locking in the global competitiveness of countries with complete supply chains, such as China. In 2023, the total digital economy of the five largest digital economies, including the United States, China, Germany, Japan, and South Korea, exceeded $33 trillion, with the digital economy accounting for an average of 60% of GDP. The United States, as the largest digital economy, had a ratio

[1] Oxford Internet Institute, "*COMPROP Case Study Launch* (June 19: London, UK)," June 2017, https://www.oii.ox.ac.uk/blog/ comprop-case-study-launch-june-19-london-uk/.

near that level[1]. In 2025, China, as the second-largest digital economy, had a digital economy scale of approximately 49 trillion yuan, accounting for about 35% of GDP. The National Development and Reform Commission stated that the "AI+" initiative is empowering China's vast scenario advantages, and that AI is accelerating its transition from the digital world to the physical world, driving explosive growth in high-end manufacturing, emerging consumption, and new business forms and models[2].

Against this backdrop, the processes involved in AI international governance are complex and interwoven, with multiple points of intersection and connection with existing Internet governance, digital governance, and cyber arms control negotiation processes. The debate on military AI governance under the UN framework is primarily concentrated in the LAWS arms control negotiation process. Since the LAWS arms control issue entered the mechanism of the Convention on Certain Conventional Weapons in 2013, states and stakeholders have been debating issues such as the definition of LAWS, whether to conclude a legally binding international treaty, the applicability of international law including International Humanitarian Law, human control, and attribution of responsibility.

By contrast, the debate on non-military AI governance under the UN framework follows a different institutional logic. The Global Digital Compact provides governance principles and guiding objectives for global digital technology development, covering key areas such as bridging the digital divide, data governance, and AI governance. The United Nations believes it can advance the AI governance process through the Global Digital Compact. This article takes the LAWS arms control process and the Global Digital Compact negotiations as case studies, focusing on the debates in two specific scenarios under the UN framework concerning military and non-military AI governance, as well as their intersection and alignment with domestic positions, in order to more accurately assess the UN's actual role and institutional boundaries in AI international governance.

I. Military AI International Governance: The LAWS Arms Control Negotiation Process

Military AI governance concerns national security, military advantage,

---

[1] China Academy of Information and Communications Technology (CAICT), *2024 Global Digital Economy White Paper*, July 2, 2024.

[2] Su Deyue, "China's Digital Economy to Account for About 35% of GDP in 2025," *People's Posts and Telecommunications News*, January 23, 2026.

and humanitarian ethics. For AI powers, prematurely formulating and accepting legally binding restrictions may undermine their capability in future intelligent warfare. For weaker states and many civil society groups, failure to prohibit or strictly regulate early on may lead to indiscriminate killing, difficulties in accountability, and erosion of the principles of *International Humanitarian Law*. The issue of military AI governance is not purely a technical matter; it involves multiple dimensions including military strength, legal responsibility, and national security.

In the field of military AI international governance, differences in national positions are mainly concentrated in the UN LAWS arms control negotiation process. Since 2013, the polarization of views among states and stakeholders on LAWS arms control within the UN framework has been extremely pronounced, far exceeding the diversity of positions in the cyber arms control negotiations conducted by the UN Group of Governmental Experts on Information Security (UN GGE) and the Open-ended Working Group on Information Security (OEWG) since 1998. On issues such as whether to conclude a legally binding international treaty, the definition and classification of LAWS, governance principles, and whether to establish a new track beyond the LAWS Group of Governmental Experts, the positions of states and stakeholders differ fundamentally. The negotiation process for LAWS arms control under the UN framework has roughly passed through two phases, namely informal expert meetings and Group of Governmental Experts sessions, and may split into two distinct mechanisms in the future: negotiations under the Group of Governmental Experts established by the Conference of States Parties to the Convention on Certain Conventional Weapons (CCW), and a separate negotiation track outside the CCW framework.

1. From Informal Expert Meetings to a Group of Governmental Experts: The Institutionalization of the LAWS Issue

On 14-15 November 2013, the Meeting of the States Parties to the CCW was held in Geneva, Switzerland. The meeting decided to establish an informal Meetings of Experts on LAWS mechanism to discuss the governance of emerging technologies related to LAWS. The full title of the CCW is the *Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*. The convention entered into force in 1983, is of unlimited duration, and has been acceded to by over 120 states. Its purpose is to prohibit or restrict the use of specific types of weapons deemed to cause unnecessary or unjustifiable suffering to combatants or to have indiscriminate effects on civilians, such as mines, booby traps, incendiary weapons, and laser

blinding weapons. In 2013, the LAWS issue was placed on the agenda of this meeting. The significance of this institutional entry point is that LAWS was not treated as a general issue of technological ethics, but was placed within the framework of conventional weapons restrictions and arms control norms, thereby determining that subsequent discussions would always revolve around the balance between military necessity and humanitarian constraints.

From early 2014 to early 2016, three consecutive informal expert meetings on LAWS were held, each producing a report. The number of participating states (including observer states) grew from 87 in 2014 to 94 in 2016, a modest increase, but non-governmental participants increased substantially from 28 to 53, with NGOs, academia, industry, and civil society all actively engaged. The growth in non-governmental participation indicates that the LAWS issue was not confined to national military establishments and arms control diplomats, but involved all stakeholders and global issues concerning ethics, law, technology, and public security.

From 13 to 16 May 2014, the first informal expert meeting on LAWS initially explored issues such as the definition of LAWS and dual-use issues, without provoking intense disagreement. From 13 to 17 April 2015, at the second informal expert meeting on LAWS, participants held widely divergent views. Most states believed that the full potential of AI in various aspects had not yet been realized and thus should not be strictly regulated, but some states expressed deep concern over LAWS and advocated a prohibitive approach. They argued that LAWS violated the fundamental principles of *International Humanitarian Law*, and that the international community should immediately conclude a binding treaty to comprehensively prohibit the development, acquisition, trade, and deployment of such weapons. The Non-Aligned Movement (NAM) broadly held this position:

> LAWS are inherently contrary to ethics and morality, lack human judgment and compassion, increase the covertness of military operations, violate International Humanitarian Law, exacerbate asymmetries in warfare, complicate tracing and punishment, generate new proliferation risks, and may lead to a new arms race. LAWS may undermine regional balance and global strategic stability, hinder progress in disarmament and non-proliferation, lower the threshold for initiating military action, facilitate escalation, fall into the hands of non-state actors, and increase the risk of terrorist attacks[1].

[1] "*Fifth Review Conference of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*," December 23,

The divergence at this stage already revealed the fundamental tension in LAWS governance: one side emphasized that the technology was not yet mature and its potential not yet fully released, and therefore should not be prematurely constrained; the other side stressed that once the technology matured and proliferated, governance would become more difficult. The former reflects a logic of "regulating while developing," while the latter reflects a logic of "preventive prohibition."

From 11 to 15 April 2016, the third informal expert meeting on LAWS emphasized the importance of non-governmental participation in the discussion, and recognized the applicability of international law, including International Humanitarian Law and international human rights law, in this domain. Although some states preferred to continue discussing the issue informally, the majority supported transitioning to a formal mechanism, establishing "an open-ended Group of Governmental Experts" to define LAWS, discuss measures to enhance transparency and mutual confidence, and develop legal principles and rules applicable to LAWS governance.

From 12 to 16 December 2016, the Fifth CCW Review Conference was held in Geneva. The conference adopted the recommendation of the informal expert meetings and formally approved the establishment of a Group of Governmental Experts on LAWS, the LAWS GGE, to discuss emerging technologies related to LAWS. This new expert group mechanism was more open and transparent in its operational procedures than the UN GGE on information security in the cyber domain, with participation open to all states upon registration and approval. Thus, the LAWS issue completed its transition from informal expert discussions to intergovernmental institutionalized negotiations.

2. The LAWS GGE Phase: From Principle-Oriented Deliberation to Core Divergence

From 13 to 17 November 2017, the first session of the LAWS Group of Governmental Experts was convened in Geneva. In advance of the meeting, the Chair of the LAWS GGE, Amandeep Singh Gill, stated that although discussions on LAWS had been transferred to a formal mechanism, this did not constitute the commencement of formal negotiations on LAWS. The session sought to present a comprehensive perspective on LAWS governance. Prior to the meeting, the United Nations Office for Disarmament Affairs (UNODA) issued a report entitled *Perspectives on Lethal Autonomous Weapon Systems* to provide a foundation for the GGE, with five authors

2016, https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/117/16/pdf/G1611716.pdf?OpenElement.

examining LAWS from legal, security, technical, ethical, and military perspectives[1]. Chair Gill, drawing on these dimensions, put forward 27 sub-questions in advance of the meeting. The final report recommended that future attention be directed toward "definition of LAWS, information-sharing, issuance of political declarations, conclusion of binding treaties, and suspension of LAWS deployment." These issues essentially covered the main contentious aspects of LAWS governance, indicating that the GGE phase was not merely a continuation of informal discussions, but rather began to incorporate technical definitions, legal pathways, and policy tools into the agenda.

On 13 November 2017, the first day of the inaugural LAWS GGE session, the Campaign to Stop Killer Robots screened a roughly eight-minute film titled *Slaughterbots* at a side event, depicting the horrifying uses of killer drones. These drones, equipped with wide-angle cameras, facial recognition, special-purpose sensors, and shaped explosives, could execute missions at speeds a hundred times faster than humans, and swarms could easily annihilate an entire city's population. The film depicted a doomsday scenario and went viral globally, making the campaign widely known. The Campaign to Stop Killer Robots is a coalition of sixteen civil society groups, including Amnesty International, Article 36, Human Rights Watch, and the International Committee for Robot Arms Control (ICRAC). Participants in the movement actively tracked national positions, reporting that 22 states had explicitly expressed support for banning the development and deployment of LAWS. These were primarily developing and less powerful states, including Algeria, Argentina, Bolivia, Brazil, Chile, Costa Rica, Cuba, Ecuador, Egypt, Ghana, Guatemala, the Holy See, Iraq, Mexico, Nicaragua, Pakistan, Panama, Peru, Palestine, Uganda, Venezuela, and Zimbabwe[2]. The involvement of civil society reinforced the humanitarian and ethical dimensions of the LAWS issue and helped disseminate prohibitive discourse in international public opinion. However, public mobilization clearly could not directly translate into intergovernmental arms control rules, as what ultimately determines whether a treaty can be concluded remains states' calculations of military advantage, technological risk, and legal liability.

From 9 to 13 April 2018, the second session of the LAWS GGE was held in Geneva. The disagreements among states and stakeholders centered

[1] United Nations Office for Disarmament Affairs, *Perspectives on Lethal Autonomous Weapon Systems*, UNODA Occasional Papers No. 30 (New York: United Nations, November 2017), https://digitallibrary.un.org/record/3929837/files/op30.pdf.
[2] "Country Views on Killer Robots," November 2017, http://www.stopkillerrobots.org/wp-content/uploads/2013/03/KRC_CountryViews_16Nov2017.pdf.

on two main issues: whether to conclude a treaty, and the definition of LAWS and the applicability of international law. First, should there be a treaty to prohibit and regulate LAWS? The Non-Aligned Movement and most civil society groups supported this proposition. The NAM position paper stated: "The LAWS GGE should initiate discussions on the establishment of a legally binding international instrument to prohibit and regulate LAWS." "Official declarations, codes of conduct, weapons review procedures, and expert committees, these proposals, while numerous, lack legal binding force and cannot substitute for the objective of concluding a legally binding treaty[1]."

China's position was more balanced and open, at that time taking a comprehensive view of the application of emerging technologies such as AI in both military and economic domains. On the one hand, China stated that LAWS lowered the threshold for war and made it difficult to distinguish soldiers from civilians, and therefore such weapons should be used prudently and with restraint until reasonable solutions could be found. On the other hand, China emphasized that AI had been widely applied in economic and social development, and called for objective, fair, and comprehensive assessment of emerging technologies, cautioning against preconceived biases and against hindering the development of AI technologies[2]. This position reflected a dual consideration in military AI governance: acknowledging the potential security and ethical risks posed by LAWS, while avoiding the stigmatization of military AI applications as a whole or prematurely closing off space for technological development.

The second contentious issue was the definition of LAWS and the applicability of international law. Russia argued that the definitions of LAWS provided by various states had become increasingly divergent, complicating the Group's deliberations. Russia stated that the definition of LAWS should be categorized, describing various weapon categories that fall under LAWS, while avoiding qualitative judgments of good or bad that cater to the political preferences of individual states; it should maintain flexibility and not be constrained by current understandings of LAWS; it should be of general applicability, comprehensible to scientists, engineers, technicians, military personnel, lawyers, and ethicists; and it should not be rushed to conclusion, so as to avoid impeding technological progress and undermining

[1] UNODA, CCW_GGE_1_2018_WP.1.pdf, submitted 28 March 2018, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts(2018)/CCW_GGE_1_2018_WP.1.pdf
[2] UNODA, CCW_GGE.1_2018_WP.7.pdf, submitted 11 April 2018, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts(2018)/CCW_GGE.1_2018_WP.7.pdf

existing research on peaceful robotics and AI[1]. The Russian position highlighted the constraints that technological uncertainty places on arms control negotiations, namely that when weapons configurations are not yet stabilized and technological boundaries remain unclear, definition itself becomes an important tool for states to preserve negotiating space and shape rule-making.

The central theme of the U.S. position paper was that the application of emerging technologies in the LAWS domain brings numerous benefits to humanity, such as reducing civilian casualties and aiding the wounded in armed conflict. The United States maintained that these emerging technologies could better protect civilians in conflict. "Smart weapons, deployable through computer and automated functions, can more precisely and effectively reduce the risk of harm to civilians." With respect to the applicability of international law, the United States argued that the principles of distinction and proportionality under *International Humanitarian Law* apply to the LAWS domain. "In counterterrorism operations in Iraq, U.S. commanders sought to minimize damage to infrastructure controlled by terrorists in order to facilitate post-war reconstruction. Weapons have become more precise and effective, balancing military effectiveness with humanitarian protection requirements." The United States also identified terms such as "self-destruct," "self-deactivation," and "self-neutralization" from existing international legal texts, emphasizing that these functions could also be realized in new weapon systems such as LAWS[2]. The U.S. position emphasized that existing international law and technological improvements were sufficient to address risks, thereby opposing the compression of future military application space through preventive prohibition.

China's position paper argued that LAWS should refer to "fully autonomous lethal weapon systems", characterized by five features: lethality, autonomy, impossibility for termination, indiscriminate killing, and evolution. With respect to the applicability of international law, in contrast to the U.S. position, China questioned the applicability of traditional principles such as distinction and proportionality under *International Humanitarian Law* in the LAWS domain, arguing that new scenarios presented many uncertainties. First, it remained unknown whether such weapon systems

[1] UNODA, CCW/GGE.1/2018/WP.6_E, submitted 4 April 2018, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.6_E.pdf.
[2] UNODA, CCW/GGE.1/2018/WP.4, submitted 28 March 2018, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.4.pdf.

possess the capability to distinguish targets. Second, such weapon systems would find it difficult to make proportionate and commensurate decisions. Third, once such weapons are deployed, attribution of responsibility would be challenging[1].

At the 2019 GGE session, the Group achieved a milestone by reaching agreement on eleven important guiding principles, including: *International Humanitarian Law* continues to apply fully to all weapon systems, including the possible development and use of LAWS; human beings must remain accountable for decisions on the use of weapon systems, and responsibility cannot be transferred to machines, this must be taken into account throughout the entire lifecycle of weapon systems; any possible policy measures discussed and adopted within the framework of the *Convention on Certain Conventional Weapons* should not hinder the advancement of autonomous technology or its peaceful use; and the *Convention on Certain Conventional Weapons* provides an appropriate framework for addressing issues related to new technologies in the LAWS domain within the scope of the Convention's objectives and purposes, seeking to strike a balance between military necessity and humanitarian considerations[2]. These principles constituted a consensus on LAWS governance, but also revealed the boundaries of that consensus: parties could accept principled expressions such as "human accountability," "applicability of *International Humanitarian Law*," and "non-hindrance of peaceful use," but had not reached agreement on the scope of prohibition, legal form, or enforcement mechanisms.

3. The Evolution of China's Position: Differentiated Governance, Opposition to a New Arms Race, and Support for the CCW Platform

Subsequently, under the UN *Convention on Certain Conventional Weapons* platform, the LAWS GGE sessions largely continued to be held twice annually. During 2020-2021, the LAWS GGE convened four sessions in a hybrid format combining small-scale in-person meetings with online participation (September 2020, August 2021, September 2021, December 2021). From 2022 to 2025 (March and July 2022, March and May 2023, March and August 2024, March and September 2025),

[1] UNODA, CCW_GGE.1_2018_WP.7.pdf, submitted 11 April 2018, https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts(2018)/CCW_GGE.1_2018_WP.7.pdf

[2] Cited in Cao Huayang, Li Xiang, Kuang Xiaohui, and Zhang Gang, "The Arms Control Process of Lethal Autonomous Weapon Systems," in *2022 International Arms Control and Disarmament* (Beijing: World Knowledge Press), pp. 203–204.

consultations continued at a pace of two sessions per year, with discussions on the issue growing deeper.

During this period, three position papers submitted by China merit particular attention. In December 2021, China submitted to the Sixth Review Conference of the *Convention on Certain Conventional Weapons* a document entitled *China's Position Paper on Regulating Military Applications of Artificial Intelligence*, which put forward recommendations in eight areas, including strategic security, military policy, legal ethics, technical security, R&D operations, risk management, rule-making, and international cooperation, regarding the responsible development and use of AI technologies within the military domain. The paper was updated in January 2024[1].

In July 2022, China submitted to that year's LAWS GGE session a working paper entitled *China's Working Paper on the Issue of Lethal Autonomous Weapon Systems*. On the question of categorizing autonomous weapon systems, the paper stated that parties should consider dividing autonomous weapon systems into "unacceptable" and "acceptable" categories, prohibiting the former and regulating the latter, so as to ensure that the relevant weapon systems are safe, reliable, and controllable, and comply with *International Humanitarian Law* and other applicable international law. The paper argued that "unacceptable" autonomous weapon systems should satisfy, but not be limited to, five basic characteristics: lethality, autonomy, impossibility for termination, indiscriminate killing, and evolution[2].

In November 2022, China submitted *China's Position Paper on Strengthening Ethical Governance of Artificial Intelligence*, stating that "as the most representative disruptive technology, artificial intelligence, while bringing tremendous potential development dividends to human society, also carries uncertainties that may give rise to many global challenges and even fundamental ethical concerns." It further argued that "at the ethical level, the international community is generally concerned that without proper regulation, the misuse and abuse of AI technologies could undermine human dignity and equality, violate human rights and fundamental freedoms, exacerbate discrimination and prejudice, and impact existing legal systems, with far-reaching implications for government administration, national

---

[1] Arms Control Department of the Ministry of Foreign Affairs of the People's Republic of China, *China's Position Paper on Regulating Military Applications of Artificial Intelligence*, updated January 2024.

[2] UNODA, "Working paper on Lethal Autonomous Weapons Systems submitted by the People's Republic of China," July 2022, https://meetings.unoda.org/ccw/convention-certain-conventional-weapons-group-governmental-experts-2022.

defense, social stability, and global governance." China accordingly put forward its own propositions in four areas: regulation, R&D, use, and international cooperation[1]. This indicates that China's position on LAWS is not merely an arms control position, but is also connected to broader AI ethical governance and global governance propositions.

Furthermore, China's position on LAWS arms control is clearly reflected in the statements made by China's Ambassador for Disarmament Shen Jian at the first session of the LAWS GGE in March 2024 and in China's submission to the UN Secretary-General on the LAWS issue in May 2024. These can be summarized in five respects[2]. First, with regard to human control: the development of AI should comply with applicable international law, and all states, especially major powers, should adopt a prudent and responsible attitude toward the development and use of AI technologies in the military domain, ensuring that AI remains under human control at all times. Second, with regard to preventing a new arms race: the LAWS issue should be addressed in accordance with the principles of equal security, common security, and universal security, balancing humanitarian concerns, military necessity, and the impact on strategic balance and stability, and opposing the use of LAWS as a tool for waging war or seeking hegemony. Third, with regard to definition and categorization: China advocates the prohibition of unacceptable autonomous weapon systems. Fourth, with regard to the legal applicability of LAWS: China argues that in light of technological development trends, the relevant legal issues require thorough research and assessment, and that considerable uncertainty remains as to whether *International Humanitarian Law* is sufficient to meet the challenges posed by LAWS. Fifth, with regard to the forum for discussing LAWS issues: China considers the CCW platform to be the most appropriate venue. By that time, the UN General Assembly had already adopted Resolution A/RES/78/241 on LAWS, requesting the UN Secretary-General to seek opinions and submit a report, thereby expanding the scope of discussion and accelerating the pace, with the venue for discussing LAWS issues having already expanded beyond the CCW platform.

II. Non-Military AI International Governance: The *Global Digital Compact* Establishes UN Authority

Non-military AI governance focuses more on development discourse and

---

[1] *China's Position Paper on Strengthening Ethical Governance of Artificial Intelligence*, November 17, 2022.

[2] Ministry of Foreign Affairs of the People's Republic of China, *China's Position Paper on Strengthening Ethical Governance of Artificial Intelligence*, May 23, 2024.

development rules, paying attention to the imbalance in AI development worldwide and identifying platform mechanisms for global AI dialogue. Non-military AI governance can easily be subsumed under issues such as the digital divide, capacity building, technology diffusion, platform monopolies, data governance, and the representation of developing countries. Within the framework of the *Global Digital Compact*, AI is not discussed as a single security risk, but rather as part of the global digital development agenda, situated in the context of inclusive development, fair participation, and capacity building. This issue setting provides a legitimacy foundation for UN involvement in AI governance.

The UN's legitimacy in non-military AI governance is primarily achieved through the negotiation and formulation of the *Global Digital Compact*. The *Global Digital Compact* covers multiple areas, including the digital divide, governance of Internet technical resources, digital human rights, cross-border data flows, online content governance, and AI governance, all of which are key issues relating to national security, economic development, global innovation, and digital trade. Many of China's propositions in the field of non-military AI international governance both balance global community interests and national interests, and are closely intertwined with texts such as the *Global Digital Compact* under the UN framework.

1. Integrating Digital Governance and AI Governance: The Agenda Expansion of the Global Digital Compact

UN Secretary-General António Guterres has defined the digital transformation issue and the climate crisis as two major issues of the 21st century. First, in September 2020, the General Assembly adopted the *Declaration on the Commemoration of the Seventy-Fifth Anniversary of the United Nations*, which stated that digital technologies have profoundly transformed society, bringing unprecedented opportunities and new challenges, and that if improperly or maliciously used, they can exacerbate divisions within and among states, increase insecurity, undermine human rights, and exacerbate inequality. It affirmed that the world depends more than ever on digital tools for connectivity and socio-economic prosperity. To this end, Secretary-General Guterres, in his report *Our Common Agenda*, proposed strengthening the UN's role in digital governance, forming a "multi-stakeholder digital technology track" involving the UN, governments, the private sector, and civil society, with the aim of finalizing the *Global Digital Compact* at the 2024 UN Summit of the Future, to outline common digital principles for all.

The process of the *Global Digital Compact* went through several twists

and turns, undergoing multiple revisions from zero draft through first, second, third, fourth, and fifth drafts, before finally being adopted by consensus at the UN General Assembly. Since the dominant economic problem in the digital world is the global monopoly of U.S. technology and digital platforms, the *Global Digital Compact* appears to address the issue of bridging the digital divide, but in practice has evolved into a process in which states seek ways to address the current landscape of U.S. digital monopolies. The *Global Digital Compact* reenacts the dynamics of the 2003-2005 UN World Summit on the Information Society (WSIS), which was originally intended to address development issues such as bridging the digital divide, but later evolved into dissatisfaction with U.S. dominance over Internet governance via the Internet Corporation for Assigned Names and Numbers (ICANN). This eventually, compounded by factors such as the Snowden revelations, led the United States in 2016 to terminate its contract with ICANN and globalize oversight of the IANA functions. This historical continuity is worth emphasizing: from WSIS to the *Global Digital Compact*, development issues are often not merely development issues in themselves, but also transform into debates over digital infrastructure resources, platform power, and the distribution of global governance authority. The entry of AI governance into the *Global Digital Compact* continues this logic.

2. The Battle over Governance Authority: From the Breaking of the Silence Procedure to the Formulation of "an Important Role"

The United States was particularly dissatisfied with the provisions of the *Global Digital Compact* concerning online content governance, data flows, and AI governance. Originally, member states had already approved the third draft of the *Global Digital Compact* on 11 July 2024, and from the following day it entered the "silence procedure," under which, if no member state broke the silence within 72 hours, the third draft would be adopted by global consensus. However, after the third draft was issued, Australia and other states broke the silence procedure, demanding amendments on data flows and AI governance. After a moderately revised fourth draft was issued, the Group of 77 and China (G77+China) and other Global South states expressed dissatisfaction with the overly broad human rights language in the fourth draft, and also broke the silence procedure, demanding further revisions. Some language was ultimately deleted, and a fifth and final draft was reached in early September 2024 and adopted by consensus. The fact that the silence procedure was broken twice indicates that the *Global Digital Compact* is not a technical consensus document, but rather a political negotiation centered on data flows, digital human rights, AI governance, and Global South representation.

On the issue of AI international governance, the U.S. administration and business community were generally unwilling to support UN involvement in AI governance through the *Global Digital Compact*. By contrast, the G77+China and other Global South states generally supported the UN playing a central role in global digital governance and rule-making, and supported expanding the UN's role in the AI field to ensure that all states fully enjoy the rights to technological development and peaceful use, and share in the benefits of AI technologies. The final outcome of the debate on whether the UN should play a role in AI governance through the *Global Digital Compact* was neither the U.S.-favored position of "play no role or little role," nor entirely the G77+China-favored position of "play a central or leading role," but rather a compromise position that leaned significantly toward the Global South: "play an important role." In other words, "an important role" is not ordinary wording, but rather a political compromise forged between the pressure of technologically powerful states and the demands of the Global South. It affirms that the UN must not be sidelined from AI governance, while refraining from framing the UN as the sole or dominant global platform for AI governance.

The zero draft of the *Global Digital Compact* fully replicated the official position of the Global South, stating: "We recognize the UN's central role in supporting international AI governance (zero draft, paragraph 47, 1 April 2024)." The first draft changed "central role" to "a vital role" (first draft, paragraph 47, 15 May). The second draft changed it to "a critical role" (second draft, paragraph 52, 26 June). The third, fourth, and final fifth drafts changed it to "an important role." This demonstrates the intensity of the contest among states and stakeholders in this domain. This textual evolution from "central role" to "important role" can be seen as the detail in the *Global Digital Compact* that best reflects the competition over AI governance authority. "Central role" would mean that the UN should become the primary institutional center for international AI governance; "vital role" and "critical role" still carry a strong connotation of institutional priority; while "important role" is a compressed form of authority recognition.

The term "non-military" first appeared in paragraph 49 of the zero draft of the *Global Digital Compact*, referring to the promotion of non-military AI safety standards. By the third draft and the final version, the Compact had completely disengaged from military AI governance, explicitly stating that "the Global Digital Compact sets out some objectives, principles, commitments, and courses of action applicable to the non-military domain". This change indicates that the *Global Digital Compact* actively avoids all military issues, positioning itself as a non-military AI governance framework rather than an alternative to the LAWS arms control negotiations. It addresses

not military issues, but rather development, capacity, safety standards, inclusive application, and governance participation.

3. China's Core Propositions in the Field of Non-Military AI International Governance

In the field of non-military AI international governance, China is replicating and updating many of its successful experiences from the Internet domain, focusing on inclusive development, equitable participation, and institution-building. China's positions on non-military AI international governance are mainly reflected in the *Global AI Governance Initiative*, the "AI Capacity Building for All Initiative," the *Global AI Governance Action Plan*, and the initiative to establish a World AI Cooperation Organization, all issued during 2023-2025. Together, these documents and initiatives constitute China's basic policy package for non-military AI governance: opposing technological monopolies and exclusionary blocs, emphasizing capacity building for developing countries, and advocating for the UN as the main channel for promoting inclusive and fair multilateral governance.

In October 2023, the Cyberspace Administration of China issued the *Global AI Governance Initiative*. Among the 11 propositions put forward, three aspects merit particular attention. The first extends the principle of non-interference in internal affairs from the digital domain to the AI domain, stating opposition to "using AI technological advantages to manipulate public opinion, disseminate false information, interfere in other countries' internal affairs, social systems, and social order, or undermine other countries' sovereignty". The second concerns supply chain security: opposition to "drawing lines based on ideology or forming exclusionary blocs to maliciously obstruct other countries' AI development", and opposition to "using technological monopolies and unilateral coercive measures to create development barriers and maliciously disrupt the global AI supply chain." The third concerns institution-building: advocating for "enhancing the representation and voice of developing countries in global AI governance, ensuring equal rights, equal opportunities, and equal rules for all countries in AI development and governance, and carrying out international cooperation and assistance for developing countries to continuously bridge the AI divide and governance capacity gaps[1]." These three aspects correspond respectively to content security, development rights, and institutional participation, reflecting a policy logic that combines AI governance with opposition to technological monopolies and the defense of developing country

[1] Cyberspace Administration of China, *Global AI Governance Initiative*, October 18, 2023.

representation.

In September 2024, in order to implement the UN General Assembly resolution on "International Cooperation on Capacity Building of Artificial Intelligence" (A/RES/78/311) and promote the implementation of the UN 2030 Agenda for Sustainable Development, China's Ministry of Foreign Affairs launched the "AI Capacity Building for All Initiative," proposing five long-term goals: promoting connectivity in AI and digital infrastructure; advancing "AI+" to empower all industries; strengthening AI literacy and talent development; enhancing AI data security and diversity; and ensuring AI is safe, reliable, and controllable. The "AI Capacity Building for All Initiative" has been promoted and publicized at multiple forums, including the UN, BRICS meetings, and the Forum on China-Africa Cooperation[1]. This indicates that China does not understand non-military AI governance merely as risk regulation, but rather places it squarely within the framework of capacity building, infrastructure connectivity, and the sustainable development agenda.

In July 2025, China issued the *Global AI Governance Action Plan* at the World Artificial Intelligence Conference and High-Level Meeting on Global AI Governance, calling on all parties to take effective action, on the basis of adhering to the goals and principles of benefiting the people, respecting sovereignty, development-orientation, safety and controllability, fairness and inclusiveness, and openness and cooperation, to jointly advance global AI development and governance. The *Global AI Governance Action Plan* includes 13 aspects, of which Article 11 advocates "actively implementing the relevant commitments of the UN *Pact for the Future* and its annex the *Global Digital Compact*, adhering to the UN as the main channel, with the goal of helping developing countries bridge the digital divide and achieve fair and inclusive development, and on the basis of complying with international law and respecting national sovereignty and development differences, promoting the construction of an inclusive and fair multilateral global digital governance system[2]." At the same conference, China proposed the establishment of a World AI Cooperation Organization, envisioning it as an important international public good that would achieve three goals: "deepening innovative cooperation and unleashing the dividends of intelligence," "promoting inclusive development and bridging the AI divide," and "strengthening collaborative governance and ensuring AI for good." It explicitly stated that it would "follow the purposes and principles of the UN Charter, support the UN in playing a central role in AI governance, and

[1] Ministry of Foreign Affairs of the People's Republic of China, *AI Capacity Building for All Initiative*, September 27, 2024.

[2] Ministry of Foreign Affairs of the People's Republic of China, *Global AI Governance Action Plan (Full Text)*, July 26, 2025.

provide useful supplementation to the efforts of the UN and its relevant agencies[1]."

III. Structural Differences and Institutional Reference Points of the Dual Tracks in AI Governance under UN Framework

The preceding sections have respectively traced the AI international governance processes in the two scenarios of the LAWS arms control negotiations and the *Global Digital Compact*. It can be observed that UN framework AI governance does not revolve around a single issue, a single mechanism, or a single rule objective, but rather has formed different governance tracks along the two axes of military security and non-military development. An analysis of these two tracks helps both to understand the institutional role that the UN can play in AI governance and to grasp its practical boundaries.

1. The Formation of the Dual Tracks

The processes involved in AI international governance are complex and interwoven, with multiple points of intersection and connection with existing Internet governance, digital governance, and cyber arms control negotiation processes. The AI governance debate under the UN framework has not proceeded along a single track, but rather along two lines: military and non-military. The LAWS arms control negotiations and the *Global Digital Compact* respectively constitute two important scenarios under the UN framework involving military AI governance and non-military AI governance.

In the field of military AI governance, the LAWS arms control negotiations have in recent years attracted considerable attention and have yielded significant results. Since the LAWS issue entered the mechanism of the *Convention on Certain Conventional Weapons*, parties have continuously discussed issues such as the definition of LAWS, the applicability of international law, human control, attribution of responsibility, and military necessity, thereby bringing military AI governance into the UN institutional agenda. At the same time, the LAWS arms control negotiations have long been marked by significant divergences, with severe polarization among states and stakeholders on issues such as whether to conclude a treaty, how to avoid a new arms race, how international law including *International Humanitarian*

[1] Xinhua News Agency, "Chinese Government Proposes Establishment of World AI Cooperation Organization," July 26, 2025.

Law should apply, whether to expand the scope of negotiations, and whether to open new negotiation tracks. Nevertheless, military AI governance, as represented by LAWS arms control, has formed a relatively stable dominant dialogue track and accumulated considerable dialogue experience. Given the high risks and destructive nature of such weapons systems, the absence of rules would be tantamount to losing control of a vehicle's steering wheel. Therefore, under the United Nations framework, member states are expected to form binding rules on key issues such as the deployment of nuclear weapons.

In the non-military governance domain, the *Global Digital Compact* has established a legitimacy foundation for UN participation in digital and AI governance. Originating from the *Declaration on the Commemoration of the Seventy-Fifth Anniversary of the United Nations*, and adopted by global consensus at the UN General Assembly in September 2024 after multiple rounds of textual revisions, the Compact has become one of the broadest global consensus documents in the field of digital cooperation to date. Its significance lies not in establishing a globally enforceable AI regulatory mechanism, but rather in affirming that the UN plays an "important role" in relevant governance and in providing new institutional space for states, especially developing countries, to participate in digital and AI governance. Although the United States has long been reluctant to let the UN dominate digital governance, it has found it difficult to completely bypass the UN's universal platform and the representativeness of the Global South. From this perspective, the *Global Digital Compact* is not merely a digital governance text, but also provides UN framework legitimacy resources for subsequent AI governance initiatives, capacity-building programs, and the provision of international public goods.

2. The Asymmetry of the Dual Tracks

The LAWS arms control negotiations and the *Global Digital Compact* together constitute the dual tracks of UN framework AI governance, but they are not fully symmetrical institutional processes. LAWS has long been embedded in the arms control framework, while the *Global Digital Compact* emerged from the digital governance process. The two differ in their temporal positioning, subject matter, stakeholder demands, and institutional objectives, and therefore should not be simplistically compared on the same scale.

In terms of temporal progression, since the LAWS issue entered the *Convention on Certain Conventional Weapons* mechanism in 2013, it has undergone long-term discussions through informal expert meetings and the Group of Governmental Experts, forming an arms control negotiation process lasting over a decade. It has achieved breakthroughs such as the "11 Guiding

Principles" on the major and complex multidimensional issue of LAWS arms control, and has developed systematic and in-depth understanding of military AI governance. However, due to the deterioration of the international security environment in recent years and the accelerated application of AI in the military domain, weaker developing countries, some European states, and civil society groups remain dissatisfied with the pace of negotiations on LAWS arms control under the CCW mechanism. It is possible that in the future the process may split into two mechanisms: GGE negotiations under the CCW and a non-CCW dialogue track. The *Global Digital Compact*, by contrast, originated from the *Declaration on the Commemoration of the Seventy-Fifth Anniversary of the United Nations* adopted in September 2020, was subsequently advanced through the *Our Common Agenda* and Summit of the Future processes, and after multiple rounds of revision was adopted by global consensus at the UN General Assembly in September 2024. The *Global Digital Compact* has a much shorter history, having been formed as a digital governance consensus text within a specific political window. The two differ significantly in their timeframes, negotiation rhythms, and levels of institutional accumulation.

In terms of issue attributes, the LAWS arms control negotiations address military AI issues, directly involving highly sensitive topics such as weapon systems, battlefield lethality, and the threshold of war. The relevant discussions bear on national security, military advantage, and future warfighting configurations, and their governance is more deeply influenced by security logic and arms control logic. Whether comprehensive and strongly binding rules can ultimately be formed depends primarily on the political will of major military powers such as the United States, Russia, and China, and on whether catastrophic events occur in this domain. The AI governance provisions in the *Global Digital Compact* fall mainly within the non-military governance category, focusing on digital cooperation, development opportunities, capacity building, and the UN's role in global digital governance. They reflect a stronger development and governance orientation, making it easier to form a certain level of consensus through language on inclusiveness, capacity building, and fair participation. The future trajectory depends on whether states and stakeholders have sufficient will and capacity to promote the implementation of relevant principles.

In terms of interest structures, the main contradiction in the LAWS arms control negotiations is between military powers seeking to establish and maintain future operational advantages and the numerous weaker states and civil society groups demanding preventive restrictions. For military powers, prematurely accepting legally binding restrictions may undermine their capability in future intelligent warfare. For weaker states and civil society

groups, failure to prohibit or strictly regulate LAWS early on may lead to a lower threshold for war and indiscriminate killing. The contradiction between the two is dialectically characterized by the tension between "force" and "reason"–"force" referring to power and capability, "reason" referring to moral justification and ethics. Many weaker states still hope to prevail through reason and moral persuasion, but find it difficult to shake the deeply entrenched ambition of military powers such as the United States to pursue military advantage and hegemony. This profound divergence does not mean that states cannot reach agreement and conclude treaties on issues of vital importance, because even great powers need rules to avoid major risks and miscalculation. To some extent, this situation also provides an opportunity for reason, ethics, and the concept of peace to prevail. The contradictions in the *Global Digital Compact* revolve more around digital development rights, governance participation rights, and rule discourse power. Although the United States and other digital powers, platform companies, developing countries, civil society, and technical communities have different demands, these can be indirectly addressed through actions such as digital development, capacity building, and the sharing of technological dividends.

In terms of institutional objectives, the LAWS arms control process directly confronts questions of the form, scope, and binding force of arms control rules, including whether to conclude a legally binding international treaty, how to understand the enormous risks posed by autonomous weapon systems and the necessity and urgency of rule-making, whether to prohibit or restrict certain categories of autonomous weapon systems, and how international law should apply. The objective of the *Global Digital Compact* is not to formulate a dedicated AI treaty, but rather to confirm the UN's important role in AI governance through a comprehensive digital governance framework, to provide principled direction and a legitimacy foundation for non-military AI governance, and to utilize the "International Scientific Panel on AI" and the "Global Dialogue on AI Governance" to promote the implementation of relevant governance principles. Thus, the progress of LAWS is mainly reflected in issue setting, principle sedimentation, and rule incubation, while the significance of the *Global Digital Compact* lies mainly in legitimacy construction, agenda integration, and consensus formation.

3. Comparing the AI Dual Tracks with the Internet Governance Process

The debate on AI governance under the UN framework has not yet achieved results comparable to those attained by the UN in the domains of cyber arms control and Internet governance.

At the military level, comparing with cyber arms control, the LAWS arms control negotiation process has replicated many of the issues faced in cyber arms control negotiations: for example, whether to conclude a treaty; whether a treaty would in practice help prevent abuse or prove counterproductive by legitimizing arms races and facilitating war initiation; how international law, including *International Humanitarian Law*, should apply; whether to expand the scope of negotiations; whether to open new negotiation tracks; and whether the goal should be anti-weaponization or anti-arms race. From 1998 to the present, cyber arms control negotiations have continued for 28 years and have achieved considerable success, and a global mechanism dialogue process is now underway. The AI arms control negotiations, originating in 2013, have entered their 13th year. In recent years, due to the exponential progress of AI technology and its broader application in war and conflict, compounded by the fact that the military-industrial complexes of some countries have identified this domain as a new source of profit growth, international attention to LAWS governance has suddenly increased, and the debate on autonomous weapon systems governance under the UN framework has accelerated. International public opinion generally holds that AI weapons are more dangerous than cyber weapons, since AI weapons are not merely lethal weapons but systems capable of manufacturing other deadly weapons, and the demand for rule-making in this domain is therefore more urgent. In this sense, although AI arms control started later than cyber arms control, its risk perception is stronger, and it may attract even more attention in the future than cyber arms control.

At the non-military level, the UN, through the outcome documents of the World Summit on the Information Society (WSIS, 2003 and 2005) and the WSIS+20 review (2025), has shaped the principles and framework of Internet governance, ensuring for over two decades a single, interconnected global Internet and supporting the digital economic development of all countries. In July 2005, the UN, through the establishment of the Working Group on Internet Governance (WGIG), put forward a definition of Internet governance: "the development and application by governments, the private sector, and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet." It also identified 18 Internet public policy issues, including domain name allocation. Unlike Internet governance, the UN has not yet provided a specific definition of AI governance in terms of its content, scope, principles, mechanisms, participating actors, or detailed issue areas. The *Global Digital Compact* has established the UN's legitimacy in AI international governance, but this is only a beginning. The outcome documents on Internet governance and information society governance formed under the UN framework are more idealistic in tone, embody the spirit of globalization, and better respect the

vision of a community with a shared future. But by the AI era, the global geopolitical and technological landscape has undergone enormous changes, and many of these concepts have become fragmented and difficult to sustain.

Overall, the Internet governance process provides an important reference point for understanding UN AI governance: on the one hand, it shows that the UN can gradually form concepts, principles, and mechanisms through long-term dialogue; on the other hand, it reminds us that AI governance faces more pronounced security risks, higher technological thresholds, and more intense great-power competition, making it difficult to fully replicate the established pathways of Internet governance. Although the LAWS arms control negotiations and the *Global Digital Compact* have respectively opened institutional entry points for military and non-military AI governance, it will still take much longer, namely more negotiation, accumulation, and institutional shaping, to form a stable rules system comparable to that in Internet governance.

Conclusion

AI governance under UN framework has not proceeded along a single track, but has formed different processes in the military and non-military directions. The LAWS arms control negotiations embody the security logic and arms control challenges of military AI governance, have greatly promoted issue institutionalization and principled consensus formation, and continue to face fundamental divergences on issues such as whether to conclude a treaty, the applicability of international law, negotiation mechanisms, and forms of constraint. Nevertheless, due to the high risks of autonomous weapon systems, the negotiations and debates in this domain under the UN framework have attracted widespread attention and are likely to achieve new breakthroughs. The *Global Digital Compact*, by contrast, embodies the development logic and legitimacy construction of non-military AI governance, providing new institutional space for UN participation in AI governance via issues such as the digital divide, capacity building, technological inclusiveness, and Global South representation.

Therefore, in its future participation in UN AI governance, China should fully assess the AI capabilities and governance philosophies of various countries, distinguish between military and non-military dimensions, summarize the differences between the Internet governance process and the AI governance process, coordinate domestic and international governance, clarify the roles of governments, enterprises, and civil society, and formulate governance strategies in a cross-ministry and multi-stakeholder manner. On this basis, China needs to take a more proactive role in influencing the AI

governance agenda within the UN framework. It will advance AI governance in a safer, more inclusive, equitable and sustainable manner, safeguard its own legitimate development rights and interests, and put forward institutional solutions with greater inclusiveness and representativeness for global AI governance.

Research Trends of Artificial Intelligence Governance by Think Tanks in the West

Lang Ping Zhang Mengxi[*]

Abstract Nowadays, artificial intelligence (AI) is profoundly reshaping national security, international competition, and the global landscape. Based on an analysis of AI-related research publications by seven influential think tanks in the West between January 2023 and April 2026, this study delineates a multidimensional analytical framework, encompassing social impact and domestic regulation, national strategy and great power competition, military application and security risk, industrial transformation and market restructuring, as well as international cooperation and global governance. Overall, think tanks in the West are adopting an increasingly pragmatic and policy-driven approach to integrating AI into national and global agendas. Their research trends can reveal the evolution of policy thinking patterns, while providing a critical lens for deciphering the future dynamics of technological competition and global governance.

Keywords Think Tank; Artificial Intelligence; Great-Power Competition; National Security; Global AI Governance

Introduction

As a general-purpose enabling technology, artificial intelligence has strategic significance that extends far beyond the economic and social development dimensions. It has become a key force shaping national security, comprehensive national power, and the construction of international order, and has rapidly evolved into a major arena for great-power strategic competition. On this important strategic issue, Western think tanks have conducted extensive research in recent years, publishing numerous important research reports and findings that have provided analytical support for policy-making in

[*] Lang Ping, Senior Fellow and Director of Security Studies Division, Institute of World Economics and Politics, Chinese Academy of Social Sciences; Zhang Mengxi, Ph.D. candidate, Class of 2024, School of Global and Regional Studies, University of Chinese Academy of Social Sciences.

Western countries and have also served as an important window for observing and analyzing their policy trajectories. Against this backdrop, this article takes well-known Western think tanks as its research subjects, conducts continuous tracking and systematic analysis of their AI-related research outputs, and focuses on examining research themes, major viewpoints, and policy recommendations, with the aim of drawing lessons and insights for China.

I. Think Tank Sample Selection

The purpose of this article is to explore the influence of think tank reports on the AI policy directions of Western countries by analyzing those reports. Accordingly, this paper comprehensively considers multiple factors and selects think tanks that possess strong international influence, high policy engagement, and robust professional research capabilities, while also taking regional complementarity into account. It ultimately selects seven think tanks in the west: the Brookings Institution, the Carnegie Endowment for International Peace, the Center for Strategic and International Studies (CSIS), the RAND Corporation, the Center for Security and Emerging Technology (CSET), the Stanford Institute for Human-Centered AI (Stanford HAI), and Bruegel. The specific details are shown in Table 1.

Table 1 Rankings and Countries of Sample Think Tanks

No.	Think Tank Name	*2020 Global Go to Think Tank Index Report* Ranking	Country
1	Brookings Institution	1 (Foreign Policy and International Affairs) 1 (Best AI Policy and Strategy Think Tanks)	United States
2	Carnegie Endowment for International Peace	2 (Foreign Policy and International Affairs)	United States
3	Center for Strategic and International Studies (CSIS)	4 (Foreign Policy and International Affairs)	United States
4	RAND Corporation	9 (Foreign Policy and International Affairs)	United States
5	Center for Security and Emerging Technology (CSET)	3 (Best AI Policy and Strategy Think Tanks)	United States
6	Stanford Institute for Human-Centered AI (Stanford HAI)	6 (Best AI Policy and Strategy Think Tanks)	United States
7	Bruegel	2 (TOP THINK TANKS WORLDWIDE, US and non-US)	Belgium

Source: James G. McGann, *2020 Global Go to Think Tank Index Report*, Philadelphia: Think Tanks and Civil Societies Program, University of Pennsylvania, 2021.

First, all selected think tanks possess strong international influence. According to the *2020 Global Go to Think Tank Index Report* published by the University of Pennsylvania's "Think Tanks and Civil Societies Program," the Brookings Institution, Carnegie Endowment for International Peace, CSIS, and RAND Corporation all rank among the top ten think tanks in the field of foreign policy and international affairs research, with RAND also ranking fourth in the science and technology policy think tank category. According to the report's ranking of global think tanks engaged in AI policy and strategy research, both CSET and Stanford HAI rank in the top ten. Bruegel ranks first in the comprehensive list of top non-U.S. think tanks, reflecting its important position in European and global policy research.

Second, the selected think tanks all have relatively high levels of policy engagement. The Brookings Institution, Carnegie Endowment for International Peace, and CSIS have long maintained close relationships with U.S. policy circles, and their research findings are frequently used in policy formulation, implementation, and evaluation processes. RAND Corporation has close ties with the U.S. military, and successive *U.S. National Security Strategy* reports have contained content overlapping with its research findings. CSET actively promotes its researchers' entry into the "revolving door" through project-based engagement, with its researchers serving as fellows in the U.S. Department of State, Department of Homeland Security, thereby channeling the think tank's research outputs to key decision-makers. Stanford HAI has put forward multiple policy recommendations around the U.S. update of its national AI R&D strategic plan, and its research on AI ethics and policy has influenced relevant U.S. government legislation. As a top European think tank in the field of international economic research, Bruegel's research findings on AI policy and economic impacts have been used as policy references by European and other countries and international organizations.

In addition, the selected think tanks also have strong professional research capabilities and regional complementarity. In terms of research output, all seven think tanks are able to leverage interdisciplinary and cross-sectoral professional knowledge platforms to continuously publish research findings on technological ethics, governance, and policy recommendations that are both academically rigorous and closely integrated with policy discussions. For example, Stanford HAI has published nine editions of its annual AI Index Report since 2017, which is regarded as one of the most credible and authoritative sources of AI data and insights globally, and has been cited by governments in multiple countries and regions including the United States, the United Kingdom, and China[1]. In terms of regional distribution, the inclusion of

[1] Stanford HAI, "AI Index," https://hai.stanford.edu/ai-index, Accessed: May 3, 2026.

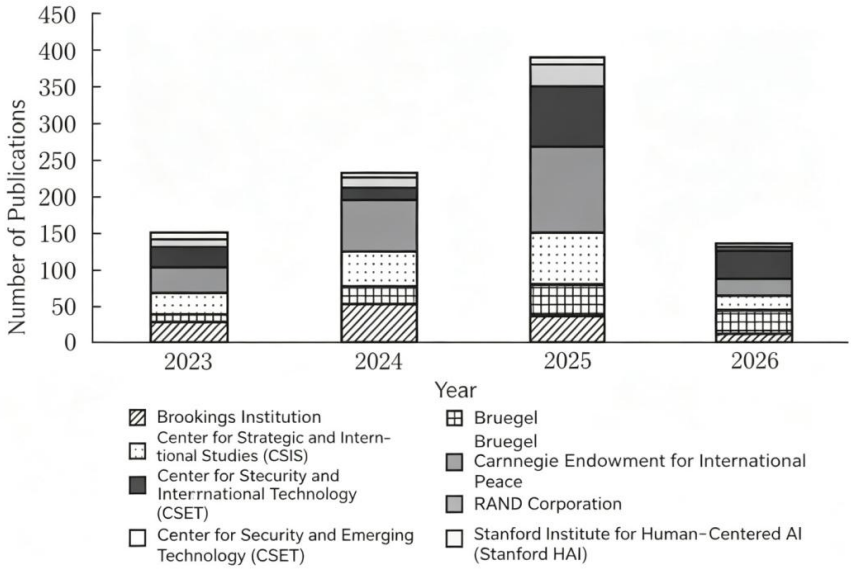
RAND Europe, the Carnegie Europe Center, and Bruegel, based in Belgium, helps to observe global AI policy issues from a European perspective, compensating for the analytical limitations of U.S.-centric perspective.

II. Research Volume and Thematic Structure

From a temporal perspective, this article primarily takes the research outputs published by the above-mentioned think tanks after 2023 as its analytical content. The reason for selecting this period is that in November 2022, generative AI technology, represented by ChatGPT, achieved breakthrough progress and sparked a new wave of AI development, making research on reports published thereafter more timely and practically relevant. In terms of specific methodology, the study systematically searched the official websites of the seven sample think tanks to obtain research reports, policy briefs, commentary articles, expert interviews, and meeting records published during this period. Through thematic retrieval, keyword keyword search, and manual screening, a total of 911 valid samples were ultimately identified.

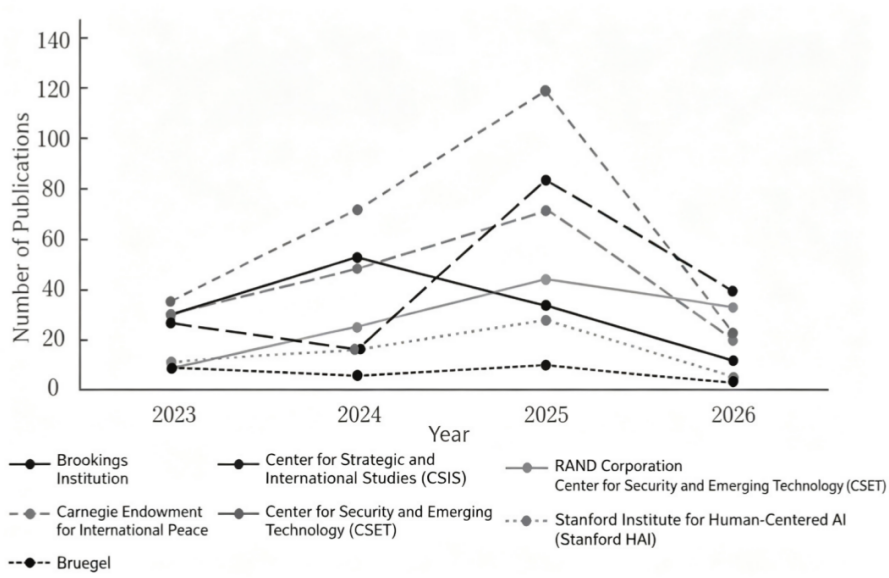
1. Research Volume

In terms of overall volume, as shown in Figure 1, the AI-related publications published by the seven think tanks exhibit a basic trend of rapid growth. During the statistical period, the seven think tanks collectively released 911 AI-related publications, with an average monthly output of 22.8 items and an average monthly output of over 3 items per think tank. Among these, 150 items were produced in 2023; the number of outputs surged to 234 in 2024, with an annual growth rate of 56.0%; 2025 saw peak output of 391 items, with an annual growth rate exceeding 67.1% and an average monthly output of approximately 32.6 items; in the first four months of 2026 alone, 136 items had already been counted, averaging 34 items per month, representing an increase of approximately 38.8% compared to the same period in the previous year, indicating that this issue remains at the core of policy debate and that output intensity continues to increase.



Source: Compiled by the author.

Figure1 Annual AI Research Output of Sample Think Tanks (2023.1–2026.4)



Source: Compiled by the author.

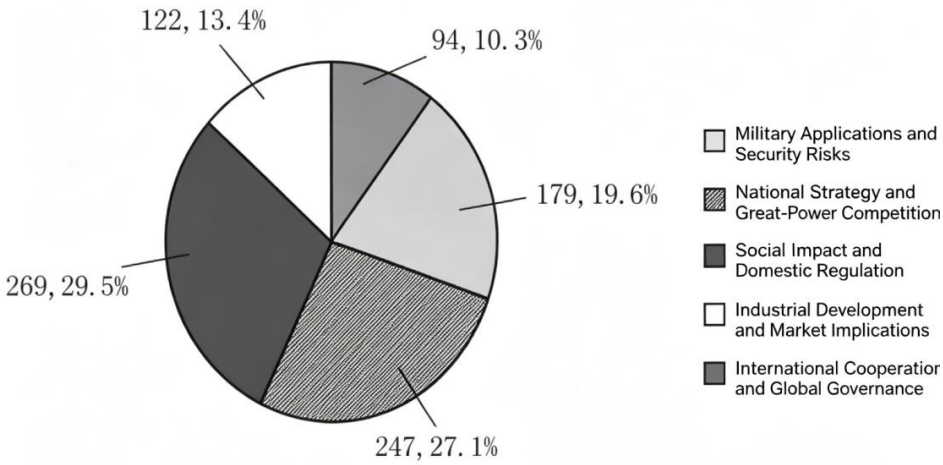
Figure 2 Trends in Annual AI Research Output of Sample Think Tanks (2023.1–2026.4)

Specifically, the seven think tanks show notable differences in both the volume of AI research outputs and the magnitude of annual changes during the statistical period. In terms of output volume, RAND Corporation produced the

most, with 249 items, accounting for 27.3%; CSET and CSIS followed with 169 and 167 items respectively, ranking second and third. Together, these three accounted for 585 items, or 64.2% of the total sample, constituting the primary source of knowledge. The Brookings Institution and Carnegie Endowment for International Peace published 127 and 111 items respectively, placing them in the middle range; Stanford HAI and Bruegel published 60 and 28 items respectively, with relatively smaller sample sizes. In terms of growth trajectories, as shown in Figure 2, RAND Corporation consistently ranked first among the seven think tanks in annual output during 2023-2025, with a compound annual growth rate of approximately 85.1%. The Carnegie Endowment for International Peace, Stanford HAI, and CSIS also maintained stable output growth, with compound annual growth rates of approximately 121.1%, 59.5%, and 53.9% respectively. Bruegel, based in Belgium, exhibited the most stable annual output, reflecting its cautious yet sustained focus on AI issues. The Brookings Institution and CSET experienced declines in annual output in 2025 and 2024 respectively, which may be attributable to internal research cycles, resource allocation, or shifts in project priorities; however, given that their outputs subsequently rebounded, their overall attention to the field has not diminished.

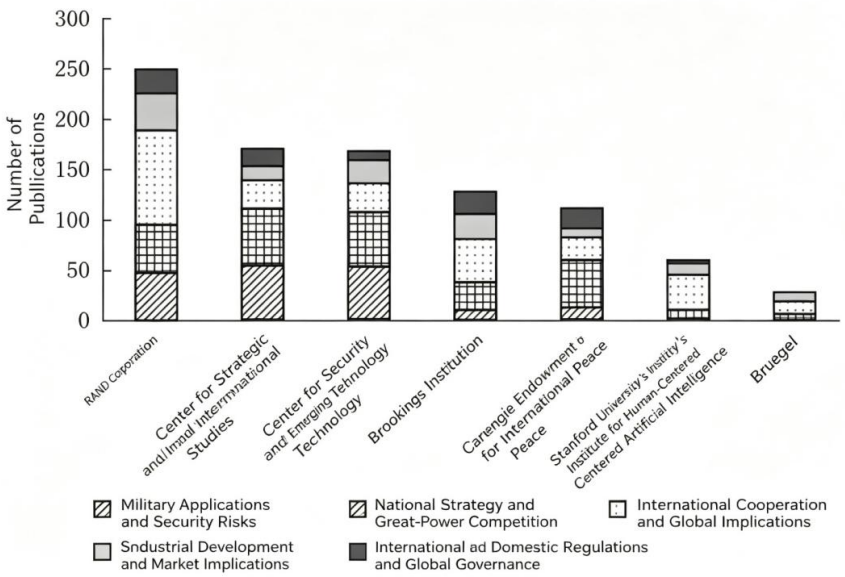
2. Thematic Distribution

Although all seven think tanks regard AI as an important research topic, there are notable differences in their focal points due to their respective organizational positioning and policy audiences. Through item-by-item reading and manual coding, this article further clusters all sample outputs into five core thematic categories, with the overall distribution shown in Figure 3. Among these, Social Impact and Domestic Regulation ranks first with 269 items, accounting for 29.5%; National Strategy and Great-Power Competition follows with 247 items, accounting for 27.1%; Military Applications and Security Risks accounts for 179 items, or 19.6%; Industrial Development and Market Implications and International Cooperation and Global Governance account for 122 and 94 items respectively, representing 13.4% and 10.3%.



Source: Compiled by the author.

Figure 3 Overall Thematic Distribution of Sample Think Tank Outputs (2023.1–2026.4)



Source: Compiled by the author.

Figure 4 Thematic Distribution of Outputs Released by Each Sample Think Tank (2023.1–2026.4)

Specifically, as shown in Figure 4, RAND Corporation has the largest volume of research and the broadest coverage, particularly excelling in the Social Impact and Domestic Regulation category, which accounts for approximately 40.9% of its total research output, primarily involving regulatory policies and system design in application scenarios such as healthcare, financial

services, climate and energy, and transportation, while also maintaining relatively balanced attention to the other four categories. CSET and CSIS have each published over 50 items in the National Strategy and Great-Power Competition and Military Applications and Security Risk categories, with these two categories combined accounting for 64.1% and 63.4% of their respective total outputs, demonstrating a distinct strategic and security research orientation. The Brookings Institution's research is notably biased toward the public policy dimension, with approximately 35.2% of its outputs concentrated in the Social Impact and Domestic Regulation category. The Carnegie Endowment for International Peace focuses its research on National Strategy and Great-Power Competition, while also addressing international governance and domestic regulatory issues to a considerable extent, which is consistent with its institutional positioning. Bruegel has a smaller sample size but a clearly defined research perspective, primarily adopting a European standpoint and focusing on economic policy, industrial competition, and regulatory framework analysis related to AI.

III. Analysis of Major Viewpoints

Based on an analysis of the 911 AI-related outputs from the seven sample think tanks, it is found that the discussions within think tanks in the West have expanded from mere technology governance to systematic examinations of national security, geopolitical competition, social transformation, industrial ecosystems, and global order. This section will further elaborate on the key viewpoints, policy debates, and strategic considerations under each of the five core thematic categories revealed in the preceding section, in order to present the core perceptions of the Western strategic research community regarding the opportunities and challenges of the AI era.

1. Social Impact and Domestic Regulation

The generality, autonomy, and lack of interpretability of AI technologies have made the social consequences they trigger and the institutional constraints they require a matter of common concern across countries. Among the entire sample, this thematic category accounts for the highest proportion, highlighting its significance at the forefront of policy and academic debate. Relevant research is primarily conducted along two dimensions: risk identification and regulatory assessment.

On the one hand, AI may give rise to numerous ethical and social risks, including algorithmic bias and discrimination, privacy violations, copyright disputes, black-box problems, and accountability challenges. The *2026 AI Index Report* released by Stanford HAI in April 2026 cites statistical data

indicating that the number of documented AI incidents in 2025, covering areas such as deepfakes, privacy breaches, and algorithmic bias, increased by more than 55% over the previous year, showing responsible AI is not keeping pace with AI capability[1]. A Brookings Institution study on "algorithmic exclusion" argues that current policy discussions on technological fairness are predominantly focused on algorithmic bias, while neglecting the more insidious but equally harmful issue of algorithmic exclusion, whereby AI systems may fail to effectively identify and assess certain groups due to missing data, thereby generating new forms of social exclusion and inequality[2]. In addition, multiple think tanks have also explored real-world challenges posed by generative AI applications such as ChatGPT and Sora, including copyright disputes[3], democratic erosion[4], and privacy violations[5].

On the other hand, the regulatory frameworks, policies, and legislation in the United States and the European Union still leave room for revision and improvement. In February 2023, the Carnegie Endowment for International Peace sought to draw lessons from Europe's horizontal and comprehensive AI Act and China's vertical, application-specific algorithmic regulation to inform U.S. policy formulation[6]. In response to U.S. state and local government actions to strengthen AI regulation, Ian Klaus, founding director of the Carnegie California Project, argued that these efforts offer future policymakers valuable opportunities to learn from best practices[7]. Regarding the executive order signed by Trump in December 2025 aimed at weakening state-level AI regulatory authority, Aalok Mehta, director of the CSIS

---

[1] Stanford HAI, "The 2026 AI Index Report," April 2026, https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf, p. 128.

[2] Catherine Tucker, "Artificial Intelligence and Algorithmic Exclusion," Brookings Institution, December 4, 2025, https://www.brookings.edu/articles/artificial-intelligence-and-algorithmic-exclusion/.

[3] Courtney C. Radsch, "The Case for Consent in the AI Data Gold Rush," Brookings Institution, January 16, 2025, https://www.brookings.edu/articles/the-case-for-consent-in-the-ai-data-gold-rush/.

[4] Rachel George and Ian Klaus, "AI and Democracy: Mapping the Intersections," Carnegie Endowment for International Peace, January 8, 2026, https://carnegieendowment.org/research/2026/01/ai-and-democracy-mapping-the-intersections.

[5] Arushi Gupta, Victor Y. Wu, Helen Webley-Brown, et al., "The Privacy-Bias Trade-Off," Stanford HAI, October 2023, https://hai.stanford.edu/assets/files/2023-10/Privacy-Bias-Trade-Off.pdf.

[6] Matt O'Shaughnessy and Matt Sheehan, "Lessons from the World's Two Experiments in AI Governance," Carnegie Endowment for International Peace, February 14, 2023, https://carnegieendowment.org/posts/2023/02/lessons-from-the-worlds-two-experiments-in-ai-governance.

[7] Ian Klaus and Ben Polsky, "Subnational Practices in AI Policy: A Working Guide," Carnegie Endowment for International Peace, December 12, 2023, https://carnegieendowment.org/research/2023/12/subnational-practices-in-ai-policy-a-working-guide.

Wadhwani Center for AI and Advanced Technologies, criticized it as "puts the cart before the horse," arguing that it runs counter to the expectations of U.S. citizens and many legislators, driven by a false sense of urgency without evidence, and adopting an unnecessarily adversarial stance toward state legislators' efforts, which could exacerbate U.S. challenges in international AI governance[1]. Bruegel has conducted multiple analyses around the EU AI Act[2] and the *Digital Markets Act*[3], discussing how Europe should balance the protection of fundamental rights with incentivizing innovation. In January 2025, RAND Corporation offered a positive assessment to the UK government's *AI Opportunities Action Plan*, describing it as "a bold experiment in AI governance—one that charts a distinct path from the European Union," and noting that "the success or failure of this targeted approach could have significant implications for how other nations balance comprehensive AI oversight with focused regulation of the most capable systems[4]."

2. National Strategy and Great-Power Competition

AI exhibits a classic Matthew Effect, enabling "winner-takes-all" outcomes through technological advantage, and thus plays a critically important role in national strategic positioning and great-power competition. Currently, think tanks in the West have reached a basic consensus that AI will become a key arena for future great-power strategic competition, and that future global AI competition will unfold primarily between the United States and China. Consequently, China, as the primary technological competitor and strategic counterparty, has become a focal point for long-term tracking and in-depth analysis by major think tanks.

First, AI is regarded as a key technology shaping future geopolitical configurations and influencing national competitiveness. In April 2025, Jacob Parakilas, research leader at RAND Europe for the Defence Strategy,

---

[1] Aalok Mehta, "Targeting State AI Laws Undermines, Rather than Advances, U.S. Technology Leadership," CSIS, December 12, 2025, https://www.csis.org/analysis/targeting-state-ai-laws-undermines-rather-advances-us-technology-leadership.

[2] Christophe Carugati, "Preparing for the Virtual AI Revolution: EU's Flexible Principles Imperative," Bruegel, October 9, 2023, https://www.bruegel.org/opinion-piece/preparing-virtual-ai-revolution-eus-flexible-principles-imperative.

[3] Bertin Martens, "Are New EU Data Market Regulations Coherent and Efficient?" Bruegel, December 18, 2023, https://www.bruegel.org/working-paper/are-new-eu-data-market-regulations-coherent-and-efficient.

[4] Paul Khullar and Sana Zakaria, "UK Government's AI Plan Gives a Glimpse of How It Plans to Regulate the Technology," RAND Corporation, January 27, 2025, https://www.rand.org/pubs/commentary/2025/01/uk-governments-ai-plan-gives-a-glimpse-of-how-it-plans.html.

Policy and Capabilities group, emphasized that AI has both direct effects on military and strategic decision-making and indirect effects on economic and social dimensions, both of which collectively shape the global competitive landscape[1]. A CSIS report also argues that AI is reshaping the ways in which economies generate wealth, militaries conduct operations, states collect and process information, and leaders engage in diplomacy[2]. In July 2025, RAND Corporation released a report entitled *How Artificial General Intelligence Could Affect the Rise and Fall of Nations: Visions for Potential AGI Futures*, which, based on simulations of eight future scenarios, identified six key factors determining the geopolitical trajectory of future AGI: degree of centralization, U.S.-China competition, Public-private relationships, capacity for AI governance, new uncertainties introduced by the challenge of AGI alignment, and economic and social disruption[3].

Second, future global AI competition will primarily unfold between the United States and China. In November 2024, Stanford HAI, after conducting a comprehensive assessment and ranking of AI vibrancy across 36 countries along eight dimensions, including R&D, economy, and education, found that from 2017 to 2023, global AI vibrancy exhibited three distinct competitive tiers: the top tier was consistently dominated by the United States and China; the middle tier was relatively stable, including the United Kingdom and India; and the lower tier was more volatile, with countries such as France, Germany, Japan, Singapore, and South Korea frequently shifting positions within this tier[4]. In June 2025, Elham Tabassi, director of the Artificial Intelligence and Emerging Technology Initiative at the Brookings Institution, predicted that the AI revolution is reshaping U.S.-China relations, and that technological development could become the primary driver of bilateral interaction over the next five years[5]. Regarding Europe's strategic positioning amid U.S.-

---

[1] Jacob Parakilas, "For Geopolitics, What AI Can't Do Will Be as Important as What It Can," RAND Corporation, April 3, 2025, https://www.rand.org/pubs/commentary/2025/04/for-geopolitics-what-ai-cant-do-will-be-as-important.html.

[2] Erica D. Lonergan and Benjamin Jensen, "AI and Grand Strategy: The Case for Restraint", CSIS, January 28, 2026, https://www. csis.org/analysis/ai-and-grand-strategy-case-restraint.

[3] Pavel, Barry, Ivana Ke, Gregory Smith, et al., "How Artificial General Intelligence Could Affect the Rise and Fall of Nations: Visions for Potential AGI Futures", RAND Corporation, July 2, 2025, https://www.rand.org/pubs/research_reports/RRA3034-2.html.

[4] Stanford HAI, "The Global AI Vibrancy Tool," November 2024, https://aiindex.stanford.edu/wp-content/uploads/2024/11/Global_AI_Vibrancy_Tool_Paper_November2024.pdf, p. 23.

[5] R. David Edelman, Diana Fu, Ryan Hass, et al., "How Will AI Influence US-China Relations in the Next 5 Years?" Brookings Institution, June 18, 2025, https://www.brookings.edu/articles/how-will-ai-influence-us-china-relations-in-the-next-5-years/.

China competition, in August 2025, CSIS invited Lucilla Sioli, Director of the EU AI Office, to discuss European AI strategy, conveying an important signal that the EU wishes to participate in the AI race and become a major player and one of the global leaders in this field[1]. In February 2026, a Bruegel commentary criticized the EU's current "plan is to rival the United States, home to most foundational AI models," arguing that it "risks abandoning core EU principles to embrace a model championed by its international competitors[2]."

On this basis, multiple think tanks have engaged in intense debate around the effectiveness of export controls toward China as a policy tool. On the one hand, in January 2025, Yasir Atalan, deputy director of the Futures Lab in CSIS's Defense and Security Department, argued in light of recent developments that the success of China's DeepSeek "does partly reflect failures of earlier implementations of U.S. export controls," but that these controls "continue to play a critical role in supporting the United States' strategy for winning the AI race against China[3]." In November, the think tank released a report entitled *The Architecture of AI Leadership: Enforcement, Innovation, and Global Trust*, in which the authors similarly argued that the United States should strengthen export controls through building a modern enforcement architecture, enhancing enforcement capabilities, expanding oversight of virtual computing, and coordinating standards with allies[4]. On the other hand, some CSIS research argues that "these tactical actions "are not a grand strategy," and that U.S. competitive advantages over China should be rooted in talent, education, infrastructure, and public innovation rather than mere export restrictions[5]. CSIS also argued that ineffective controls could stimulate technological advancement among U.S. adversaries, facilitate technology proliferation, and weaken the U.S. government's ability to monitor the destinations and end-uses of technology

[1] CSIS, "Inside Europe's AI Strategy with EU AI Office Director Lucilla Sioli," August 28, 2025, https://www.csis.org/analysis/inside-europes-ai-strategy-eu-ai-office-director-lucilla-sioli.
[2] Mario Mariniello, "Europe's Artificial Intelligence Strategy Should Be Built on European Strengths," Bruegel, February 19, 2026, https://www.bruegel.org/first-glance/europes-artificial-intelligence-strategy-should-be-built-european-strengths.
[3] Yasir Atalan, "DeepSeek's Latest Breakthrough Is Redefining AI Race," CSIS, February 3, 2025, https://www.csis.org/analysis/deepseeks-latest-breakthrough-redefining-ai-race.
[4] Charles Wessner and Shruti Sharma, "The Architecture of AI Leadership: Enforcement, Innovation, and Global Trust," CSIS, November 6, 2025, https://www.csis.org/analysis/architecture-ai-leadership-enforcement-innovation-and-global-trust.
[5] Benjamin Jensen, "Winning the Tech Race with China Requires More than Restrictions," CSIS, April 17, 2025, https://www.csis.org/analysis/winning-tech-race-china-requires-more-restrictions.

flows[1]. A commentary co-authored by Carnegie Endowment researchers suggested that the significance of export controls toward China lies in buying time for the United States and its allies to compete for the global AI market, but that to maintain its leading edge, the United States needs to accelerate the implementation of a full-stack diffusion strategy[2].

3. Military Applications and Security Risks

As a quintessential dual-use technology, AI is driving an intelligent military transformation that is deepening globally. However, technological innovation always has two sides, and AI is no exception. With the rapid evolution of technology and its real-world deployment and application, multiple Western think tanks have continuously deepened their research on military applications and security risks, with relevant analyses focusing on four main dimensions.

First, AI has become a critical enabling technology for future battlefields. In January 2025, a RAND Corporation report entitled *Artificial General Intelligence's Five Hard National Security Problems* argued that the deep application of AGI in the military domain could lead to the emergence of "wonder weapons" that would undermine existing military balances of power[3]. A CSIS Defense and Security Department report, *War and the Modern Battlefield: Insights from Ukraine and the Middle East*, also emphasized that recent conflicts in Ukraine and the Middle East mark a leap forward for advanced technologies on the battlefield, and that "the next generation of warfare will not be defined solely by who possesses the most advanced technology, but by who can integrate, adapt, and counter it the fastest[4]." Based on these assessments, multiple think tanks have used recent conflicts as testing grounds to observe how AI drives precision strikes and autonomous operations. In July 2025, Lauren A. Kahn of CSET, formerly Policy Advisor for Force Development and Emerging Capabilities in the Office of the Under Secretary

[1] William Alan Reinsch, Emily Benson, Thibault Denamiel, et al., "Optimizing Export Controls for Critical and Emerging Technologies," CSIS, May 31, 2023, https://www.csis.org/analysis/optimizing-export-controls-critical-and-emerging-technologies.
[2] Sarosh Nagar, Scott Singer and Nitarshan Rajkumar, "The Closing Window to Win: American AI Leadership Requires a Global Strategy for Full-Stack Diffusion," May 7, 2025, https://www.thefai.org/posts/the-closing-window-to-win-american-ai-leadership-requires-a-global-strategy-for-full-stack.
[3] Jim Mitre and Joel B. Predd, "Artificial General Intelligence's Five Hard National Security Problems," RAND Corporation, February 10, 2025, https://www.rand.org/pubs/perspectives/PEA3691-4.html.
[4] Aosheng Pusztaszeri and Emily Harding, "Technological Evolution on the Battlefield," CSIS, September 16, 2025, https://www.csis.org/analysis/chapter-9-technological-evolution-battlefield.

of Defense for Policy at the U.S. Department of Defense, and Michael C. Horowitz, formerly a Deputy Assistant Secretary of Defense, published an op-ed in *Foreign Affairs* exploring how recent drone operations in Ukraine and Israel mark a turning point in modern warfare, arguing that these actions demonstrate the growing impact of low-cost AI systems on traditional military platforms[1]. CSIS researchers observed Iran's drone retaliatory operations during "Operation Epic Fury", concluding that drones are no longer ancillary strike systems but central instruments of modern air campaigns, capable of imposing sustained economic, psychological, and operational pressure on adversaries at relatively low cost while preserving missile assets for specific targets[2].

Second, AI technology has a dual-edged impact on intelligence, information warfare, and cognitive security. On the one hand, a CSIS report analyzed the public security risks posed by synthetic media produced using deepfake technology, arguing that it could be misused to disrupt public opinion in adversary or rival countries, fabricate enemy positioning information, confuse enemy assessments of actual military capabilities, manipulate public sentiment, and influence political elections[3]. In April 2026, the Carnegie Endowment for International Peace further noted that although AI warfare has become dominant in Ukraine, Gaza, and Iran, this technology may also create a battlefield "fog" more hazardous than the uncertainties stemming from the scarcity of traditional information. This fog refers to the "false clarity" that emerges when the pace of AI-generated decision-making recommendations far outpaces human evaluative capacity, concealing the "technological black box[4]" inherent to AI. On the other hand, a RAND Corporation report on the future of information warfare in the Indo-Pacific region, published in May 2024, argued that the development of AI technologies, particularly advanced language models, while greatly expanding the scope of disinformation dissemination activities, also provides states with tools for analysis, characterization, and visualization to remain vigilant and detect and eliminate disinformation propaganda early. Future information warfare strategies must be dynamic, collaborative, and proactive, combining vigilance, rapid response capabilities,

---

[1] Michael C. Horowitz, Lauren A. Kahn and Joshua A. Schwartz, "What Drones Can—and Cannot—Do on the Battlefield," *Foreign Affairs*, July 4, 2025, https://www.foreignaffairs.com/united-states/what-drones-can-and-cannot-do-battlefield.

[2] Kateryna Bondar, "Unpacking Iran's Drone Campaign in the Gulf: Early Lessons for Future Drone Warfare," CSIS, March 10, 2026, https://www.csis.org/analysis/unpacking-irans-drone-campaign-gulf-early-lessons-future-drone-warfare.

[3] Di Cooke, Abigail Edwards, Alexis Day, et al., "Crossing the Deepfake Rubicon," CSIS, November 1, 2024, https://www.csis.org/analysis/crossing-deepfake-rubicon.

[4] Raluca Csernatoni, "The Fog of AI War," Carnegie Endowment for International Peace, April 16, 2026, https://carnegieendowment.org/europe/strategic-europe/2026/04/the-fog-of-ai-war.

and multinational cooperation to effectively address threats[1].

Third, AI may exacerbate risks to strategic stability and crisis escalation. A Stanford HAI study assessed the escalation risks of large language models in high-stakes military and diplomatic decision-making from an interdisciplinary perspective, concluding that "turning high-stakes decisions over to autonomous LLM-based agents can lead to significant escalatory action," and in some cases even to the use of nuclear weapons[2]. In January 2025, CSIS experts held a live debate on the integration of AI into U.S. nuclear command, control, and communications (NC3). Proponents argued that AI integration has already become a reality, and that relevant tools and technologies can accelerate decision-making, enhance early warning and pre-launch systems, and analyze complex intelligence data. Critics, however, drew on historical lessons of automation failures, loss of human control, and nuclear accidents to argue that AI has technical flaws, contributes to the spread of disinformation, and increases the risks of accidental escalation and cyber vulnerabilities[3].

Finally, AI is becoming a multiplier for non-traditional security threats such as terrorism, cybercrime, and biosecurity. In March 2025, Jacob Parakilas, research leader at RAND Europe for the Defence Strategy, Policy and Capabilities group, hyper-empowering, warned that under certain circumstances, AI could provide non-state armed groups with hyper-empowering capabilities: it can generate malicious computer code, help develop novel biological weapons, and provide cheap alternatives for guided munitions, which were once the exclusive domain of superpower militaries and introduce destabilizing factors and profound unpredictability into geopolitics[4]. Another RAND report directly listed "empower nonexperts to develop weapons of mass destruction" as one of the five hard national security problems posed by AGI, arguing that foundation models could help non-experts rapidly acquire complex knowledge, and that AI-assisted gene-editing tools would substantially reduce the cost of synthesizing lethal viruses. Although nuclear weapons manufacture still depends on sophisticated

---

[1] Raluca Csernatoni, "The Fog of AI War," Carnegie Endowment for International Peace, April 16, 2026, https://carnegieendowment.org/europe/strategic-europe/2026/04/the-fog-of-ai-war.

[2] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, et al., "Escalation Risks from LLMs in Military and Diplomatic Contexts," Stanford HAI, May 2024, https://hai.stanford.edu/assets/files/2024-05/Escalation-Risks-Policy-Brief-LLMs-Military-Diplomatic-Contexts.pdf, p. 1.

[3] CSIS, "PONI Live Debate: AI Integration in NC3," January 24, 2025, https://www.csis.org/analysis/poni-live-debate-ai-integration-nc3.

[4] Jacob Parakilas, "For Geopolitics, What AI Can't Do Will Be as Important as What It Can," RAND Corporation, April 3, 2025, https://www.rand.org/pubs/commentary/2025/04/for-geopolitics-what-ai-cant-do-will-be-as-important.html.

industrial infrastructure, the thresholds for biological and cyber weapons are continuously declining[1]. In April 2026, a CSIS blog post, drawing on Anthropic's newly released Claude Mythos model, emphasized that the era of AI autonomously identifying and weaponizing vulnerabilities is accelerating, and that AI's core contribution to the cyber threat domain lies not in innovation, but in efficiency[2].

4. Industrial Development and Market Implications

As the strategic engine driving the new round of technological and industrial transformation, AI's innovative development and industrial translation have become long-term, overarching critical issues. Since 2023, Western think tanks have conducted multidimensional and systematic discussions around the various factors required throughout the full lifecycle of AI.

First, electricity and energy, infrastructure, and critical minerals constitute the material foundation for AI industrial development. CSIS research clearly states that modern AI systems rely on massive parallelism, continuous data flows, and sustained operation across large clusters, making energy availability more important than peak chip performance. "Power is now the significant binding constraint on AI progress. National competitive advantage increasingly depends on who can most efficiently convert electricity into sustained, real-world AI output, rather than who builds the smallest transistors[3]." A Brookings Institution commentary emphasizes that non-discriminatory access to critical infrastructure is a prerequisite for competition, which is particularly crucial in the AI era[4]. A study warns against the rising political impulse surrounding sovereign cloud and sovereign AI construction, arguing that greater control also tends to mean higher costs, slower growth, and less innovation. "Sovereign infrastructure has a poor track record of success and often becomes stranded investment, burdening economies[5]." CSIS

[1] Jim Mitre and Joel B. Predd, "Artificial General Intelligence's Five Hard National Security Problems," RAND Corporation, February 10, 2025, https://www.rand.org/pubs/perspectives/PEA3691-4.html.
[2] Jiwon Lim, "Beyond Autonomous Attacks: The Reality of AI-Enabled Cyber Threats," CSIS, April 29, 2026, https://www.csis.org/blogs/strategic-technologies-blog/beyond-autonomous-attacks-reality-ai-enabled-cyber-threats.
[3] Maryam Cope, "The Rise of 'Watt's La' and Why Power Could Put a Ceiling on American Innovation," CSIS, February 13, 2026, https://www.csis.org/analysis/rise-watts-law-and-why-power-could-put-ceiling-american-innovation.
[4] Tom Wheeler, "OpenAI Floats Federal Support for AI Infrastructure—What Should the Public Expect?" Brookings Institution, November 18, 2025, https://www.brookings.edu/articles/openai-floats-federal-support-for-ai-infrastructure-what-should-the-public-expect/.
[5] Bill Whyman, "Sovereign Cloud–Sovereign AI Conundrum: Policy Actions to Achieve Prosperity and Security," CSIS, December 4, 2025,

research on critical minerals shows that rare earth elements enabling AI, drones, and defense systems are not only indispensable components of U.S. economic strength, but also vital parts of its security infrastructure, and that the ability to deter, respond, and prevail in future conflicts will depend on them[1].

Second, policy orientation and capital investment profoundly affect market structures and innovation vitality. A July 2024 Bruegel policy brief noted that the exponential growth of model training costs has created market entry barriers for AI startups, which often need to partner with big tech firms to access computing infrastructure and end markets. While such cooperation can enhance overall innovation efficiency, it may also reinforce the market advantages of large firms and distort competitive dynamics, making it necessary for governments to revise the entire competition policy setting for AI industries[2]. Navin Girishankar, president of the Economic Security and Technology Department at CSIS, argued that the United States urgently needs to completely reset its tariff policy as a national program to achieve full-stack leadership in AI, with one important element being that the government's proposed sovereign wealth fund design should pay particular attention to AI and other critical emerging technologies, with the dual objectives of accelerating the translation of AI and AI-driven breakthrough technologies from lab to market, and deploying capital to remove bottlenecks to AI diffusion throughout the economy[3]. RAND Corporation also used simulation studies to analyze whether government investment in AI yields returns, thereby enhancing the future robustness of U.S. AI[4].

Third, workforce structural transformation is an inevitable requirement for adapting to the AI economy. CSIS research based on modeling points out that AI infrastructure construction could create new labor demand. By 2030, the United States may need approximately 140,000 electricians, HVAC technicians, and welders to support AI infrastructure construction. Without rapid expansion of training and apprenticeship systems, labor instead of chips

---

https://www.csis.org/analysis/sovereign-cloud-sovereign-ai-conundrum-policy-actions-achieve-prosperity-and-security.

[1] Jake Kwon and Benjamin Jensen, "Deterrence Runs on Rare Earths," CSIS, July 24, 2025, https://www.csis.org/analysis/deterrence-runs-rare-earths.

[2] Bertin Martens, "Why Artificial Intelligence Is Creating Fundamental Challenges for Competition Policy," Bruegel, July 18, 2024, https://www.bruegel.org/policy-brief/why-artificial-intelligence-creating-fundamental-challenges-competition-policy.

[3] Navin Girishankar, "From Tariffs to Tech Power: The Pivot the United States Needs Now," CSIS, May 15, 2025, https://www.csis.org/analysis/tariffs-tech-power-pivot-united-states-needs-now.

[4] Brian A. Jackson and Pauline Moore, "Reality Checking a Major National R&D Investment in AI Trustworthiness, Safety, and Security," RAND Corporation, March 12, 2026, https://www.rand.org/pubs/research_reports/RRA4718-1.html.

or electricity could become the key bottleneck restricting U.S. AI leadership[1]. In April 2026, a Carnegie Endowment paper entitled *The AI Labor Debate: Three Views on the Future of Work* systematically synthesized the debate on AI's employment impact over the previous two years into three camps: the alarmed, which strongly believes AI will replace humans; the patient, which believes the impact will unfold slowly; and the excited, which believes AI will create more new jobs. The authors further pointed out that the true divergence among the three camps centers on five empirically testable variables, and that policymakers can use this to improve data collection, pilot wage insurance and retraining programs, and monitor three categories of key indicators, thereby preparing response tools for all scenarios[2]. In addition, multiple studies from CSIS[3], the Brookings Institution[4], and CSET point out that talent, education, and public innovation are the true sources of America's long-term competitive advantage, and that the answer to maintaining technological leadership "is not to erect more barriers, but to build more bridges," including retaining global talent, expanding educational opportunities, and developing research infrastructure[5].

5. International Cooperation and Global Governance

In the AI era, how to effectively conduct AI governance, maximize the development potential of AI technologies, and address the security risks posed by AI applications has become a major challenge facing all countries. Currently, Western think tanks have produced a relatively limited volume of discussion on foreign policy, focusing primarily on two aspects: first, international collaboration pathways oriented toward strategic positioning and capacity cooperation; and second, global governance systems centered on risk management and rule construction.

---

[1] Navin Girishankar and Karl Smith, "GenAI's Human Infrastructure Challenge—Can the United States Meet Skilled Trade Labor Demand Through 2030?" CSIS, September 16, 2025, https://www.csis.org/analysis/genais-human-infrastructure-challenge-can-united-states-meet-skilled-trade-labor-demand.

[2] Teddy Tawil, "The AI Labor Debate: Three Views on the Future of Work," Carnegie Endowment for International Peace, April 23, 2026, https://carnegieendowment.org/research/2026/04/the-ai-labor-debate-three-views-on-the-future-of-work.

[3] Benjamin Jensen, "Winning the Tech Race with China Requires More than Restrictions," CSIS, April 17, 2025, https://www.csis.org/analysis/winning-tech-race-china-requires-more-restrictions.

[4] Darrell M. West, "Improving Workforce Development and STEM Education to Preserve America's Innovation Edge," Brookings Institution, July 26, 2023, https://www.brookings.edu/articles/improving-workforce-development-and-stem-education-to-preserve-americas-innovation-edge/.

[5] Julia Shapero, "Trump's $100K H-1B Visa Fee Rattles Silicon Valley," *The Hill*, September 24, 2025, https://thehill.com/policy/technology/5518278-h1b-visa-fee-shockwaves/.

First, there is a focus on international cooperation and global strategic landscape. In February 2026, Carnegie Endowment researchers focused on two strategic clusters: treaty allies and critical swing states in the Global South[1]. This also reflects the common concerns of multiple think tanks. In terms of alliance and partnership coordination, CSIS research argues that although Washington has long favored offensive realism, the most logical strategy of restraint in the AI era requires replacing traditional political-military alliances with new economic networks that connect states, corporations, and sovereign wealth funds to meet the insatiable demand for infrastructure investment[2]. With regard to the Global South, RAND Corporation, based on its continuous tracking Chinese government-supported development finance projects that utilized or enabled AI technology in the Global South since 2000, points out that AI exports facilitated by such arrangements could further consolidate China's position in global technology supply chains and trade networks, and enhance its influence in technology standard-setting and governance rule-making, making it necessary for the United States to remain vigilant and adjust its international cooperation and standard-setting strategies accordingly[3]. Noam Unger, director of CSIS's Sustainable Development and Resilience Initiative, and colleagues argue that although the Trump administration's push for full-stack technology package exports to reduce international competitors' technological dependence is a direction worth affirming, the geopolitical competition between the United States and its adversaries is often contested within developing countries in the Global South, particularly Africa, with its projected labor demographics and growth[4]. The think tank further emphasized in subsequent analysis that emerging and developing economies tend to be more optimistic about AI than developed countries, but are less prepared for technology adoption, and this gap could provide a "golden opportunity" for cooperation with the United States on AI standards, skills, and exports[5].

---

[1] Anton Leicht, Sarosh Nagar, Liam Patell et al., "An Allied World on the American AI Stack: A Strategy for Export Leadership," Foundation for American Innovation, February 6, 2026, https://www.thefai.org/posts/an-allied-world-on-the-american-ai-stack-a-strategy-for-export-leadership.

[2] Erica D. Lonergan and Benjamin Jensen, "AI and Grand Strategy: The Case for Restraint," CSIS, January 28, 2026, https://www.csis.org/analysis/ai-and-grand-strategy-case-restraint.

[3] Jennifer Bouey, Lynn Hu, Keller Scholl, et al., "China's AI Exports Database (CAIED)," RAND Corporation, October 15, 2025, https://www.rand.org/pubs/tools/TLA2696-1.html.

[4] Navin Girishankar, Kirti Gupta, Matt Pearl, et al., "Experts React: Unpacking the Trump Administration's Plan to Win the AI Race," CSIS, July 25, 2025, https://www.csis.org/analysis/experts-react-unpacking-trump-administrations-plan-win-ai-race.

[5] Navin Girishankar and Andrea Leonard Palazzi, "Is the Global South More Optimistic Than Prepared for AI?", CSIS, October 23, 2025, https://www.csis.org/analysis/global-south-more-optimistic-prepared-ai.

Second, there is attention to catastrophic risks and global governance. Think tanks currently hold different views on the catastrophic risks that AI may bring. On the one hand, RAND research states that humans still cannot understand AI's decision-making processes and have limited understanding of the second-and third-order effects of instructions given to AI systems, meaning that the motivations and objectives of AI operations could diverge significantly from human intentions, harboring strong destructive potential[1]. CSET researchers warned in media statements that the AI race is making a "Hindenburg-style disaster a real risk[2]." CSIS experts also stated that human extinction deserves to be taken seriously[3]. On the other hand, some views remain cautious about catastrophic risk narratives. For example, CSIS published a commentary by cybersecurity expert James Lewis arguing that constantly talking about impending catastrophic risks could create false panic and lead to bad policies[4]. RAND scientist Michael J. D. Vermeer and colleagues also specifically assessed AI's capacity to generate extinction threats using three technologies, namely nuclear weapons, biological pathogens, and climate change, ultimately concluding that while the possibility cannot be ruled out, extinction threats would require substantial time to take effect in each scenario, which might allow humanity time to respond and mitigate the threats[5].

On this basis, think tanks have further examined the current state and future of multilateral rules and international collaboration. In February 2025, Li Feifei, co-founding director of Stanford HAI and "godmother of AI," stated that "now more than ever, AI needs a governance framework[6]." CSIS researchers, in an article published in October 2025, discussed the UN Global Dialogue on AI Governance, arguing that its significance matters less for its immediate regulatory impact, than as a powerful symbolic signal of political will. At the same time, amid U.S. disengagement and China's increasingly

---

[1] Barry Pavel, Ivana Ke, Michael Spirtas, et al., "AI and Geopolitics: How Might AI Affect the Rise and Fall of Nations?" RAND Corporation, November 3, 2023, https://www.rand.org/pubs/perspectives/PEA3034-1.html.

[2] CSET, "AI Race Making a 'Hindenburg-Style Disaster a Real Risk'," CSET, February 18, 2026, https://cset.georgetown.edu/article/ai-race-making-a-hindenburg-style-disaster-a-real-risk/.

[3] Michael Frank, "Managing Existential Risk from AI without Undercutting Innovation," CSIS, July 10, 2023, https://www.csis.org/analysis/managing-existential-risk-ai-without-undercutting-innovation.

[4] James Andrew Lewis, "AI and Rumors of Impending Doom," CSIS, July 5, 2023, https://www.csis.org/analysis/ai-and-rumors-impending-doom.

[5] Michael J. D. Vermeer, Emily Lathrop and Alvin Moon, "On the Extinction Risk from Artificial Intelligence," RAND Corporation, May 6, 2025, https://www.rand.org/pubs/research_reports/RRA3034-1.html.

[6] Fei-Fei Li, "Now More than Ever, AI Needs a Governance Framework," *Financial Times*, February 8, 2025, https://www.ft.com/content/3861a30a-50fc-41c9-9780-b16626a0d2e8.

assertive stance, it has become a testing ground for deep-seated international contradictions, as well as an important barometer for measuring geopolitical dynamics in AI governance[1]. Multiple studies have also conducted real-time tracking and evaluation of recent governance efforts at AI safety summits[2] and in regions such as Southeast Asia[3], Europe[4], and Africa[5]. Regarding the future of international governance, in August 2024, Michael Vermeer, a technology expert at the RAND Corporation, drew on governance models for nuclear technology, the internet, cryptographic products and genetic engineering, fields that have faced dilemmas analogous to those confronting AI governance since World War II, and put forward governance proposals for three distinct scenarios. When access to and deployment of AI technologies demand substantial resources and carry serious risks of broad harm, stakeholders may adopt the governance model for nuclear technology, reaching consensus on the safe application of such technologies and establishing an international organization analogous to the International Atomic Energy Agency. If AI generates limited developmental risks, governments may delegate governance authority to the private sector by following internet governance practices. Where AI development faces nearly no competitive barriers yet poses substantial risks to governments and human well-being, policymakers may learn from the governance framework for cryptographic products to build or strengthen public-private partnerships[6]. However, Lauren A. Kahn of CSET

[1] Laura Caroli and Matt Mande, "What the UN Global Dialogue on AI Governance Reveals About Global Power Shifts," CSIS, October 7, 2025, https://www.csis.org/analysis/what-un-global-dialogue-ai-governance-reveals-about-global-power-shifts.

[2] Joshua Meltzer and Paul Triolo, "The Bletchley Park Process Could Be a Building Block for Global Cooperation on AI Safety," Brookings Institution, October 7, 2024, https://www.brookings.edu/articles/the-bletchley-park-process-could-be-a-building-block-for-global-cooperation-on-ai-safety/; Mariano-Florentino (Tino) Cuéllar, "The UK AI Safety Summit Opened a New Chapter in AI Diplomacy," Carnegie Endowment for International Peace, November 9, 2023, https://carnegieendowment.org/posts/2023/11/the-uk-ai-safety-summit-opened-a-new-chapter-in-ai-diplomacy; Joanna Wiaterek, Jared Perlo and Sumaya Nur Adan, "AI Safety and Security Can Enable Innovation in Global Majority Countries," Brookings Institution, September 22, 2025, https://www.brookings.edu/articles/ai-safety-and-security-can-enable-innovation-in-global-majority-countries/.

[3] Lyantoniette Chua, Philip Tham and Chinasa T. Okolo, "AI Safety Governance, the Southeast Asian Way," Brookings Institution, August 26, 2025, https://www.brookings.edu/articles/ai-safety-governance-the-southeast-asian-way/.

[4] Raluca Csernatoni, "Charting the Geopolitics and European Governance of Artificial Intelligence," Carnegie Endowment for International Peace, March 6, 2024, https://carnegieendowment.org/research/2024/03/charting-the-geopolitics-and-european-governance-of-artificial-intelligence.

[5] Chinasa T. Okolo, "AI Is Not Afric's Savior: Avoiding Technosolutionism in Digital Development," Brookings Institution, August 1, 2025, https://www.brookings.edu/articles/ai-is-not-africas-savior-avoiding-technosolutionism-in-digital-development/.

[6] Michael J. D. Vermeer, *Historical Analogues That Can Inform AI Governance*, RAND

co-authored an article released in June 2025 to counter this viewpoint. She argues that nuclear non-proliferation constitutes an inappropriate framework for AI governance, given the fundamental disparities between AI and nuclear technology. Governing AI based on nuclear analogies may raise the international community's expectations for capacity to control model proliferation, overemphasize the deployment of such technology in weaponry, and overlook its potential to drive economic prosperity and bolster national security. Instead of drawing inspiration from the International Atomic Energy Agency, the most suitable blueprints for AI governance lie in international and national standard-setting bodies and agreements such as the U.S. Food and Drug Administration, the International Civil Aviation Organization, the International Telecommunication Union, and the United Nations Convention on the Law of the Sea[1]. In April 2026, the Carnegie Endowment for International Peace offered a reality-based analysis. The United States' fierce opposition to all multilateral AI governance initiatives at the AI Impact Summit hosted in India demonstrates that countries at the cutting edge of AI innovation currently have little geopolitical incentive to establish a global oversight body dedicated to AI safety and security[2].

IV. Recommendations for Maintaining Strategic Advantage in the AI Era

In response to the upheavals brought by the above AI technologies, Western think tanks, based on their assessments of technological evolution and great-power dynamics, have proposed a series of policy recommendations aimed at consolidating and expanding strategic advantages across five key dimensions: strategic coordination, technology diffusion, military applications, innovation ecosystems, and international cooperation.

First, build on existing AI leadership advantages and strengthen national strategic coordination. How to formulate national strategy for the AI era amid technological iteration and great-power competition has been a hot topic of think tank discussion. In September 2025, RAND Corporation published a commentary by Jeffrey Ding, assistant professor of political science at George

---

Corporation, August 19, 2024, https://www.rand.org/pubs/research_reports/RRA3408-1.html.

[1] Michael C. Horowitz and Lauren A. Kahn, *Nuclear Non-Proliferation Is the Wrong Framework for AI Governance*, AI Frontiers, June 27, 2025, https://ai-frontiers.org/articles/nuclear-non-proliferation-is-the-wrong-framework-for-ai-governance.

[2] Nidhi Singh, Tejas Bharadwaj, Shruti Mittal, et al., *For People, Planet, and Progress: Perspectives from India's AI Impact Summit, Carnegie Endowment for International Peace*, April 30, 2026, https://carnegieendowment.org/research/2026/04/for-people-planet-and-progress-perspectives-from-indias-ai-impact-summit.

Washington University, criticizing the AGI sprint anxiety in the U.S. national security community and arguing that AI competition is a marathon testing endurance, systematicity, and inclusiveness, and that the United States must urgently shift its strategic and policy focus from pursuing model breakthroughs to promoting technology diffusion[1]. CSIS's Navin Girishankar and colleagues explicitly stated that winning the technology race does not depend on a single breakthrough; what truly determines long-term leadership is "ecosystem strength," composed of Economy-wide fundamentals, Technology-specific enablers, Ecosystem governance and Enterprise strategies[2]. In terms of methodological reference, a RAND report emphasized that the United States must transcend fragmented policies, draw lessons from the past, and build a comprehensive strategy drawing on six historical strategic experiences including the Marshall Plan and the Apollo Program[3]. Some experts have also provided reference points for U.S. open-source AI strategy by examining the different frameworks of U.S. and Chinese AI action plans[4]. Bruegel, taking a pragmatic approach, emphasized that Europe's AI strategy should be built on its own strengths and should adhere to building an open, competitive, and resilient AI system[5].

Second, optimize export controls and technology diffusion, and expand global market share. Since the intense debate over the effectiveness of export control policy triggered by DeepSeek in early 2025, CSIS has recommended that "these blanket restrictions should give way to more detailed and targeted export-control systems," paired with better enforcement[6]. Subsequently, some studies, building on reflection and critique, have attempted to propose alternative approaches. Researchers from RAND Corporation and the Center for a New American Security co-authored an article proposing that the United States could choose to rent, rather than sell, world-class AI chips, while

---

[1] Jeffrey Ding, "Running the Right Artificial Intelligence Race," RAND Corporation, September 30, 2025, https://www.rand.org/pubs/perspectives/PEA4165-1.html.
[2] Navin Girishankar, Mark P. Dallas, Sree Ramaswamy, et al., "Tech Edge: A Living Playbook for America's Technology Long Game," CSIS, January 20, 2026, https://www.csis.org/analysis/tech-edge-living-playbook-americas-technology-long-game.
[3] Michael J. Mazarr and Anca Agachi, "Building a U.S. National Strategy for the Artificial Intelligence Era," RAND Corporation, November 17, 2025, https://www.rand.org/pubs/perspectives/PEA4297-1.html.
[4] Owen J. Daniels and Hanna Dohmen, "Open Models, Soft Power, and the Spectrum of U.S.-China Artificial Intelligence Competition," RAND Corporation, March 26, 2026, https://www.rand.org/pubs/perspectives/PEA4686-1.html.
[5] Mario Mariniello, "Europe's Artificial Intelligence Strategy Should Be Built on European Strengths," Bruegel, February 19, 2026, https://www.bruegel.org/first-glance/europes-artificial-intelligence-strategy-should-be-built-european-strengths.
[6] Yasir Atalan, "DeepSeek's Latest Breakthrough Is Redefining AI Race," CSIS, January 20, 2026, https://www.csis.org/analysis/tech-edge-living-playbook-americas-technology-long-game.

retaining the power to withdraw access at any time[1]. CSIS's Navin Girishankar argued that the United States should condition access to leading-edge chips on binding commitments to settle AI-enabled exports in dollars, thereby "turning the AI revolution into dollar dominance[2]." A July 2025 Carnegie Endowment study pointed out that the United States is in a critical window to consolidate its leadership in global AI infrastructure. In the face of Chinese solutions in the Global South, the United States should implement five strategies: first, Pick Smart Battles and Tailor AI Offerings; second, Unlock Strategic Capital; third, Bargain Smartly; fourth, Embrace Sub-frontier Open-Source ; and fifth, Fewer Lectures, More Problem-Solving[3]. In February 2026, the think tank further summarized its approach in a report entitled *An Allied World on the American AI Stack: A Strategy for Export Leadership*, advocating for focused and high-impact export projects targeting two major strategic clusters, namely treaty allies and critical swing states in the Global South, and effectively utilizing coordination, diplomacy, and financing as three tools to close the gap between market delivery and U.S. strategic requirements[4].

Third, accelerate military R&D and deployment while strengthening security risk management. In January 2026, RAND Corporation released a report entitled *How Artificial Intelligence Could Reshape Four Essential Competitions in Future Warfare*, putting forward three recommendations to the U.S. government on how to gain first-mover advantages in AI: first, invest more in research and experimentation for new capabilities that enable the greater use of mass and deception (such as attritable runway-independent uncrewed aircraft and new AI tools for deception); second, allocate scarce resources under the assumption that it will face sophisticated and adaptive adversaries, avoiding blind reliance on AI first-mover advantages; third, accelerate a plan to manage the transition to an AI-enabled force[5]. In September 2024, RAND Europe completed a research report commissioned by

---

[1] Janet Egan and Lennart Heim, "America Should Rent, Not Sell, AI Chips to China," RAND Corporation, August 15, 2025, https://www.rand.org/pubs/commentary/2025/08/america-should-rent-not-sell-ai-chips-to-china.html.

[2] Navin Girishankar, "Turning the AI Revolution into Dollar Dominance," CSIS, December 8, 2025, https://www.csis.org/analysis/turning-ai-revolution-dollar-dominance.

[3] Marianne Lu and Sam Winter-Levy, "The Other AI Race: An Export Promotion Strategy for the Global South," Carnegie Endowment for International Peace, July 21, 2025, https://carnegieendowment.org/research/2025/07/the-other-ai-race-an-export-promotion-strategy-for-the-global-south?lang=en.

[4] Anton Leicht, Sarosh Nagar, Liam Patell et al., "An Allied World on the American AI Stack: A Strategy for Export Leadership," Foundation for American Innovation, February 6, 2026, https://www.thefai.org/posts/an-allied-world-on-the-american-ai-stack-a-strategy-for-export-leadership.

[5] Zachary Burdette, Dwight Phillips, Jacob L. Heim, et al., "How Artificial Intelligence Could Reshape Four Essential Competitions in Future Warfare," RAND Corporation, January 22, 2026.

the UK Ministry of Defence and the Foreign, Commonwealth & Development Office, proposing three categories of feasible measures to mitigate AI-related risks: Mechanisms to boost AI adoption and benefits for Defence (accelerate investment in and adoption of AI across Defence, while also seeking to increase resilience against hostile usage of AI);Mechanisms to restrict AI adoption and benefits for adversaries (restrict, slow or increase the costs to nonstate and terrorist actors, and hostile states of deploying military AI); Mechanisms to shape and influence governance arrangements (focus on easier issues, nonbinding mechanisms or smaller groups, then building trust and momentum towards eventual consolidation of a more robust governance architecture[1]). In June 2025, CSIS, based on AI vulnerabilities and the challenges of military applications, recommended that the U.S. Department of Defense promote the development of quantitative and qualitative tools to reduce the difficulty of assessing, managing, and controlling algorithms, and systematically build operational concepts for effectively integrating AI into warfare[2].

Fourth, balance innovation incentives and regulatory rules, and break through electricity and talent bottlenecks. In April 2026, CSIS published a commentary entitled *Toward a Mature AI Economy: Policy Priorities for the Road Ahead*, arguing that the state of the AI economy has not yet matured, and that the next step requires shifting from embedding AI into existing processes for efficiency gains to using AI as a catalyst to drive innovation and build sustainable business models. The article put forward three specific recommendations for U.S. policymakers: first, shift policy attention away from the "tech sector" to industry-specific practical AI applications; second, fill the regulatory vacuum in data privacy; and third, investing in education and training of AI fundamentals[3]. Facing the obstacles posed by regulatory uncertainty, a December 2024 Brookings commentary pointed out that balance is critical, as poorly designed laws could hinder investment, impede startups, and undermine U.S. AI leadership. Through flexible policies to promote innovation and by cultivating and retaining top talent, the United States can maintain its leading position in global AI competition while ensuring

[1] James Black, Mattias Eken, Jacob Parakilas, et al., "Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence," RAND Corporation, September 6, 2024, https://www.rand.org/pubs/research_reports/RRA3295-1.html.
[2] Carol Kuntz, "Artificial Intelligence and War: How the Department of Defense Can Lead Responsibly," CSIS, June 26, 2025, https://www.csis.org/analysis/artificial-intelligence-and-war.
[3] Yinuo Geng, "Toward a Mature AI Economy: Policy Priorities for the Road Ahead," CSIS, April 22, 2026, https://www.csis.org/analysis/toward-mature-ai-economy-policy-priorities-road-ahead.

accountability and ethical development[1]. In response to the talent gap, RAND Corporation pointed out that in the global competition to develop AGI, a carefully tailored and pragmatic approach to AI talent management will be crucial for the United States, with specific measures including preserving homegrown AI talent, capturing foreign AI talent, and targeting adversarial AI talent[2]. In response to electricity and energy bottlenecks, RAND argued that relying on renewables such as wind and solar alone cannot solve the problem; the United States must advance grid modernization while balancing emissions reduction targets, power supply reliability, and AI development, and rapidly release grid capacity through policy coordination and technological investment, proposing the most feasible single-site example and two medium-potential AI data center sites[3].

Fifth, conduct bilateral standards and security cooperation to shape the AI governance process. In January 2024, researchers from Carnegie Endowment's Global Order and Institutions Program published an article envisioning a regime complex for global AI governance, characterizing BRICS and the Shanghai Cooperation Organization as "rival coalitions" that could promote norms distinct from Western ones and accelerate fragmentation of the global economy and world order. In response, the United States and other Western developed market democracies need to cooperate with democratic countries in the Global South (such as India) to develop standards and coordinate regulatory policies[4]. In April 2026, Kyle Chan of the Brookings Institution testified before the U.S. House Select Committee on the Strategic Competition Between the United States and the Communist Party of China, stating that although the United States is engaged in intense competition with China in the AI field and has taken measures to protect the United States and U.S. companies from risks associated with Chinese AI development, the two countries should still consider developing common AI safety protocols and even establishing communication channels at points of converging national interests[5]. In practical terms, in April 2025, a group of scholars including

[1] Sarah Kreps, "The Global AI Race: Will US Innovation Lead or Lag?" Brookings Institution, December 6, 2024, https://www.brookings.edu/articles/the-global-ai-race-will-us-innovation-lead-or-lag/.
[2] Iskander Rehman, "The Battle for Brilliant Minds: From the Nuclear Age to AI," RAND Corporation, August 1, 2025, https://www.rand.org/pubs/external_publications/EP71021.html.
[3] Ismael Arciniegas Rueda, Ryan J. Bain, Hye Min Park, et al., "Accelerating Large-Scale Grid Infrastructure Projects to Win the AI Race," RAND Corporation, December 4, 2025, https://www.rand.org/pubs/perspectives/PEA4501-1.html.
[4] Emma Klein and Stewart Patrick, "Envisioning a Global Regime Complex to Govern Artificial Intelligence," Carnegie Endowment for International Peace, March 21, 2024, https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence.
[5] Kyle Chan, "Competing AI Strategies for the US and China," Brookings Institution,

Turing Award laureate and "godfather of deep learning" Yoshua Bengio, from institutions including Stanford University, RAND Corporation, and the Carnegie Endowment for International Peace, jointly published an article entitled *In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?*, proposing that research on AI verification mechanisms and shared protocols, as areas of lower risk and challenge, could be suitable domains for U.S.-China AI cooperation[1]. Scott Singer, a researcher in technology and international affairs at the Carnegie Endowment, outlined three prudent approaches for UK engagement with China: first, establish emergency preparedness protocols and institutions, including methods for convening AI safety authorities across countries and a minimum set of safety preparedness measures; second, establish a safety assurance framework to ensure that models exhibiting specific risk levels are not further developed or deployed; and third, fund independent global AI safety and verification research[2].

V. Conclusion

Through an analysis of 911 AI-related outputs published by seven think tanks between January 2023 and April 2026, it is found that Western think tank research on AI has transcended discussions of technological ethics or economic impacts, and is now deeply embedded in narratives of national security, great-power competition, and global order restructuring. Overall, the tone of Western think tank discourse on AI issues is becoming increasingly pragmatic, manifesting in three specific aspects: first, in terms of strategic objective perception, shifting from pursuing technological breakthroughs in AGI and frontier models to building long-term systemic advantages encompassing computing power, energy, data centers, critical minerals, talent, and regulatory certainty; second, in terms of great-power competition strategy, shifting from discussing how to contain competitors to expanding Western technology stack influence, consolidating alliance and partnership networks, and seizing global markets while restricting the outflow of key technologies; and third, in terms of technology governance approaches, shifting from principled debates such as "will AI take jobs" and "should AI be regulated" to operational governance focused on what indicators to use for monitoring

---

April 16, 2026, https://www.brookings.edu/articles/competing-ai-strategies-for-the-us-and-china/.

[1] Ben Bucknall, Saad Siddiqui, Lara Thurnherr, et al., "In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?" *arXiv*, April 17, 2025, https://arxiv.org/abs/2504.12914.

[2] Scott Singer, "How the UK Should Engage China at AI's Frontier," Carnegie Endowment for International Peace, October 18, 2024, https://carnegieendowment.org/posts/2024/10/lammy-china-ai-safety-cooperation.

military application risks, what rules to use for reducing policy uncertainty, and what institutions to use for cushioning social impacts.

Given that relevant research findings have entered the policy process through channels such as expert appointments, policy consulting, congressional hearings, government-commissioned research, and public communication, future interactions between China and the West on AI may unfold along three parallel tracks: first, in areas such as high-end chips, cloud computing, frontier model capabilities, and AI-enabled military deployment, forming more institutionalized and finely calibrated security controls around the acquisition and diffusion of key capabilities; second, in areas such as Global South capacity building and infrastructure export, engaging in market and standards competition around who can provide more usable, more affordable, and more locally embeddable solutions; and third, on high-politics issues such as model verification, catastrophic risks, and nuclear strategic stability, preserving limited safety dialogue and technical cooperation space out of the practical need to prevent loss of control and reduce miscalculation. To this end, China should rationally and objectively view U.S.-China competition and cooperation in the AI field, and while promoting the healthy and orderly development of China's AI, actively share AI governance practical outcomes through the "Belt and Road" digital cooperation network, provide solutions for the Global South, and in broader international cooperation, develop higher regulatory compatibility and mutual recognition capacity, expanding the global influence of its technology ecosystem and governance rules.

Actors and Pathways

Research on the Role and Strategies of International Organizations in the AI Governance Process

*Feng Shuai Zhang Tianyu**

Abstract In the current global AI governance process, international organizations play an important role. There are three types of international organizations participating in AI governance, universal intergovernmental organizations represented by the United Nations; multilateral national mechanisms represented by the G20; and various functional technical organizations. International organizations play an important role as rule makers, platform builders, resource coordinators, and supervision promoters in the AI governance process. They use four major strategies: policy guidance, project cooperation, capacity building, and publicity and promotion to play their role in the governance process. International organizations mainly adopt three methods to participate in governance activities, namely, building platforms and forums, relying on neutrality and appeal to build consensus; formulating standards and ethical frameworks, relying on authority and professionalism to establish global rules; and promoting capacity building and project piloting, based on inclusive and development missions to bridge the digital divide. These three methods are connected to each other and form a full-chain governance system of "consensus formation-rule establishment-implementation." Although international organizations have played an irreplaceable role in promoting global AI governance, they still face structural limitations such as insufficient execution, differences of interests, and concentration of resources. There is still a need to strengthen institutional integration and innovation,

* Feng Shuai, Deputy Director of the Institute for International Strategy and Security, and Secretary-General of the Research Center for Cyberspace Governance, Shanghai Institutes for International Studies (SIIS), Associate Research Professor; Zhang Tianyu, Master's Degree Candidate, Shanghai Institutes for International Studies. This paper is a phased outcome of the National Social Science Foundation project "Global AI Governance Process and China's Participation Pathways" (Project No.: 24BGJ004).

further leverage the role of international organizations, and promote the global AI governance process.

Keywords AI Governance; International Organization; Role; Strategy; Full-chain Governance System

Introduction

AI governance has become a core issue in global public policy and global governance in the digital age. As an institutional arrangement and action systems for regulating the R&D, deployment, application, and risk control of AI technologies, AI governance employs a diverse range of governance instruments, including legal regulations, regulatory frameworks, ethical guidelines, industry standards, multilateral agreements, and social oversight, to reconcile and balance multiple value objectives, such as technological innovation, industrial development, public safety, human rights protection, and social equity[1]. Due to the inherent characteristics of artificial intelligence, such as borderless diffusion, cross-domain penetration, and systemic risk spillover, key challenges including cross-border data flows, algorithmic discrimination and bias, security of cutting-edge AI models, cyber threats, and lethal autonomous weapon systems all transcend the jurisdictional capacity and governance boundaries of individual sovereign states. This means that AI governance has been inherently transnational and global from its inception, making effective regulation unattainable through fragmented national oversight alone. As a result, multi-stakeholder collaborative governance has become an inevitable choice for addressing the systemic risks posed by artificial intelligence[2]. Consequently, various types of international organizations play a crucial role in the AI governance process.

However, compared with the voluminous research on AI governance by sovereign states, analyses of the role of international organizations in this process remain relatively insufficient. Existing studies mostly focus on case analyses of specific international organizations' participation in AI governance, but lack systematic assessment of their effectiveness, limitations, and constraints[3]. Moreover, many institutional blueprints for building

[1] Allan Dafoe, *AI Governance: A Research Agenda*, Oxford: Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2018, p. 4.
[2] Beatrice Christie et al., "Regulating Lethal Autonomous Weapon Systems: Exploring the Challenges of Explainability and Traceability", *AI & Ethics*, Vol. 4, No. 2, 2024, pp. 229-245.
[3] Relevant studies include: Hadrien Pouget et al., "Global AI Governance: Barriers and Pathways Forward," *International Affairs*, Vol. 100, No. 3, 2024, pp. 1275-1296.; Tatevik Davtyan, "The U.S. Approach to AI Regulation: Federal Laws, Policies, and Strategies Explained," *Journal of Law, Technology & the Internet*, Vol. 16, No. 2, 2025, pp. 223-

governance systems around international organizations fail to adequately connect with the actual organizational behavior, strategies, and operational mechanisms in practice. This disconnection produces a weak dialogue between theory and practice, and insufficient exploration of the roles and functions of international organizations in AI governance.

Given this, this article focuses on various international organizations that are already broadly engaged in global AI governance and play key coordinating roles, aiming to explore their functions and governance strategies in the global AI governance. At the practical level, this research aims to provide empirical references for policymakers across different countries to improve global collaborative governance systems, fill policy gaps, and optimize governance pathways, thereby contributing to the completion of the critical landscape of global AI governance. At the theoretical level, it seeks to connect existing theoretical models with institutional blueprints, bridge the gap between empirical research and future design, further enrich interdisciplinary research at the intersection of AI governance and international organization theory, and provide academic support and pathway envisioning for building a more inclusive, efficient, equitable, and sustainable future global AI governance system, and ultimately promote the safe, orderly, and inclusive development of AI technologies within the framework of global governance.

I. International Organizations in the AI Governance Process: Categories and Characteristics

In the contemporary international relations system, with the increasing diversification of international actors, different subjects hold significantly divergent perceptions, positions, and evaluations of international organizations. The academic community and policy practitioners have yet to

---

259.; Alejandro Bellogín, Oliver Grau, Stefan Larsson, Gerhard Schimpf, Biswa Sengupta, and Gürkan Solmaz, "The EU AI Act and the Wager on Trustworthy AI," *Communications of the ACM*, Vol. 67, No. 12, 2024, pp. 21-25.; Hipolito Calero, "An Analysis of China's AI Governance Proposals," CSET, September 12, 2024, https://cset.georgetown.edu/article/an-analysis-of-chinas-ai-governance-proposals/.; Olajide Olugbade, "In Search of a Global Governance Mechanism for Artificial Intelligence (AI): A Collective Action Perspective," *Global Public Policy and Governance*, Vol. 5, 2025, pp. 139-161.; Matthijs Maas and Jose Jaime Villalobos, *International AI Institutions: A Literature Review of Models, Examples, and Proposals*, AI Foundations Report 1, Legal Priorities Project, 2023, pp. 1-48.; Olivia J. Erdelyi and Judy Goldsmith, "Regulating Artificial Intelligence: Proposal for a Global Solution," *Proceedings of the 2018 AAAI/ACM Conference on AI*, Ethics, and Society, 2018, pp. 95-101.; Mark Robinson, "The Establishment of an International AI Agency: An Applied Solution to Global AI Governance," *International Affairs*, Vol. 101, No. 4, 2025, pp. 1483-1502.

reach a unified and universally accepted definitional standard. To avoid conceptual disagreements interfering with subsequent analysis, and to ensure consistency and rigor in the analytical framework, this article first classifies and introduces the international organizations involved in AI governance, thereby delineating the core scope of this research.

Given the highly specialized and inherently transnational nature of AI technology, its governance structure exhibits a logical twofold dependency. On the one hand, the cross-border flow, global diffusion, and impact of AI on international security and economic systems render its governance inevitably dependent on inter-state coordination mechanisms. On the other hand, the complexity and professional threshold of AI technology make it difficult for any single country to independently complete rule-making and risk assessment, thereby necessitating reliance on professional communities and standard-setting bodies with technical authority. Against this backdrop, AI governance should not be understood merely as a traditional intergovernmental cooperation issue, but rather as a composite system simultaneously embedded in both state coordination and technical governance. Foreign policy makers likewise need to attach due importance to AI industry standards developed by authoritative bodies, as such standards may also influence their negotiations in international settings. Therefore, in studying international organizations in AI governance, it is necessary to break through the conventional perspective that focuses exclusively on intergovernmental organizations and incorporate both intergovernmental organizations and functional international organizations, such as standard-setting bodies, into the analytical framework simultaneously.

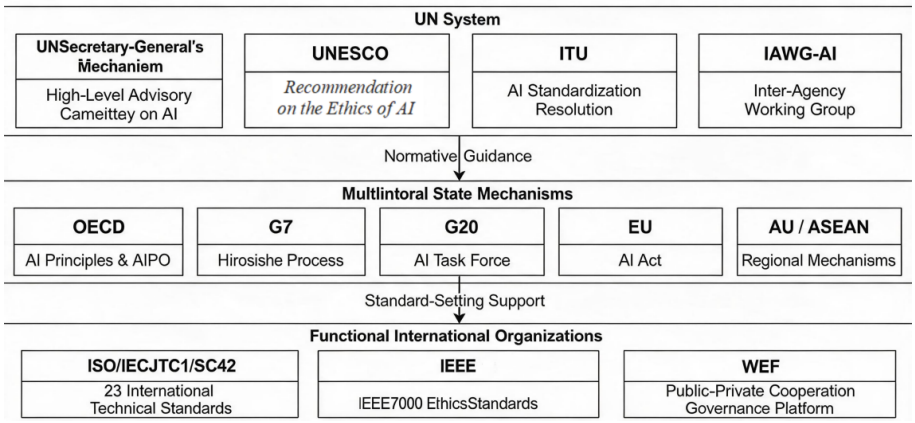
Within intergovernmental international organizations, significant differences persist, warranting further subdivision. Based on membership scope and influence structure, they can be divided into universal international organizations represented by the UN system, and multilateral state coordination mechanisms represented by the Organisation for Economic Co-operation and Development (OECD), the Group of Twenty (G20), and the Group of Seven (G7). The former typically possess relatively comprehensive institutional structures and broad membership bases, while the latter differ in membership size, representativeness, and issue focus. For instance, the G7 is composed primarily of developed countries, exhibiting a high degree of consensus and policy coordination efficiency. In contrast, the G20 encompasses both developed countries and emerging economies, offering broader representativeness while facing higher complexity in policy coordination. This differentiation reflects structural distinctions among different types of interstate mechanisms in terms of governance scope and influence.

Although from a strict legal perspective, the G7 and G20 do not fully conform to the traditional definition of international organizations, as they lack treaty bases and permanent secretariats, and thus are regarded as "informal mechanisms" or inter-state forums. However, with the evolution of the global governance system, such informal mechanisms have in practice become important platforms for international policy coordination. For example, the G20, as the premier forum for international economic cooperation, exerts extensive influence on global economic, climate, and digital governance issues. Therefore, excluding them solely on the basis of legal form would fail to accurately reflect the institutional structure of contemporary global governance.

Meanwhile, the role of functional international organizations in AI governance is equally indispensable. Standard-setting institutions represented by the International Organization for Standardization (ISO) are non-governmental international organizations whose core function is to formulate globally applicable technical standards. These standards provide a "common language" for digital technologies and play a foundational role in ensuring technological compatibility and security. In the field of AI, technical standards directly affect algorithm design, data governance, and system interoperability, thereby effectively participating in the rule-making process. Consequently, such organizations are not merely technical implementers but constitute a critical source of authority within the global governance system.

In summary, given the transnational nature and technical complexity of AI governance, it is necessary to construct an analytical framework encompassing multiple types of actors. Specifically, it should simultaneously include:

Universal intergovernmental organizations represented by the UN system; multilateral state mechanisms represented by the OECD, G7, and G20; and functional technical organizations represented by ISO and the International Electrotechnical Commission (IEC) (as shown in Figure 1). This classification not only reflects differences among organizations in terms of membership composition, functional positioning, and scope of influence, but also highlights two core dimensions of AI governance: first, the political organizations that primarily participate in diplomatic decision-making; and second, the standard-setting organizations that address current AI compliance issues through standardized. This is even more in line with the basic structural features of contemporary AI governance as a multi-level institutional complex.



Source: Compiled by the author.

Figure 1 Classification of International Organizations Participating in AI Governance

1. The UN System

In articulating the construction of a global AI governance system, the United Nations is an indispensable mechanism that is destined to play a significant role in the AI governance architecture. UN Secretary-General António Guterres has consistently emphasized the far-reaching implications of emerging technologies such as AI and has actively promoted the development of a trustworthy, human-rights-based, safe, sustainable, and peace-promoting AI framework. In 2018, Guterres spearheaded the establishment of the High-level Panel on Digital Cooperation, dedicated to addressing global challenges posed by the Internet, AI, and other digital technologies[1]. In June 2020, the Panel's outcomes were formally released as the *Roadmap for Digital Cooperation*, which proposed the creation of a multi-stakeholder advisory body for global AI cooperation[2], and in 2022, Amandeep Singh Gill was formally appointed as the Technology Envoy, tasked with coordinating global digital cooperation and integrating AI governance into the core international agenda, thereby strengthening the UN's top-level coordination function in this domain[3]. In 2024, the UN further established the High-level Advisory Body on Artificial Intelligence, composed of 39 cross-disciplinary experts, focusing on building a global AI governance framework

[1] "Secretary-General's High-level Panel on Digital Cooperation," United Nations, July 12, 2018, https://www.un.org/en/sg-digital-cooperation-panel.
[2] United Nations Secretary-General, *Roadmap for Digital Cooperation*, New York: United Nations, 2020, p. 25.
[3] United Nations Secretary-General, "Secretary-General Appoints Amandeep Singh Gill of India Envoy on Technology," United Nations, June 10, 2022, https://india.un.org/en/193551-secretary-general-appoints-amandeep-singh-gill-india-envoy-technology.

and studying the establishment of an international AI regulatory mechanism modeled on the International Atomic Energy Agency (IAEA)[1]. The Advisory Body explicitly asserts that the governance framework must be grounded in scientific evidence, adopt interdisciplinary collaborative models, and ensure that developing countries have substantive participation in rule-making, thereby remedying the representational imbalance inherent in traditional multilateral mechanisms.

Various UN agencies have also formed a synergistic network in AI governance, demonstrating significant cognitive authority and operational autonomy. In 2021, the United Nations Educational, Scientific and Cultural Organization (UNESCO) adopted the *Recommendation on the Ethics of Artificial Intelligence*, which became the world's first universally guiding ethical framework for AI, establishing eleven core principles including human rights protection, transparency, and accountability, and explicitly prohibiting high-risk applications such as social scoring and mass surveillance[2]. To date, the Recommendation has guided numerous member states in formulating their domestic AI ethics policies and has continuously monitored implementation effects through the "Ethical Impact Assessment" mechanism. The International Telecommunication Union (ITU), as the UN's specialized agency for information and communication technologies, plays a key role in technical standard-setting. In 2024, ITU issued its first dedicated resolution on AI, accelerating the global AI standardization process and collaborating with ISO to develop deepfake detection tools[3]. Concurrently, in 2024, UNESCO and ITU jointly created the "AI for Good" collaboration platform, which has cataloged 53 global AI application demonstration cases and facilitated coordinated action by over 40 UN agencies in the governance domain, effectively enhancing policy coherence and cross-agency coordination efficiency[4].

Moreover, given the large number of participating entities and the broad coverage of governance actions within the UN system, and to further integrate dispersed governance resources and improve coordination efficiency, the UN system has been advancing institutional consolidation. In March 2021, it launched the Inter-Agency Working Group on Artificial Intelligence (IAWG-

---

[1] UN Secretary-General's High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report*, New York: United Nations, 2024, pp. 7-10.
[2] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris: UNESCO, 2021, pp. 17-23.
[3] International Telecommunication Union, *Resolution 101: Standardization Activities of the ITU Telecommunication Standardization Sector on AI Technologies*, Geneva: ITU, 2024, p. 4.
[4] International Telecommunication Union, *AI for Good Global Summit 2024 Snapshot Report*, Geneva: ITU, 2024, p. 2.

AI), co-led by UNESCO and ITU through an adopted term of reference. In May 2024, it released the *United Nations System White Paper on AI Governance*, systematically mapping institutional models, functional boundaries, and existing international normative frameworks[1]. The Working Group integrates resources from UNESCO, ITU, UNCTAD, and other entities to coordinate across three major domains, including AI ethics, economic equity, and industrial application, driving governance concepts from principled advocacy toward practical implementation.

Overall, leveraging its existing global governance architecture and adopting an approach of high-level mechanism leadership, multi-agency coordination, and full-chain coverage, the UN has formed a multi-layered, multi-dimensional action network in AI governance. Despite challenges such as diversified initiatives and complex procedures, the UN, with its broad membership base, cognitive authority, and multilateral convening power, remains a core platform for building global AI governance consensus, ensuring equitable development, and preventing technological risks, playing an irreplaceable integrative and coordinating role in the fragmented global governance landscape.

2. Multilateral State Mechanisms

In the global AI governance process, multilateral state mechanisms play an irreplaceable and significant role. The OECD, with its formidable cognitive authority and normative capacity, plays a central and leading role in global AI governance. As early as 2016, the OECD Committee on Digital Economy Policy began exploring the necessity of formulating AI principles, and in May 2018 established an expert group. In 2019, the OECD released the world's first intergovernmental AI policy guidelines, the *OECD AI Principles*, which were updated in 2024, establishing core principles including inclusive growth, respect for human rights, transparency and explainability, robustness and security, and clear accountability, thereby laying the institutional foundation for trustworthy AI. Beyond OECD member countries, Argentina, Brazil, Costa Rica, Malta, Peru, Romania, and Ukraine have also endorsed these principles, demonstrating their broad international influence[2]. Launched in 2020, the "AI Policy Observatory (AIPO)" serves as an open knowledge platform providing data, case studies, and policy analysis support to assist countries in implementing the OECD AI Principles[3]. Additionally, the OECD

[1] UN Inter-Agency Working Group on Artificial Intelligence, *United Nations System White Paper on AI Governance*, Geneva: ITU/UNESCO, 2024, p. 15.
[2] OECD, *Recommendation of the Council on Artificial Intelligence*, Paris: OECD, 2024, pp. 8-9.
[3] OECD, *OECD AI Policy Observatory (OECD.AI)*, Paris: OECD, 2020, p. 12.

Network of Experts on AI (ONE AI), as a cross-disciplinary, multi-stakeholder collaboration platform, is dedicated to providing specialized AI policy recommendations and promoting international cooperation, while engaging in deep collaboration with the G7, ISO/IEC, and other mechanisms to facilitate regulatory interoperability and ideational diffusion. As of 2024, over 70 countries have committed to adhering to the *OECD AI Principles*.

The G7 and G20, as important multilateral coordination platforms, differ in membership composition and issue focus, yet both exert profound influence on global AI governance. The G7, comprising predominantly developed countries, has progressively elevated AI issues to the leaders' summit level since 2016, adopting documents such as the *Charlevoix Common Vision* to establish a people-centered, safe, and trustworthy development direction, while proposing twelve ethical and safety commitments and promoting the formation of the Global Partnership on Artificial Intelligence (GPAI)[1]. In the 2023 "Hiroshima AI Process," the G7 further issued *International Guiding Principles* and *a Code of Conduct for Developers*, focusing on full-lifecycle risk management of advanced models; this platform was the first to establish a self-regulatory framework for frontier models[2]. Although the G20 started somewhat later, it has incorporated AI into the core agenda of global economic governance, placing greater emphasis on the integration of technological innovation with the Sustainable Development Goals (SDGs). Driven by its Digital Economy Working Group, the G20 Rio de Janeiro Leaders' Declaration in 2024 reaffirmed bottom-line requirements including ethics, fairness, and human rights protections, while emphasizing the use of AI to improve public services and maintain information integrity[3]. The 2025 South African summit further established a "G20 Task Force on AI" to advance policy coordination and standard mutual recognition, while paying attention to the equitable participation of developing countries in supply chains[4]. Overall, the G7 places greater emphasis on rule-shaping and frontier risk governance, whereas the G20 prioritizes inclusive development and multilateral cooperation, incorporating the concerns of the Global South and efforts to bridge the digital divide into the broader framework.

---

[1] G7 Leaders, *Charlevoix Common Vision for the Future of Artificial Intelligence*, Charlevoix: G7, 2018, pp. 2-3.
[2] G7 Leaders, "G7 Leaders' Statement on the Hiroshima AI Process," October 30, G7 Information Centre, 2023, https://www.g7.utoronto.ca/summit/2023hiroshima/231030-ai.html.
[3] G20 Leaders, *G20 Rio de Janeiro Leaders' Declaration*, Rio de Janeiro: G20, 2024, pp. 20-21.
[4] G20 South Africa Presidency, "First Meeting of Task Force 3 on Artificial Intelligence, Data Governance and Innovation for Sustainable Development," G20, February 25, 2025, https://www.g20.org.za/track-news/first-meeting-of-the-task-force-on-artificial-intelligence-data-governance-and-innovation-for-sustainable-development/.

The European Union (EU), as a pioneer in regulatory innovation, has established a binding and forward-looking institutional system through its supranational governance model. The European Commission has demonstrated a high degree of autonomy, consistently staying ahead globally from its 2018 strategic blueprint to the establishment of a High-Level Expert Group, and then to the 2019 *Ethics Guidelines for Trustworthy AI*[1]. The EU AI Act, which entered into force in 2024, implements systematic risk-tiered regulation: it comprehensively prohibits unacceptable applications such as social scoring and mass surveillance, imposes stringent compliance obligations on high-risk scenarios, and sets transparency requirements for general-purpose AI models[2]. Concurrently, it established the European AI Office to strengthen implementation oversight and technical assessment[3]. Beyond regulation, the EU also launched the Invest AI Initiative, mobilizing a total of €200 billion, including investment in computing infrastructure and talent development programs, effectively addressing computing power shortages and enhancing enterprise competitiveness. This regulatory-innovation nexus model not only reinforces the protection of human rights but also creates a demonstration effect for the paradigm of trustworthy AI[4].

At the regional level, the African Union (AU) and the Association of Southeast Asian Nations (ASEAN) have incorporated AI into their core regional development agendas, pursuing localized and inclusive pathways through strategic planning and capacity building. The AU, through its Continental AI Strategy, places driving sustainable development and strengthening local capacity at the center, deploying computing centers and cultivating local talents through high-level dialogues, with a focus on improving service efficiency in agriculture, healthcare, and education[5]. ASEAN has issued governance guidelines and launched the "AI Ready ASEAN" program, conducting large-scale digital literacy training covering 5.5 million people, while establishing working groups and smart city pilots that provide operational guidance on transparency, fairness, and human-

[1] EU High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, Brussels: European Commission, 2019, p. 26.
[2] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act)," *Official Journal of the European Union*, August 1, 2024, OJ L 2024/1689.
[3] European Commission, "Commission Decision of 24 January 2024 establishing the European Artificial Intelligence Office," European Commission, February 14, 2024, https://eur-lex.europa.eu/eli/C/2024/1459/oj.
[4] European Commission, "EU Launches InvestAI Initiative to Mobilise €200 Billion Investment in Artificial Intelligence," European Commission, February 11, 2025, https://digital-strategy.ec.europa.eu/en/news/eu-launches-investai-initiative-mobilise-eu200-billion-investment-artificial-intelligence.
[5] African Union, *Continental Artificial Intelligence Strategy*, Addis Ababa: African Union, 2024, pp. 1-35.

machine collaboration. These regional mechanisms adopt a phased and pragmatic approach, offering adaptable and long-term sustainable practical models for the global community[1].

3. Functional International Organizations

In the global AI governance landscape, in addition to the core role played by intergovernmental mechanisms, international standard-setting organizations, professional technical bodies, and non-state actors constitute equally critical pillars. Leveraging technical expertise, institutional flexibility, and industry cohesion, they open up diversified governance pathways in standard-setting, ethical norms, risk governance, and safe R&D, serving as crucial links between government frameworks and industrial practice. The work of these organizations is often highly technical in nature. Nevertheless, standard-setting, particularly international standards, undoubtedly exerts a profound impact on the R&D and diffusion of AI technologies, which extends to the corresponding regulatory and governance domains. Moreover, the cognitive authority of these standard-setting bodies directly or indirectly influences the policymaking processes of other actors. The joint ISO/IEC subcommittee on AI, ISO/IEC JTC 1/SC 42, has been systematically advancing global AI technology standardization since 2017, and to date has published 23 authoritative international standards[2], covering core areas such as algorithmic performance evaluation, data lifecycle management, security, and explainability. Among these, the ISO/IEC 42001 standard establishes a unified methodological framework for AI system security assessment, which has been widely adopted by global technology enterprises such as Tesla and NVIDIA, effectively enhancing cross-border technical coordination efficiency and providing underlying technical support for the standardized development of the AI industry[3]. As the core global technical standard-setting organization, ISO/IEC consistently adheres to a "holistic ecosystem philosophy," balancing technical specifications with ethical and social concerns, and engages in deep collaboration with global governance actors such as the OECD and the European Commission to promote the two-way convergence of standards and governance rules.

The Institute of Electrical and Electronics Engineers (IEEE) is the most influential professional standards organization in the field of global AI ethics

---

[1] ASEAN, *ASEAN Guide on AI Governance and Ethics*, Jakarta: ASEAN, February 2024, pp. 11-18.
[2] ISO/IEC JTC 1/SC 42, "Artificial Intelligence," ISO, https://www.iso.org/committee/6794475.htm.
[3] ISO/IEC, ISO/IEC 42001:2023: *Information Technology—Artificial Intelligence—Management System*, Geneva: ISO/IEC, 2023, pp. 1-64.

and technical governance, and its work complementarily supports that of ISO/IEC. Since 2016, IEEE has formally launched a global initiative on AI ethics, with the core objective of promoting "ethically aligned design" of autonomous and intelligent technologies, aiming to build global consensus and prioritize human well-being in technological development[1]. Within this framework, IEEE introduced the landmark IEEE 7000 series of standards, establishing a full-lifecycle AI ethics design system and systematically proposing 11 key technical specifications, among which core indicators such as explainability by design, robustness by design, fairness assurance, and value alignment have become global industry benchmarks[2]. To facilitate standard implementation, IEEE has established a comprehensive ethics compliance certification system, having completed assessment and verification of over 300 AI projects from global technology companies including IBM and Intel. Concurrently, IEEE has continued to expand its governance network, establishing in 2018 the Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS), which brings together over 70 institutions worldwide to create a cross-border, cross-disciplinary platform for exchange and collaboration on ethical standards. It has also set up dedicated research funds focusing on frontier topics such as AI value alignment, algorithmic safety, and human-machine collaboration, embedding ethical principles deeply throughout the entire process of technology R&D, testing, and deployment, codifying ethical requirements through technical standards and providing actionable, verifiable practical solutions for global trustworthy AI[3].

The World Economic Forum (WEF), on the other hand, innovates governance practices through public-private partnership models. Through the Davos AI Governance Summit, it builds a tripartite collaboration platform among government, industry, and academia. In 2024, it facilitated the signing of the *AI Responsible Innovation Pledge* by technology companies including Microsoft and Google, promoting the implementation of industry self-regulation mechanisms. Concurrently, the *AI Risk Map* released that year systematically identified twelve categories of core risks including technology misuse and labor market restructuring, providing empirical evidence for global policymaking and enhancing risk governance capacity through flexible

[1] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Piscataway, NJ: IEEE, 2016, pp. 1-280.
[2] IEEE Standards Association, "IEEE P7000 Standards Projects," IEEE, March 14, 2026, https://ethicsstandards.org/p7000/.
[3] IEEE et al., "Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS)," OCEANIS, March 17, 2026, https://ethicsstandards.org/.

mechanisms[1].

Overall, international standardization organizations and professional non-state actors represented by ISO/IEC and IEEE, have developed a full-chain governance capacity covering standards, ethics, risk, and security through the formulation of technical specifications, establishment of ethical benchmarks, promotion of industry self-regulation, and advancement of safety technologies. Complementing and coordinating with intergovernmental mechanisms, they drive regulatory convergence and practical implementation within the fragmented global governance landscape, continuously strengthening the professionalism, inclusiveness, and operability of AI governance.

In sum, against the backdrop of a polycentric, multi-level governance architecture that has become an academic consensus, international organizations are not merely an alternative for future governance but rather the central hubs and practical carriers connecting national governance with global co-governance. Various international organizations, including the UN, OECD, G7, G20, GPAI, ISO, and IEC, have already undertaken AI governance practices at global, regional, and thematic levels, continuously playing critical roles in developing governance norms and building technical standards, serving as important nexuses connecting disparate national governance systems and promoting cross-national collaborative cooperation. Situated in the practical arena of global governance, they not only embody the political will of states to reach consensus but also continuously adapt their governance strategies to technological change and geopolitical circumstances; they provide a wealth of observable and assessable governance action models while also exposing institutional shortcomings and collaborative dilemmas in real-world operations.

II. International Organizations Driving the AI Governance Process: Roles and Strategies

In the AI governance process, international organizations often leverage their multilateral coordination advantages to assume four core roles, namely rule-setter, platform-builder, resource-coordinator, and supervisor-promoter, thereby establishing a governance framework with clear responsibilities and synergistic complementarity. Through four key strategies, namely policy guidance, project cooperation, capacity building, and advocacy and outreach, international organizations translate their role orientations into concrete

[1] World Economic Forum, *AI Governance Alliance: Briefing Papers*, Geneva: WEF, 2024, pp. 10-51.

governance actions. These actions not only delineate ethical and safety baselines for AI technological innovation, but also provide support for technological inclusivity and balanced global development, promoting the safe, orderly, and inclusive global collaborative development of AI.

1. Role Shaping of International Organizations

Against the backdrop of rapidly advancing global AI technology and increasingly urgent governance demands, international organizations have proactively assumed multiple core roles, primarily through rule-setting, platform-building, resource-coordination, and supervision and promotion. These four dimensions provide comprehensive support for the healthy, orderly, and sustainable development of AI. Each role has a clear division of labor and works in synergy, collectively constructing an important support system for global AI governance.

The first is the crucial role of rule-setter, whose core function is to formulate the ethical, legal, and technical standards for AI, thereby providing safeguards for the development of AI technology. The *Recommendation on the Ethics of Artificial Intelligence* adopted by UNESCO laid the foundation for global AI ethical norms, establishing 11 principles and 11 action areas[1], building broad consensus at the intergovernmental level, and emphasizing core values such as human rights, fairness, inclusiveness, transparency, and accountability. ISO, on the other hand, focuses on technical standard-setting, covering key areas such as AI terminology, model evaluation, and data management, effectively enhancing interoperability and compatibility among different systems. For example, its intelligent transportation standards ensure the collaborative operation of equipment from different manufacturers, providing support for efficient and safe smart transportation[2]. These standards not only prevent technology misuse, protect human rights, and prevent personal data breaches and algorithmic discrimination, but also reduce corporate R&D costs and market access barriers, enhance market trust, and promote the large-scale and standardized development of the AI industry.

Second, as platform-builders, international organizations facilitate cooperation and exchange in the AI field among countries through various platforms, promoting technological innovation and progress. The United Nations convenes AI-related conferences, bringing together representatives from governments, enterprises, research institutions, and social organizations

---

[1] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris: UNESCO, 2021, pp. 1-43.

[2] ISO/IEC, ISO/IEC 42001:2023: *Information Technology—Artificial Intelligence—Management System*, Geneva: ISO/IEC, 2023, pp. 1-9.

to explore opportunities for AI applications in sustainable development, climate change, public health, and other areas. The World Economic Forum (WEF), through its various seminars and workshops, provides channels for global elites from politics, business, and academia to exchange ideas and share experiences, discussing the economic, social, and cultural impacts of AI and leading industry trends. These platforms effectively break down information barriers, enable the sharing of technological achievements and application experiences across countries, foster complementarity of strengths, tackle technical challenges, and promote deep integration of industry, academia, and research, helping enterprises grasp market demand and accelerate the commercialization of scientific and technological outcomes[1].

At the same time, leveraging their unique resource advantages, international organizations also play the role of resource-coordinator, providing strong support for AI technology innovation and application. The United Nations Industrial Development Organization (UNIDO) focuses on promoting AI applications in the industrial sector, integrating resources from various parties to help developing countries improve their industrial intelligence levels. Its joint establishment with Huawei and other enterprises of the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM Global) creates a collaborative sharing platform to promote innovative applications of AI technology in the industrial sector. Alliance members jointly invest funds and technology, carry out related R&D projects, provide technical training to industrial enterprises in developing countries, and help them enhance production efficiency and competitiveness. This resource integration not only provides sufficient R&D funding for AI enterprises and research institutions, accelerating technological iteration, but also enables technology sharing and complementarity, expanding AI technology into more fields[2].

Finally, international organizations act as supervisors and promoters, focusing on the safe, reliable, and sustainable development of AI technology. They strengthen full-process supervision through the formulation of policies and measures to ensure that technology benefits humanity. The International Telecommunication Union (ITU) strengthens AI supervision through standard-setting and policy development. At the 2024 World Telecommunication Standardization Assembly, it issued its first resolution on AI, emphasizing the construction of a responsible, safe, and inclusive AI system, clarifying the principles and standards that technological development must follow, and

---

[1] World Economic Forum, *AI Governance Alliance: Briefing Papers*, Geneva: WEF, 2024, pp. 1-59.

[2] UNIDO and Huawei, "UNIDO and Huawei Launch the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM Global)," Huawei, July 6, 2023, https://www.huawei.com/en/news/2023/7/aim-global-announcement.

requiring that the design, development, and application of AI systems consider safety, privacy, and ethics throughout the process[1]. Such supervision can not only promptly identify potential risks such as algorithmic discrimination and data security issues, thus avoiding negative social impacts, but also guide the R&D of environmentally friendly and energy-efficient AI technologies, promoting coordinated development between technology, society, and the environment.

2. Strategy Selection of International Organizations

Having clarified the multiple roles of international organizations, it is necessary to further examine the specific pathways through which they translate role orientations into practical actions. Different strategic choices produce varying effects in reality. International organizations primarily employ four types of tools: policy guidance, project cooperation, capacity building, and advocacy and outreach. These tools form a functionally complementary and synergistically linked implementation system.

Policy guidance is the core strategy for international organizations to promote AI development, aiming to provide directional guidance for national AI development through the issuance of policy documents and ethical guidelines, thus assisting countries in formulating development strategies tailored to their national conditions. The European Union, as a pioneer in AI policy-making, has had a profound global impact with its relevant policies. Its 2020 *White Paper on Artificial Intelligence* proposed the construction of a safe and trustworthy AI regulatory framework, laying the foundation for subsequent policies[2]. In March 2024, the European Parliament adopted the EU AI Act, the world's first comprehensive AI regulatory legislation, which implements a risk-based classification system for AI systems, specifying compliance requirements for high-risk systems. This not only regulates the EU's AI industry but also provides a regulatory model that can be referenced by other countries worldwide[3]. In addition, the OECD plays an important role in policy guidance, with its AI-related reports and guidelines providing in-depth analysis of industry trends and policy environments, offering scientific evidence for member states' policymaking. The AI ethical principles adopted by member states in 2019 clearly established core requirements such as

---

[1] International Telecommunication Union, *Resolution 101: Standardization Activities of the ITU Telecommunication Standardization Sector on AI Technologies*, Geneva: ITU, 2024.

[2] European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, Brussels: European Commission, 2020, pp. 11-27.

[3] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act)," *Official Journal of the European Union*, August 1, 2024, pp. 1-144.

human-centeredness, transparency, explainability, fairness, and security, providing ethical guidance for AI R&D and application in member countries. Through policy dialogue and experience-sharing, the OECD also helps member states improve their own AI policy systems[1].

Project cooperation strategies focus on resource integration and technology implementation. International organizations launch various cooperative projects to pool technologies, funds, and talents from different countries, moving AI from concept to practice. The Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM Global), spearheaded by UNIDO, serves as a core vehicle for project cooperation. With collaborative innovation at its core, the alliance focuses on scenario-based applications and commercialization of AI technologies in the industrial sector. Within the framework of the alliance, parties jointly advance special actions for industrial intelligence upgrading. Relying on technological support from companies such as Huawei, they provide customized and intelligent solutions for industrial enterprises in developing countries, helping these enterprises optimize production processes and upgrade automation, thereby significantly improving production efficiency and product quality.

Capacity building strategies aim to address the gaps in AI development among countries. International organizations enhance the AI technical capabilities and talent quality of various countries through training, education, and technical assistance. IEEE continues to carry out multi-level AI education and training. Its IEEE Academy on Artificial Intelligence provides classic and modern AI knowledge modules covering diverse application scenarios, and learners can obtain certificates upon completion. Meanwhile, through online courses, workshops, and seminars covering topics such as chip design and edge computing, IEEE helps researchers and technical engineers stay abreast of industry frontiers[2]. UNESCO focuses on the needs of developing countries. In 2024, it organized a conference for African teachers and education officials in Dakar, Senegal, promoting the AI competency framework through "South-South Cooperation," supporting 25 African countries in formulating digital and AI education policies, and providing teacher training and teaching resources. At the same time, it has established AI application demonstration centers in several developing countries, offering technical consulting and training to help local areas apply AI in agriculture, health, education, and other fields, thereby

[1] OECD, *Recommendation of the Council on Artificial Intelligence*, Paris: OECD, 2024, pp. 1-12.
[2] IEEE, "IEEE Academy on Artificial Intelligence", IEEE Learning Network, March 22, 2026, https://iln.ieee.org/public/contentdetails.aspx?id=A9C06EC22F0C4EB8B869FD1C7355C1B1.

improving production efficiency and service levels[1].

Advocacy and outreach strategies emphasize raising public awareness and building consensus on AI development. International organizations popularize AI technical knowledge and application value through events and research reports. The WEF's AI seminars and industry summits invite global elites from politics, business, and academia to engage in in-depth discussions on AI industry trends and application scenarios, effectively guiding public attention to AI development. The ITU, through thematic research reports and case compilations, systematically showcases AI implementation results across various sectors. Its report *AI for Good Impact Report* details innovative AI practices in agriculture, healthcare, transportation, and other fields, thoroughly analyzing its role in promoting global sustainable development. This not only provides practical references for governments, enterprises, and research institutions but also broadens public understanding of AI's value, facilitating widespread promotion and consensus-building around AI technology[2].

Table 1 Relationship between Roles and Strategies of International Organizations in the AI Field

Core Role	Core Function of the Role	Applicable Strategies	Explanation of Strategy Application
Rule-setter	Formulate ethical, legal, and technical standards; establish AI governance norms	Policy guidance; Advocacy and outreach	Issue ethical principles and regulatory frameworks through policy guidance; promote standards and consensus through advocacy and outreach to expand rule influence
Platform-builder	Build multilateral communication platforms; facilitate transnational cooperation and consensus-building	Advocacy and outreach; Policy guidance	Create dialogue spaces through forums and summits as forms of advocacy and outreach; guide the formation of common governance directions through policy dialogue
Resource-coordinator	Integrate funds, technology, and talent; promote optimal resource allocation	Project cooperation; Capacity building	Pool global resources for joint R&D through project cooperation; deliver technical and talent support to developing countries through capacity building
Supervisor-promoter	Monitor AI safety risks; promote safe and sustainable development	Policy guidance; Capacity building; Advocacy and outreach	Set regulatory requirements through policy guidance; enhance compliance capacity through capacity building; reinforce governance concepts through advocacy and outreach

Source: Compiled by the author.

[1] UNESCO, *AI Competency Framework for Teachers*, Paris: UNESCO, 2024, pp. 1-52.
[2] International Telecommunication Union, *AI for Good Impact Report*, Geneva: ITU, 2024, pp. 1-80.

III. Approaches and Logic of International Organization Participation in AI Governance

The preceding sections have delineated the roles and strategies undertaken by international organizations from a functional perspective. Description of behavior alone, however, falls short of explaining the sources of its effectiveness. Why do international organizations choose particular pathways over others? what intrinsic logics of alignment exist between these pathways and their own organizational attributes? and what observable governance effects ultimately emerge from such alignment? Advancing the analysis to this level aims to bridge the gap between behavioral description and mechanistic explanation, thereby providing a deeper theoretical foundation for understanding the irreplaceability of international organizations in global AI governance.

It should be clarified that international organizations' selection of three pathways, namely building platforms and organizing forums, formulating standards and ethical frameworks, and promoting capacity building and pilot projects, is by no means arbitrary. Rather, it is a rational choice grounded in the core challenge of global AI governance, the full governance cycle, and the functional divisions inherent to international organizations themselves. These three pathways are mutually reinforcing and complementary, jointly constituting a full-chain governance system that covers the entire spectrum from consensus building through rule establishment to implementation. The interrelated links are interdependent and indispensable.

1. Building Platforms and Organizing Forums: The Logic of Neutrality and Convening Power

AI governance involves diverse actors including states, enterprises, academic institutions, and civil society organizations, among which significant information asymmetries and deep trust deficits persist. The AI governance ecosystem thus presents a complex structure of interests and communication challenges. Against this backdrop, a neutral public sphere is needed to provide all stakeholders with a space for dialogue free from domination by particular political or commercial forces, thereby helping parties establish initial trust, explore grounds for consensus, and pave the way for subsequent cooperation. International organizations, by virtue of their inherent neutrality, transnational coordination capacity, and institutionally accumulated credibility over many years, have become the ideal subjects for assuming this responsibility. They are able not only to transcend the parochial interests of states and markets and build dialogue platforms from a more comprehensive and inclusive perspective, but also to provide structured

communication channels for diverse actors lacking mutual trust through well-established deliberative mechanisms and international norms, thereby effectively fulfilling their roles as bridges and coordinators in global governance.

The realization of this approach is highly contingent upon the alignment between the attributes and advantages of different international organizations. The World Economic Forum (WEF), for example, as a non-governmental organization, derives its core resources not from coercive legislative authority but from its robust public-private partnership networks and high-level convening capacity. The WEF's role orientation is that of "agenda-setter" and "relationship catalyst," rather than rule-maker. Its platforms, including the Davos AI Governance Summit it hosts, perfectly align with the organization's mission of "improving the state of the world" and its operational mechanism of advancing major issues through public-private collaboration[1]. High-level economic cooperation forums like the G20, by contrast, leverage the representativeness and political influence of their member economies to play distinctive roles at the macro level. While they do not directly formulate technical standards, they are able, through leaders' declarations and joint statements of principles, to establish inviolable "red lines" and core norms, including ethics, fairness, and transparency, for global AI governance, thereby providing critical political impetus and clear strategic direction for standard-setting organizations such as ISO and ITU[2]. This functional complementarity and division of labor enable various types of international organizations to play multi-layered and systematic roles in coordination within the global AI governance architecture.

The success of international organizations in AI governance through building platforms and organizing forums, and the broad positive impact they generate, hinge crucially on their neutrality, which is widely recognized by the international community. Neither the United Nations nor the WEF is perceived as an advocate for the interests of any particular state or corporation. This detached status offers a rare and necessary "safe zone for dialogue" to parties in competition or disagreement. Member states and enterprises are willing to participate actively in such forums not only because they can gain insights into cutting-edge policy signals and technological trends, but also because these forums serve as efficient platforms for showcasing their

[1] World Economic Forum, "World Economic Forum Launches AI Governance Alliance Focused on Responsible Generative AI," World Economic Forum, June 15, 2023, https://www.weforum.org/press/2023/06/world-economic-forum-launches-ai-governance-alliance-focused-on-responsible-generative-ai/.
[2] G20 Leaders, *G20 Rio de Janeiro Leaders' Declaration*, Rio de Janeiro: G20, November 18, 2024, https://www.g20.utoronto.ca/2024/241118-declaration.html.

governance propositions, influencing the formation of international rules, and embedding themselves in global cooperation networks. For international organizations themselves, hosting such forums is not merely a fulfillment of their function of building multilateral dialogue channels, but also a strategic avenue for expanding their influence, incorporating diverse perspectives, and consolidating their central position in the global governance system. Beyond neutrality, these organizations also rely on their brand reputation, agenda-setting capacity, and operational excellence to sustain the vitality and attractiveness of these mechanisms.

The positive effects of this organizational approach are far-reaching and multi-dimensional. First, the network effects are significant: forums bring together the world's most important policymakers, industry leaders, and academic experts, facilitating numerous informal contacts and opportunities for collaboration. Many important transnational AI governance initiatives and cooperative projects originate from such "serendipitous encounters" and dialogues at these high-level occasions. Second, the risk early warning function is prominent. The Global Risks Report published by the WEF, for example, has not only succeeded in elevating the ethical and safety risks of AI to the global decision-making agenda, but has also played a crucial role in agenda anchoring and policy guidance at critical moments, directing attention toward long-term challenges rather than immediate interests[1]. Finally, international organizations' own influence is reinforced through these platforms: by continuously hosting high-profile summits that capture global attention, they not only significantly enhance their brand visibility and policy influence, but also substantively amplify their advocacy and coordination capacity in global governance, thereby forming a self-reinforcing virtuous cycle.

2. Formulating Standards and Ethical Frameworks: The Logic of Authority and Professional Expertise

As a borderless technology, AI's inadequately regulated applications are continuously generating ethical conflicts, creating new technological barriers, and producing systemic risks in security and privacy. Against the backdrop of this widening governance gap, international organizations, by leveraging their professional authority, coordination mechanisms, and cross-domain resources, are able not only to provide urgently needed public goods to member states and exercise their distinctive advantages in norm-setting and international consensus-building, but also to expand their influence through agenda-setting, consolidate their central position in the global governance system, and secure

[1] World Economic Forum, *The Global Risks Report 2024*, Geneva: WEF, 2024, pp. 1-123.

sustained political and financial support for organizational maintenance and operation. In fulfilling their roles to establish globally applicable rules, international organizations therefore work on two fronts: on the one hand, they aim to curb unfair competition among states and reduce the institutional costs of global industrial chains; on the other, they seek to ensure that technological development aligns with the fundamental values of fairness, transparency, and human rights protection, guaranteeing that the evolution of AI technology genuinely serves the common well-being of all humanity.

This governance approach is highly congruent with the functional characteristics of the organizations themselves. UNESCO, for instance, whose mission is to promote international cooperation in education, science, and culture, derives a solid foundation for drafting global ethical recommendations from its institutional traditions and international prestige in norms and ethics. UNESCO adopted the *Recommendation on the Ethics of Artificial Intelligence* in 2021. Endorsed by 193 member states, the Recommendation establishes principles including transparency, fairness, accountability, privacy protection, and human rights and dignity, while encompassing multiple policy action areas such as data governance, gender, equity, environment and ecosystems, education and research, and health and social well-being[1]. It provides not only ethical principles but also mechanisms for monitoring and evaluation, readiness assessment, and ethical impact assessment to facilitate the translation of these principles into policies and practices. At the same time, the ISO/IEC 42001:2023 Artificial Intelligence Management System (AIMS), jointly issued by ISO and IEC, represents a more technically oriented and operationally robust standard-setting instrument. This standard specifies how organizations should establish, implement, maintain, and continually improve their AI management systems in the development, provision, or use of AI products or services. The standard includes specific requirements such as risk identification and management, impact assessment, system lifecycle management, transparency and accountability, and third-party supply chain oversight, all designed to ensure that organizations balance innovation and governance in their use of AI while enhancing trust and compliance[2].

These concrete cases, including the UNESCO Recommendation and ISO/IEC 42001, have reinforced the persuasiveness and practical utility of standards and ethical frameworks in practice. The UNESCO Recommendation, for example, not only sets forth ethical principles but also

[1] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris: UNESCO, 2021, pp. 1-43.
[2] ISO/IEC, ISO/IEC 42001:2023: *Information Technology—Artificial Intelligence—Management System*, Geneva: ISO/IEC, 2023, pp. 1-52.

establishes practical tools including monitoring and evaluation mechanisms, readiness assessments, and ethical impact assessments, enabling member states and organizations to translate abstract principles into actionable policies and institutions. The case of ISO/IEC 42001 is exemplified by the fact that after adopting the standard internally, different organizations, including large corporations and public institutions, establish AIMS that clearly define leadership responsibilities, planning processes, transparency policies, and monitoring and improvement mechanisms, thereby reducing risks in AI R&D and deployment while demonstrating their commitment to responsible AI to markets and the public.

The success of international organizations in formulating standards and ethical frameworks and the broad positive impact they generate hinge on the fact that their outputs do not rely on mandatory directives, but rather on constructing "global public goods" endowed with high authority and broad industry consensus. These standards are widely embraced by member states and enterprises, including Tesla and NVIDIA, for two key reasons. On the one hand, the standard-setting process is highly professional and inclusive, typically involving transnational expert communities, government representatives, industry actors, and academic institutions, which ensures technical feasibility and policy applicability. On the other hand, voluntary adoption of such standards by enterprises helps enhance product international compatibility, increase user and market trust, and avoid additional compliance costs and trade barriers arising from fragmented standards. In terms of regulatory mechanisms, in addition to member states' translation of international standards into domestic regulations, greater reliance is placed on market-driven mechanisms, including conformity certification, industry self-regulation and oversight, and supply chain constraints, to achieve soft yet effective standard implementation and continuous monitoring.

The positive effects of this governance approach are primarily manifested in three dimensions. First, authority is established: successfully leading the development of global standards greatly enhances the leadership position and discursive power of relevant international organizations in AI governance, making them indispensable core institutions in rule evolution. Second, value is realized: these organizations are able to organically inject their traditional core missions, such as UNESCO's protection of human rights or ISO/IEC's facilitation of global trade, into the emerging field of AI, thereby maintaining their relevance and influence amid technological transformation. Third, integration is promoted: by providing a unified conceptual system, ethical norms, and technical standards, they offer a common language and collaborative foundation for potentially fragmented global AI governance, significantly reducing the institutional costs and

technical barriers of international coordination, and laying a solid foundation for building an inclusive, trustworthy, and sustainable global AI governance ecosystem.

3. Promoting Capacity Building and Project Pilots: The Logic of Inclusiveness and the Development Mission

The fundamental motivation driving international organizations to participate in AI governance through capacity building and pilot projects is to prevent the further widening of the global digital divide. If the technological dividends and development opportunities of AI were to be enjoyed solely by a few developed countries and their technology enterprises, this would not only exacerbate technological asymmetries among states but also profoundly worsen global economic inequality and social injustice. Faced with this urgent challenge, international organizations whose core mission is to promote global development, including UNIDO, ITU, and UNESCO, must take proactive action. Their commitment to ensuring that developing countries and transition economies are "not left behind" in the AI era fulfills the responsibilities stipulated in their organizational charters and aligns with the widespread international expectation for inclusive development and the sharing of technological dividends. From a deeper strategic perspective, such action not only helps maintain global stability and reduce regional risks arising from developmental imbalances, but also creates more sustainable long-term markets and collaborative ecosystems for technology-exporting countries and enterprises, thereby achieving mutual benefit for all parties.

This pathway demonstrates a high degree of alignment with the functional positioning and resource capabilities of the relevant international organizations. UNIDO, whose core mission is to promote inclusive and sustainable industrial development, exemplifies this alignment. Through launching the Global Alliance on Artificial Intelligence for Industry and Manufacturing and implementing industrial intelligence demonstration projects in countries including Ethiopia, UNIDO effectively integrates the project management experience and local networks accumulated through its traditional industrial assistance with emerging AI technologies, achieving a seamless transition from concept to practice[1]. Similarly, ITU has for decades been committed to helping developing countries build and improve their communications infrastructure. Its training, technical assistance, and case dissemination activities under the AI for Good Global Summit framework

[1] UNIDO and Huawei, "UNIDO and Huawei Launch the Global Alliance on Artificial Intelligence for Industry and Manufacturing (AIM Global)," Huawei, July 6, 2023, https://www.huawei.com/en/news/2023/7/aim-global-announcement.

represent a natural extension of this traditional function into the AI era, not only continuing the core work direction of capacity building but also fully leveraging its established international cooperation channels and technical expert networks to achieve governance objectives at relatively low cost and high efficiency[1].

The success of international organizations in promoting capacity building and pilot projects and the broad impact they generate hinge on their unique resource integration capabilities and implementation models, particularly the central role played by the powerful credibility endorsement they provide. As highly neutral and credible third parties, these organizations are able to transcend national and commercial interests, coordinate multi-actor action from a fair and transparent position, and effectively reduce transaction costs and risk concerns for all cooperating parties. Drawing on their long-accumulated international reputation and cross-sectoral coordination capacity, they are able not only to aggregate technical, financial, and human resources from governments, enterprises, academia, and non-profit organizations, but also to help resolve policy barriers, institutional differences, and cultural obstacles that may arise during project implementation. More importantly, international organization participation confers international recognition and sustainability support on cooperative projects. For example, by leveraging their global networks to disseminate successful experiences, integrating local pilots into international technical assistance systems, or promoting the large-scale replication of mature models across different regions, they ensure that abstract development concepts and macro strategies are translated into concrete and tangible economic and social benefits. This mode of organizational action thereby effectively helps developing countries improve their AI application capabilities and ultimately wins sustained trust and broad cooperation from all participating parties.

The positive effects of this governance pathway are manifested at multiple levels. First, it effectively delivers on international organizations' commitments to "promoting development" and "technological inclusiveness," enhancing their organizational legitimacy and credibility through tangible outcomes, and demonstrating to the international community that they are not merely advocacy platforms but also actors capable of producing real change. Second, by cultivating technical talent in developing countries and demonstrating viable application scenarios, international organizations are not only cultivating future markets and application ecosystems for the global AI industry but also significantly expanding the participant base and supporter

[1] International Telecommunication Union, *AI for Good Global Summit 2024 Snapshot Report*, Geneva: ITU, 2024, pp. 16-22.

networks of their own initiatives, encouraging more stakeholders, including governments, enterprises, and local communities, to join the governance frameworks they promote. Finally, the firsthand data and practical experience accumulated through pilot projects provide valuable empirical evidence for these organizations to subsequently refine policy recommendations and adjust standard-setting. This shifts their work from theoretical design toward problem-driven approaches and forms a virtuous cycle of practice, feedback, and optimization, which continuously enhances the applicability and effectiveness of their governance measures.

The role of international organizations in AI governance is by no means accidental. Their success can be attributed to a highly rational strategic adaptation: selecting the most appropriate governance approaches based on their own "identities," including their mandated functions, resource endowments, and comparative advantages, thereby effectively responding to global governance demands. Whether through building platforms to build consensus, formulating standards to establish rules, or promoting inclusiveness through project pilots, these strategic pathways have not only efficiently advanced the global AI governance process but have also generated profound value for international organizations themselves. By successfully addressing frontier issues in the digital age, they have reshaped their practical relevance, consolidated their authority in the global system, and significantly expanded their influence, ultimately forming a positive cycle in which effective governance and self-reinforcement mutually promote each other. It follows that the influence of international organizations fundamentally derives from their ability to accurately translate their core organizational advantages into public goods that address global problems, and this constitutes the fundamental logic by which they grasp their roles, deploy their strategies, and ultimately construct effective implementation pathways in the AI era.

Conclusion

With the rapid development of AI technology, the demand for its governance has grown increasingly urgent worldwide. International organizations, as the central hubs of the global AI governance system, are playing an increasingly important role. From promoting rule-making and establishing global standards to building multilateral dialogue platforms, and from resource coordination to security supervision, international organizations are actively addressing the challenges posed by global AI through multiple governance strategies.

This article has systematically analyzed the multiple roles of

international organizations in promoting AI development, with particular emphasis on their core functions in rule-setting, platform-building, resource coordination, and supervision and facilitation. Various international organizations, including the United Nations, OECD, IEC, and ISO, have formed a closely coordinated collaborative network worldwide, providing important support for AI governance. Through issuing ethical and legal frameworks, promoting the development of global technical standards, facilitating transnational cooperation, and supporting capacity building in developing countries, these organizations are actively advancing the healthy, orderly, and sustainable development of AI globally.

That said, the role of international organizations in AI governance faces inherent limitations. Structural issues including insufficient enforcement capacity, diverging interests among member states, and the concentration of governance resources in developed countries remain real obstacles constraining their effectiveness. Future research could further focus on how to strengthen the binding force of international organizations' governance outputs, how to enhance the substantive participation of developing countries in rule-making, and how to promote higher-level institutional integration within a fragmented multilateral system. Only by confronting these challenges squarely can international organizations truly transform from "advocacy platforms" into "effective actors" and play their due historical role in global governance in the AI era.

In conclusion, international organizations play an indispensable role in global AI governance. They are not only a core force driving the harmonization of global rules but also promoters of technical standard-setting and international cooperation. Going forward, international organizations can continue to leverage their distinctive multi-level coordination advantages to improve the global AI governance framework, address the new challenges brought by this technological transformation, and ensure the inclusiveness, transparency, and security of AI technologies.

The "Brussels Mirage" under the EU's Digital Sovereignty Strategy: A Case Study of the Technological Disempowerment Dilemma in the Governance of the EU AI Act

*Yu Nanping Shu Zhongsheng**

Abstract Driven by its digital sovereignty strategy, the European Union has taken the lead in regulatory legislation across three dimensions—data, technology, and markets—while extending its reach through extraterritorial jurisdiction rules. The EU Artificial Intelligence Act, the world's first systematic AI regulation, leverages the scale of the EU single market to replicate the regulatory pathway of the General Data Protection Regulation (GDPR). It seeks to generate a "Brussels Effect" that establishes the "EU model" as the global paradigm for AI governance. In practice, however, the EU trails significantly behind China and the United States in AI computing capacity, industrial investment, and engineering capability. Consequently, within the governance framework of the AI Act, technical standard-setting, regulatory capacity, and corporate willingness to comply manifest pronounced fragmentation and heterogeneity. When technological strength cannot sustain the global diffusion of rules, even a textually rigorous regime may run idle, and the intended Brussels Effect risks degenerating into a "Brussels Mirage" stemming from technological disempowerment. Against the current backdrop of accelerating technological competition and diverging regulatory regimes in global digital governance, China should therefore examine the spillover effects of the EU's AI legislative model with sober judgment. Only when technological supply, industrial diffusion, and national regulatory capacity reinforce one another can regulatory frameworks avoid devolving into a form of technological disempowerment marked by "norms without

* Yu Nanping, Professor at the School of Politics and International Relations, East China Normal University; Shu Zhongsheng, Master of International Affairs, School of Politics and International Relations, East China Normal University.

technology, order without foundation," thereby safeguarding national strategic security.

Keywords European Union; Artificial Intelligence Act; Brussels Effect; Technological Disempowerment; National Strategic Security

I. Problem Statement

In recent years, the European Union has systematically advanced digital legislation as a means of exporting international public goods, with "digital sovereignty" serving as its strategic compass. From the *General Data Protection Regulation* (GDPR) of 2016, through the *Digital Services Act* (DSA) and the *Digital Markets Act* (DMA) of 2022, to the Artificial Intelligence Act (hereinafter "the Act"), which formally entered into force in 2024, the EU has sought to reinforce regional governance through a "digital fortress" mechanism, aiming to achieve autonomous control over critical nodes in data, technology, and rules, while leveraging market access as a lever to diffuse its regulatory norms outward[1]. In the AI domain, the EU has adopted a distinct strategic sequence: first, establishing stringent AI governance rules through a comprehensive regulatory framework; second, leveraging the attractiveness of its unified internal digital market to compel foreign enterprises to comply with these rules; and finally, generating a "Brussels Effect" through regulatory spillover[2]. The concept of the "Brussels Effect" was first articulated by Anu Bradford, a professor of law at Columbia University, in 2012, and was subsequently systematized in her 2020 monograph[3]. It refers specifically to the EU's capacity, by virtue of its vast market size, robust regulatory capacity, stringent regulatory standards, inelastic regulatory targets, indivisible regulatory rules, and first-mover legislative advantages, to pressure multinational corporations through market access conditions to adjust their global operations in accordance with EU standards, thereby completing the closed loop of "rule-making, market-driven compliance, and demonstration effects[4]."

---

[1] Xiong Guangqing and Chen Fei, "Digital Sovereignty, Digital Fortress and the Brussels Effect—How EU Digital Governance Policy Shapes the Global Digital Governance Model," *Teaching and Research*, No. 12, 2025, p. 91.

[2] David Bach and Abraham L. Newman, "The European Regulatory State and Global Public Policy: Micro-Institutions, Macro-Influence," *Journal of European Public Policy*, Vol. 14, No. 6, 2007, pp. 827-846.

[3] Anu Bradford, "The Brussels Effect," *Northwestern University Law Review*, Vol. 107, No. 1, 2012, pp. 1-69.

[4] Anu Bradford, *The Brussels Effect: How the European Union Rules the World*, Oxford: Oxford University Press, 2020, pp. 25-65.

However, the EU Artificial Intelligence Act formally entered into force on 1 August 2024, with its regulatory provisions rolled out in sequential phases. Prohibitive risk clauses took effect first in February 2025, followed by binding rules governing general-purpose AI models in August of the same year. The European Commission then linked compliance requirements for high-risk AI systems to harmonized technical standards via the *Digital Omnibus* initiative, and the legislative text entered a streamlined amendment round in March 2026. Despite this layered, multi-stage regulatory rollout, the well-documented "Brussels Effect" has failed to generate expected positive cross-border regulatory spillover. Instead, it has exhibited a turn toward technological disempowerment under the "Brussels Mirage[1]." The underlying causes of this reversal are threefold: at the industrial investment level, the EU lacks dedicated industrial policies and targeted funding support for the AI sector; at the standard-setting level, the EU depends on third-party standardization organizations, resulting in a transfer of standard-setting authority; and at the enforcement level, the binding force of rules has been progressively diluted under the dual pressures of great-power technological competition and corporate lobbying. More specifically, at the industrial foundation level, according to statistics from Stanford HAI, approximately 73 percent of globally leading AI large models in 2025 were released by U.S. enterprises, 15 percent by Chinese enterprises, and the share of EU enterprises remains far below its economic weight. European cloud companies hold less than 5 percent of the market share, and only 6 percent of global AI start-up financing flows to EU-based companies[2]. At the standard-setting level, the formulation of key European standards has been delegated to European Standardisation Organisations (CEN/CENELEC) and soft-law mechanisms such as "codes of practice". At the corporate lobbying level, multinational and local enterprises have simultaneously exerted pressure on the Act's implementation. In July 2025, Meta publicly refused to sign the *GPAI Code of Practice*, characterizing it as "overregulation[3]." Concurrently, European local enterprises including Airbus, ASML, Mistral AI, and Siemens Energy have issued joint appeals for a moratorium on key risk standards under the Act to reduce compliance burdens[4].

Against this backdrop, the EU's aspiration to project "normative power"

---

[1] Matti Ylönen, "Reconceptualising the Brussels Effect," *JCMS*: Journal of Common Market Studies, Vol. 64, No. 1, 2026, p. 109.

[2] R. Doan, A. Levy and V. Storchan, "Financing Infrastructure for a Competitive European AI," Feb. 10, 2025; N. Maslej et al., *Artificial Intelligence Index Report 2025*, Stanford HAI Institute, 2025.

[3] Joel Kaplan (Meta Chief Global Affairs Officer), Statement on EU GPAI Code of Practice, July 18, 2025; Jess Weatherbed, "Meta Snubs the EU's Voluntary AI Guidelines," *The Verge*, July 21, 2025.

[4] "Stop the Clock—Open Letter," AI Champions, July 3, 2025, https://aichampions.eu/.

globally through the Artificial Intelligence Act appears unlikely to materialize, nor will it generate a broad "Brussels Effect" beyond EU borders. More critically, the traditional narrative of the "Brussels Effect," long invoked by EU political elites, amplified through media discourse, and reproduced through academic research, has generated a pervasive "halo effect." Yet whether this "effect" can still be effectively embedded in EU strategic competition in AI, whether the Act can directly address the more fundamental challenge of balancing technological regulation, rule-making, and innovation, and whether the EU's traditional unified regulatory efficacy remains viable amid global AI strategic competition, all warrant critical examination and discussion.

II. The Regulatory Framework and Practical Dilemmas of the EU Artificial Intelligence Act

The inherent inexplicability and algorithmic black-box phenomenon of AI technology, coupled with the rapid iteration of AI products, continuously undermine the stability of existing risk-based regulatory frameworks. Against this backdrop of pursuing "digital sovereignty" and accelerating AI technological advancement, the EU proactively pushed for the adoption and phased implementation of the Artificial Intelligence Act, aiming to sustain its established pathway of shaping markets through rules and extending rules through markets. However, given the fluidity of technology, the divisibility of markets, and the heterogeneity of member states' enforcement capacities, whether the "single market advantage" underpinning the Act remains viable warrants critical reexamination. The following analysis deconstructs the institutional logic and practical dilemmas of the EU regulatory framework through three dimensions: risk-based classification, regulatory sandboxes, and phased implementation.

1. Risk-Based Regulatory System and Implementation Delay Pressures

The risk-tiered classification standard of the EU Artificial Intelligence Act is primarily grounded in the potential degree of harm to citizens' health, safety, and fundamental rights. It divides AI systems into four categories, namely unacceptable risk, high risk, limited risk, and minimal risk, with differentiated compliance obligations assigned accordingly. For AI practices posing unacceptable risks, such as manipulative subliminal techniques or biometric law enforcement based on sensitive attributes, the Act imposes a direct prohibition. For high-risk AI systems, the Act establishes a comprehensive set of obligations spanning the entire system lifecycle, including risk management, data governance, transparency, robustness, and

conformity assessment. For limited-risk systems, regulation is primarily achieved through transparency requirements, while minimal-risk systems are in principle subject to no additional regulatory restrictions. Furthermore, the Act constructs a parallel regulatory framework specifically for general-purpose AI models.

Table 1 Risk-Based Classification System and Specific Categories under the EU Artificial Intelligence Act

Risk Category	Definition and Typical Examples	Core Obligations
Unacceptable Risk	AI systems that pose a clear threat to human safety, livelihoods, and fundamental rights. Examples: social scoring systems, real-time remote biometric identification in public spaces, subliminal techniques manipulating human behavior, etc.	Comprehensive prohibition.
High Risk	Application scenarios that may cause significant negative impacts on life, health, fundamental rights, and security. Examples: safety components of critical infrastructure, education and vocational training, employment recruitment and screening, assessment of basic public services and welfare, judicial administration and democratic processes.	Strict compliance obligations, including but not limited to: robust risk management systems, use of high-quality and unbiased training datasets, comprehensive technical documentation and operational logging, assurance of system accuracy, robustness, and cybersecurity, implementation of appropriate human oversight, and registration with competent authorities.
Limited Risk	AI systems that may pose deception or confusion risks when interacting with users. Examples: chatbots, deep synthesis (deepfake) technologies, emotion recognition systems.	Transparency disclosure obligations: users must be explicitly informed that they are interacting with an AI system, and AI-generated or manipulated content must be disclosed.
Minimal or No Risk	The majority of AI systems with negligible or no impact on user rights. Examples: spam filters, AI in video games, AI-driven algorithmic recommendation systems in social media.	No additional obligations: freely circulable in the market; voluntary codes of conduct are encouraged.

Source: Compiled by the author based on European Parliamentary Research Service, "EU Legislation in Progress: Artificial Intelligence Act".

It should be recognized that the implementation of this risk-based regulatory mechanism faces multiple pressures. First, risk classification itself is not an automatic trigger of static rules but an ongoing activity requiring the interweaving of legal judgment and technical assessment[1]. Given the high variability in the purposes and technical solutions of AI systems, the same technical solution may be embedded in different application domains, usage scenarios, and user groups, thereby being categorized under different risk classes. For systems with multi-purpose characteristics or those that

[1] Carolina Cancela-Outeda, "The EU's AI Act: A Framework for Collaborative Governance," *Insights of Things*, Vol. 27, 2024.

continuously evolve through ongoing learning, they may even traverse different risk categories over their lifecycle[1]. Particularly when confronting the risk governance logic of regulatory "AI agents," the risks arising from their autonomous actions are by no means equivalent to those of the underlying large models. Forcibly conflating the two under a single legal framework will inevitably create confusion in future tort liability attribution and contract breach determination. The crux of the problem lies not in how agents "apply" old rules, but in the fact that their unique technical characteristics have already undermined the existing legal presumption space, likely leading to risk mismatch and regulatory failure within the EU. Second, the Act employs evaluative language such as "appropriate," "adequate," and "effective" extensively in formulating compliance obligations. While such phrasing avoids locking in specific technical paths and preserves flexibility in implementation, it also shifts the task of defining compliance risks further toward the implementation phase, given that supporting documents such as technical standards have not yet been fully developed[2]. More critically, this monolithic compliance-oriented, risk-based regulatory system has resulted in significant implementation delays. According to the Act's original phased implementation timeline, high-risk provisions should have taken effect on August 2, 2026. However, due to notable delays in both the standard-setting work supporting high-risk requirements and the establishment of competent authorities in member states, the rules could not be smoothly implemented at the scheduled time. In November 2025, the European Commission was compelled, through the *Digital Omnibus Regulation* proposal, to postpone the effective date for standalone high-risk AI systems to December 2, 2027, and for high-risk AI systems embedded in products subject to existing EU product safety regulations, such as machinery, medical devices, and lifts, to August 2, 2028[3]. This strategic adjustment indirectly confirms the immense implementation pressures facing the Act in practice and reflects the high dependency of its complex compliance obligations on administrative resources and corporate carrying capacity.

2. Regulatory Sandbox Flexible Governance and Resource Constraints

[1] Martin Ebers, "Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act," *European Journal of Risk Regulation*, Vol. 16, No. 2, 2025, pp. 684-703.
[2] Zhang Qi, "The Implementation Pressure and Rule Spillover of the EU Artificial Intelligence Act," *Journal of University of Electronic Science and Technology of China (Social Sciences Edition)*, March 2026, pp. 2-3.
[3] European Commission, *Digital Omnibus Regulation Proposal*, December 19, 2025, https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal.

The regulatory sandbox mechanism represents a technical requirement driven by the rapid advancement of AI technology. As an innovative instrument under the Act, EU member states are required to ensure that their national competent authorities establish at least one AI regulatory sandbox, which should be operational within 24 months of the Act's entry into force. Logically speaking, the stated objectives of establishing regulatory sandboxes include enhancing legal certainty for compliance purposes, supporting best practice sharing through regulatory cooperation, and promoting innovation and competitiveness by accelerating the market entry of AI systems, particularly those provided by small and medium-sized enterprises, including start-ups. The critical question, therefore, is whether the regulatory sandbox mechanism can genuinely encourage original innovation, given that this remains an issue of the sequence between regulation and implementation. Innovators are required to conduct experiments within a prescribed "box", and the design of this box, particularly its tolerance margin, presents a complex technical challenge, quite apart from the resource bottlenecks it poses for SMEs.

First, the effective operation of flexible governance tools such as sandboxes depends heavily on the administrative input of regulatory authorities and cross-border implementation coordination. From sandbox design, project selection, and risk assessment to real-time supervision during operation, regulatory authorities must continuously invest human resources and matching technical resources. Member states have exhibited notable divergences in the pace of sandbox establishment. Denmark established a regulatory sandbox in 2024 with a relatively clear implementation plan following the Act's introduction; Finland and Hungary have also initiated related planning, though no corresponding feedback has been yet available; while some member states like Greece have yet to announce specific plans[1]. Second, regulatory agencies of EU member states are themselves preoccupied with enforcing existing regulations and lack both AI expertise and the capacity to engage effectively with enterprises. Moreover, the sandboxes provide no substantive compliance exemptions, offering only free legal consultation to businesses, which makes them largely unattractive to SMEs and thus incapable of fulfilling their intended innovation-support function.

3. Multi-Level Phased Implementation and Functional Coordination Pressures

The implementation of the EU Artificial Intelligence Act relies on a multi-level, multi-actor governance system. At the EU level, the Act involves the AI Office, the AI Board, the Scientific Panel, and the Advisory Forum. At the member state level, countries are required to designate notified bodies and

[1] The Future of Life Institute, "AI Regulatory Sandbox Approaches: EU Member State Overview," May 12, 2025.

conformity assessment bodies, market surveillance authorities, and, where fundamental rights are implicated, to involve data protection authorities, equality bodies, and other fundamental rights protection entities in the assessment process[1]. Implementation thus depends entirely on the enforcement capacities of member states and multiple subordinate regulatory bodies. Particularly given the poor international interoperability of AI technology itself, its extraterritorial application will not only exacerbate global regulatory divergences but may also provoke disputes over data security[2]. Thus, the EU's multi-level governance structure, while originally an internal necessity of European integration and a mechanism for its maintenance, follows the same pattern in the multi-level governance architecture established under the AI Act. Although the Act incorporates consideration of AI governance's technical characteristics (see Figure 1)[3], the multi-level governance hierarchy nonetheless introduces inherent complexities to regulatory uniformity.

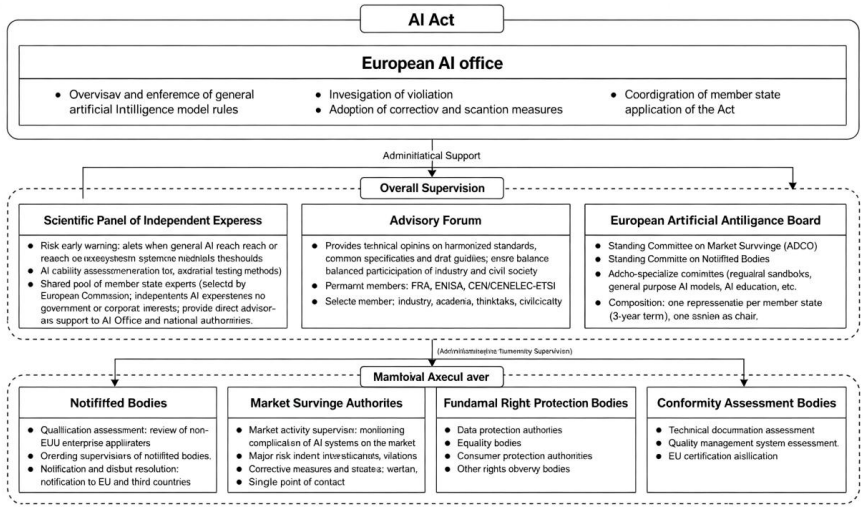


Figure 1 Multi-Level Governance Functional Structure of the Artificial Intelligence Act

The complex organizational architecture and the institutional environment of multiple parallel laws have already manifested significant operational difficulties in the early implementation phase. On the one hand, the Act does not clearly distinguish between the responsibilities of the AI Office and those

[1] Claudio Novelli et al., "A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities," *European Journal of Risk Regulation*, Vol. 23, 2024.

[2] Yu Nanping and Luan Xinwei, "On AI Regulation: The Technological Power of Artificial Intelligence from a State-Market Relations Perspective," *World Economics and Politics*, Vol. 6, No. 6, 2025, p. 39.

[3] The author has compiled and produced this figure based on the management architecture involved in the EU Artificial Intelligence Act.

of the European Commission. Some provisions assign regulatory responsibilities directly to the Commission, while others designate the AI Office as the implementing body. This ambiguity and variation in the articulation of responsible entities and their mandates create blurred lines of accountability in the early stages of institutional operation[1]. On the other hand, the autonomy and regulatory capacity differences at the member state level render rule implementation vulnerable to fragmentation risks. In terms of notified bodies, most member states have designated single entities, while a few, such as Germany and Luxembourg, have established multiple notified bodies. Regarding market surveillance authorities, countries including Czechia, Italy, and Poland have adopted centralized models, whereas Spain, Luxembourg, Finland, and others rely on multiple regulatory authorities for coordinated implementation [2] . These divergent institutional choices, compounded by disparities in administrative resources and regulatory experience, have led to widening gaps among EU member states in institutional development, staffing, and regulatory pacing.

More critically, the EU Artificial Intelligence Act does not operate in isolation but is nested within the EU's extensive legal system. As a general-purpose foundational technology, AI systems frequently trigger the simultaneous application of the GDPR, the DSA, the DMA, and multiple existing product safety regulations in their deployment scenarios. This parallel legal framework, while perhaps sustainable in the traditional era of lower technical demands, now requires regulatory authorities, under the complex conditions of AI, to continuously delineate the scope of application, division of responsibilities, and procedural requirements among different laws with technical implications, while simultaneously subjecting market actors to the compounding pressures of understanding and meeting multiple compliance obligations. For SMEs and start-ups with limited resources and expertise, these layered and highly complex regulatory requirements can easily become barriers to entry and obstacles to innovation[3].

III. From the "Brussels Effect" to the "Brussels Mirage": Constructing a New Analytical Framework

The implementation dilemmas of the EU Artificial Intelligence Act

[1] H. Graux et al., "Interplay between the AI Act and the EU Digital Legislative Framework," Publications Office of the European Union, 2025.
[2] S. Pinto, "Who Oversees AI in the EU? The Designation of National Competent Authorities under the EU AI Act," *HAL Archives*, October 2025.
[3] V. L. Raposo, "Reflections on the AI Act: Brussels, Do We Have a Problem?" in *The European Artificial Intelligence Act: Promises and Perils*, Springer, 2025.

discussed above are not merely a technical misalignment between institutional design and enforcement capacity; they also conceal a deeper transformation in the governance paradigm of the EU's digital sovereignty strategy in the age of AI. The five criteria underlying the traditional "Brussels Effect" theory, including single market size, regulatory capacity, preference for strict rules, inelastic regulatory targets, and indivisibility of standards, face severe challenges in the AI era. To comprehend this profound shift, it is necessary to systematically revisit the theoretical foundations and historical evolution of the Brussels Effect, examine whether the EU Artificial Intelligence Act satisfies the necessary conditions for its emergence, and on this basis construct a new analytical framework for the "Brussels Mirage."

1. Theoretical Foundations and Formative Conditions of the "Brussels Effect"

The "Brussels Effect," as a theory explaining the EU's projection of normative power through unilateral regulation, was first articulated by Professor Anu Bradford of Columbia Law School in 2012 and systematically elaborated in her 2020 monograph *The Brussels Effect*: How the European Union Rules the World[1]. Bradford conceptualized this phenomenon as the EU's capacity, by virtue of its institutional endowments and single-market advantages, to unilaterally externalize its internal regulatory standards globally in the absence of formal international agreements or coercive diplomatic instruments, thereby compelling or inducing enterprises, consumers, and even legislators in other jurisdictions to adopt "European standards", effectively making the EU a "global rule-maker[2]."

Bradford systematically distilled the generative mechanism of the "Brussels Effect" into five interconnected and indispensable core elements, which she characterized as the "necessary and sufficient" conditions for the effect to operate[3]. First, a single and vast internal market size. The EU, as one of the world's most significant consumer markets, derives its advantage not only from population scale but also from a high-quality unified market comprising approximately 500 million high-income consumers[4]. This immense

[1] Anu Bradford, "The Brussels Effect," *Northwestern University Law Review*, Vol. 107, No. 1, 2012, pp. 1-68; Anu Bradford, The Brussels Effect: How the European Union Rules the World, Oxford: Oxford University Press, 2020.
[2] Wang Xinchen and Liu Yun, "Conceptual Clarification and Misunderstanding Rectification of the Brussels Effect—Also on China's Response to the Globalization of EU Data Regulatory Standards," *Journal of Graduate School of China University of Political Science and Law*, Vol. 1, 2025, pp. 3-21.
[3] Anu Bradford, *The Brussels Effect: How the European Union Rules the World*, 2020, pp. 25-65.
[4] World Bank, "GDP (current US$)—European Union," World Bank Group, April 20,

market attractiveness forces multinational enterprises to bear corresponding compliance costs to gain access to the EU market, thereby laying the economic foundation for the global diffusion of regulatory standards. Second, robust regulatory capacity. Market size alone does not automatically translate into regulatory influence. As Bradford observed: "Becoming a regulatory power is a conscious choice of a state, not an inherent result of its market size[1]." Third, a preference for strict rules. The EU's "stringent orientation" toward emerging technology regulation stems both from its internal governance logic and from Europeans' cultural preferences regarding government regulation. Compared with market-driven spontaneous order and ex post private litigation remedies, European citizens generally place greater trust in government capacities for risk management[2]. Fourth, a predisposition to regulate inelastic targets. The sustainability of regulatory standards depends largely on whether regulated entities can evade regulation through simple cross-border relocation. The EU's capacity to exert unilateral influence derives precisely from the fact that regulatory targets in its single market are primarily consumer-oriented markets such as product safety, food safety, and data protection. Such consumer products with relatively low elasticity typically do not relocate to other jurisdictions in response to stricter regulation due to factors including market size[3]. Fifth, the indivisibility of standards. This constitutes the most critical condition for determining whether regulatory rules can generate spillover effects. When multinational corporations face indivisibility across different markets, or when the scale effects of adopting uniform regulatory standards exceed the cost savings of maintaining differentiated standards in more lenient markets, enterprises will proactively extend EU rules as global standards. Bradford subdivided indivisibility into three types: legal indivisibility, such as the integrity of merger review in antitrust enforcement; technical indivisibility, such as the inability to physically separate EU data from other data in cross-border data flows; and economic indivisibility, where the costs of maintaining differentiated products exceed the costs of uniform compliance[4].

Bradford's other significant theoretical contribution lies in her precise distinction between the *de facto* Brussels Effect and the *de jure* Brussels Effect. The "de facto Brussels Effect" refers to the process by which multinational enterprises, to save compliance costs, voluntarily apply EU rules to their global

---

2026, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

[1] Anu Bradford, "The Brussels Effect," p. 12.

[2] Wang Xinchen and Liu Yun, "Conceptual Clarification and Misunderstanding Rectification of the Brussels Effect—Also on China's Response to the Globalization of EU Data Regulatory Standards," *Journal of Graduate School of China University of Political Science and Law*, Vol. 1, 2025, pp. 6-7.

[3] Anu Bradford, "The Brussels Effect," p. 17.

[4] Anu Bradford, "The Brussels Effect," p. 19.

operations, thereby making EU standards the global industry benchmark at the factual level. Under this mechanism, the diffusion of EU norms does not require legislative changes in other countries; multinational corporations "voluntarily align" based on commercial interests to achieve global extension of regulatory effects. The "de jure Brussels Effect", by contrast, refers to the process by which other jurisdictions, inspired by the EU regulatory model or pressured by EU negotiations, proactively incorporate European-style rules into their domestic legislation. The two are interrelated yet distinct: the former emphasizes substantive penetration of rules at the commercial practice level, while the latter focuses on legal transplantation at the institutional level[1]. To some extent, the "de jure Brussels Effect" can also be viewed as developing countries "benchmarking advanced practices" in the context of economic globalization.

2. Conceptual Definition and Analytical Dimensions of the "Brussels Mirage"

The "Brussels Effect" anticipated by the EU Artificial Intelligence Act has long transcended the realm of pure academic analysis, having undergone a transformation from "explanatory tool" to "policy instrument" in the discursive practices of EU political elites. When the political agreement on the Artificial Intelligence Act was reached in December 2023, Ursula von der Leyen declared on social media: "The AI Act is a global first, as it delivers a unique, trustworthy legal framework for AI development[2]." The President of the European Parliament hailed the Act as "groundbreaking" legislation[3]. European Parliament rapporteurs emphasized it as "the world's first comprehensive AI legislation" and "the world's first robust precedent for AI regulation[4]." On this basis, the "Brussels Effect" is gradually undergoing a turn toward "self-fulfilling prophecy" as the "Brussels Mirage." That is, even before the legislative text of the Artificial Intelligence Act has entered full

[1] Feng Yujun and Wei Hongguang, "The 'Brussels Effect' Theory of the GDPR and Its Critique—An Analysis of Extraterritorial Influence of Legislation," *Journal of Yantai University (Philosophy and Social Sciences Edition)*, No. 6, 2023, pp. 38-40.
[2] Ursula von der Leyen's statement on X (formerly Twitter), 8 December 2023, quoted in "EU Agrees 'Historic' Deal on Artificial Intelligence Regulation," Anadolu Agency, April 24, 2026, https://www.aa.com.tr/en/europe/eu-agrees-historic-deal-on-artificial-intelligence-regulation/3078187.
[3] European Parliament, "Artificial Intelligence Act: MEPs Adopt Landmark Law," 13 March 2024; also see "World's First Major Act to Regulate AI Passed by European Lawmakers," CNBC, April 24, 2026, https://www.cnbc.com/2024/03/13/european-lawmakers-endorse-worlds-first-major-act-to-regulate-ai.html.
[4] European Parliament Press Release, "Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI," April 24, 2026, https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

implementation, the self-positioning of "global standard-setter" has already achieved discursive diffusion through official EU channels, personal statements by political elites, and the transatlantic policy commentary network, forming a normative consensus that precedes empirical verification.

Anna-Julia Saiger, in an article published in the *European Journal of International Law*, was the first to explicitly employ the concept of the "Brussels Mirage" to describe the limitations of the Artificial Intelligence Act in generating rule spillover and emulative legislation. Saiger argued that the AI Act functions more as a symbolic EU regulatory instrument, having failed to generate a genuine Brussels Effect globally, and that its rule spillover is more appropriately characterized as a "Brussels Mirage[1]." Saiger identified three levels of this mechanism: first, the Act's extensive reliance on technical standards developed by European Standardisation Organisations to guide specific implementation renders it susceptible to "reverse shaping" by countries that have taken the technological lead in AI; second, countries are increasingly asserting their own regulatory stances in AI governance rather than simply emulating the EU; third, the United States, the United Kingdom, Australia, Japan, Singapore, India, and others have adopted innovation-supportive and less restrictive approaches to AI regulation. For instance, Japan's *AI Promotion Act* follows a facilitative legislative path, while Singapore and India have primarily issued non-binding AI policy documents of a recommendatory and guiding nature[2]. Evidently, the legislative trajectory of AI regulation has shifted from initial emulation of the EU's risk-based approach toward more universal and principle-based regulatory paths, implying that the "Brussels-style" strict regulation embodied in the Act is being diluted in other jurisdictions. Almada and Radu analogize this negative impact as the "side effects of the Brussels Effect": even if the EU AI Act may generate certain normative diffusion effects, such diffusion will inevitably be accompanied by the simultaneous spillover of regulatory deficiencies, potentially undermining the European values the Act purports to promote[3]. Lu Chuanying and Yang Liwei have similarly observed that the EU AI industry is exhibiting a trend of contraction, lacking emerging technology enterprises at the forefront of the AI industrial chain, and thus unable to grasp dominance in international AI rule-making[4].

---

[1] Anna-Julia Saiger, "Brussels Mirage: The EU AI Act's Subtle Shine across International Borders," *EJIL: Talk!*, June 26, 2025, https://www.ejiltalk.org/brussels-mirage-the-eu-ai-acts-subtle-shine-across-international-borders/.

[2] Zhi Zhenfeng, "Major Concerns and Future Trends in Global Artificial Intelligence Legislation," Comparative Law Research, No. 6, 2025, p. 30.

[3] Marco Almada and Anca Radu, "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy," German Law Journal, Vol. 25, No. 4, 2024, pp. 646-663.

[4] Lu Chuanying and Yang Liwei, "The 'Security Paradox' and Regulation in the Global

Building on the foregoing discussion, this article further specifies three core analytical dimensions of the "Brussels Mirage": the hollowing-out of technical standard-setting authority, the mismatch between EU regulatory capacity and corporate compliance willingness, and the failure of legal transplantation across different institutional contexts under the dual pressures of great-power competition and corporate lobbying. These three testable dimensions will serve to determine whether the technological disempowerment of the EU Artificial Intelligence Act is real.

IV. Three Dimensions of Technological Disempowerment in the EU Artificial Intelligence Act

Whether the EU Artificial Intelligence Act can replicate the "Brussels Effect" of the *General Data Protection Regulation* in the AI domain has generated relatively systematic scholarly discussion[1]. As an analytical concept, the theoretical penetration of the "Brussels Mirage" lies in its capacity to reveal the "technological disempowerment" dilemma facing the EU in AI governance. Specifically, the EU Artificial Intelligence Act exhibits disempowerment across three dimensions: technical standard-setting, regulatory enforcement, and rule diffusion. This technological disempowerment is not merely a matter of inadequate policy implementation; it is a practical manifestation of the constrained spillover of the EU's traditional "normative power" under the compounded forces of the EU's relative technological backwardness, monopolization of the AI technology stack by multinational digital platforms, and intensifying geopolitical competition and corporate lobbying.

1. The Hollowing-Out of Technical Standard-Setting Authority: Transfer of EU AI Harmonized Standards to Private Actors

In its legislative technique, the EU Artificial Intelligence Act follows the tradition of the "New Legislative Framework" (NLF) for EU product safety regulation, which previously provided practical regulatory tools for EU member states in product rule-making, helping them build a unified and efficient market surveillance system by translating legally prescribed "essential requirements" into operational technical specifications through harmonized standards[2]. Under this framework, however, the EU Artificial Intelligence Act

---

Diffusion of Artificial Intelligence," Exploration and Contention, No. 12, 2025, p. 115.

[1] Su Kezhen and Shen Wei, "Will the EU Artificial Intelligence Governance Scheme Produce a 'Brussels Effect'?—An Analysis Based on the EU Artificial Intelligence Act," Germany Studies, No. 2, 2024, pp. 66-88.

[2] Marco Almada and Anca Radu, "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy," *German Law Journal*, Vol. 25, No. 646, 2024, pp. 646-663.

has delegated the core rule-making authority originally belonging to the legislature to private-actor-led European Standardisation Organisations (ESOs) in the name of "technological neutrality." Under the coordination of the ESOs, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) jointly established the CEN-CENELEC Joint Technical Committee 21 (JTC 21) on June 1, 2021, to accelerate the development of multiple standards supporting the implementation of the AI Act, covering quality management, risk management, engineering implementation, social ethics, and cybersecurity. Although these standards are formulated by non-governmental organizations and explicitly characterized as "voluntary" under EU standardization law, they possess de facto binding force in practice[1]. The committee currently comprises five working groups, bringing together over 300 experts from more than twenty countries, with the Danish Standards organization serving as the secretariat[2]. Yet precisely this ostensibly neutral, professional, and efficient technical standard-setting mechanism constitutes the root cause of the hollowing-out of standard-setting under the "Brussels Mirage."

First, there is the crisis of "expert representativeness" among standard-setting actors. While JTC 21 formally requires each national standardization body to include multiple stakeholders, including government, industry, and civil society, in its committee composition, the participation mechanism, which requires membership fees as a precondition, effectively excludes civil society organizations, human rights institutions, and consumer groups from the core deliberative process. Even before the draft AI Act was introduced, European Standardisation Organisations were identified as having regulatory expertise concentrated in product safety, electrotechnical engineering, and mechanical engineering, with a manifest lack of interpretive capacity and experience regarding fundamental rights, non-discrimination, and data protection for AI products in the AI era[3]. When the AI Act requires the technical committee to formulate specific technical specifications on matters involving core provisions of the Charter of Fundamental Rights of the European Union, such as "risk coefficients," "data governance," "transparency obligations," and "bias detection," technical experts are compelled to assume functions of balancing

[1] Xu Jingting, "On the Jurisprudential Boundaries and Interest Balance Paths of Standard Copyright Protection—With Reference to the Standard Practices of China, the United States, and Europe," *Journal of Tianjin Administrative College*, No. 2, 2026, pp. 82-95.
[2] CEN-CENELEC, "About the Joint Technical Committee," April 25, 2026, https://jtc21.eu/about/; see also European Commission, "Standardisation of the AI Act," https://digital-strategy.ec.europa.eu/en/policies/ai-act-standardisation.
[3] Michael Veale and Frederik Zuiderveen Borgesius, "Demystifying the Draft EU Artificial Intelligence Act—Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach," *Computer Law Review International*, Vol. 22:4, 2021, pp. 97-112.

fundamental rights that properly belong to constitutional courts or legislatures. As scholars have explicitly noted, a technical committee can formulate norms on "how to minimize the risk of worker injury caused by mechanical equipment," but can hardly define the limits of privacy infringement in workplace injury prevention scenarios[1].

Second, there is the absence of "transparency" and "accountability" in the standard-setting process. JTC 21 has long refrained from disclosing its membership list and expert composition, lacks an information disclosure mechanism open to the public regarding standard-setting progress, and makes relevant draft texts subject to copyright protection and available only for a fee, rendering effective public oversight difficult. Although this issue was partially remedied in the 2023 case of ASTM v. Public.Resource.Org, Inc., which clarified that harmonized standards, though drafted by private bodies, nonetheless "constitute part of EU law" and accordingly presumed that the European Commission bears the obligation to make these standards freely available to the public like other "laws[2]," the judgment compelled JTC 21 to respond to the transparency crisis in 2025 by establishing an "Inclusiveness Task Group" and opening social media channels as "informal avenues." Yet such "ex post remedial" mechanisms can neither remedy the procedural deficiencies in standard-setting nor fundamentally construct a standard-setting mechanism conforming to the EU's tripartite legitimacy framework of "input legitimacy, output legitimacy, and procedural legitimacy."

Finally, there is the "systematic decoupling" and "strategic deferral" of the standard-setting process. The European Commission formally submitted a standardization request to JTC 21 on May 22, 2023, requiring the completion of drafting harmonized standards supporting the AI Act by April 30, 2025. However, in August 2024, Sebastian Hallensleben, Chair of the JTC 21 Working Group on Risk and Accountability, publicly confirmed an eight-month delay. To ensure the predictability of rule implementation, the European Commission, through the *Digital Omnibus Regulation* published on November 19, 2025, directly postponed the compliance deadlines for high-risk AI systems to December 2027 for standalone systems and August 2028 for systems embedded in products, explicitly linking this deferral to the "availability" of harmonized standards. This means that absent the timely completion of standards, the core obligation provisions of the AI Act will remain in a state of

[1] Quoted from Filip's interview in the Inclusive AI Governance study, see Ada Lovelace Institute, "Inclusive AI Governance," April 25, 2026, https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/.
[2] Case C-588/21 P, Public.Resource.*Org and Right to Know v Commission* (CJEU 5 March 2024); prior case Case C-613/14, *James Elliott Construction* (CJEU 27 October 2016).

"suspension." A "framework law" lacking complete technical specifications not only suffers substantially diminished normative force in outward rule transmission but also harbors implicit strategic security concerns for the EU.

2. "Implementation Divergence" in Regulatory Capacity: Structural Misalignment between Institutional Supporting Mechanisms and Corporate Compliance Willingness

First, there is the mismatch between regulatory resources and the technical complexity of AI. The EU Artificial Intelligence Act attempts, drawing on the established tradition of product safety regulation, to incorporate AI systems into the existing framework of conformity assessment and notified bodies. However, AI systems differ fundamentally from traditional mechanical or medical devices: factors such as training data scale, model weight transparency, and inference process explainability lack uniformly quantifiable measurement benchmarks. Reports from some AI enterprise research departments have frankly stated that AI, unlike electricity or food, lacks "universal measurement thresholds," and the technical committee, relying on the collaboration of over 200,000 volunteer experts, is institutionally ill-equipped to keep pace with the rapid iteration of frontier AI technologies[1]. Technical complexity thus poses an unavoidable new challenge to regulatory capacity and resources.

Second, there is the "passive resistance" and "selective compliance" of corporate willingness to comply. The provisions concerning general-purpose AI models under Chapter V of the EU AI Act took effect on August 2, 2025. To assist enterprises in meeting compliance requirements, the European Commission issued the voluntary *GPAI Code of Practice* on July 10, 2025, as a guiding instrument. However, U.S. technology giants have adopted markedly opposing stances regarding whether to sign the GPAI Code of Practice. Joel Kaplan, Meta's Chief Global Affairs Officer, stated that such rules would stifle technological innovation, warning that "Europe is on the wrong track in AI, and the Code creates a series of uncertainties for model developers, with compliance measures extending far beyond the scope of the AI Act itself[2]." Elon Musk's xAI has adopted a "partial signing" strategy, signing only the "Safety and Security" chapter while explicitly refusing to be bound by the

[1] Mejuvante analysis report: "Why CEN/CENELEC Fails on the AI Act: When Standards Come After Regulation," September 23, 2025, https://mejuvante.ai/blog/news-2/why-cen-cenelec-fails-on-the-ai-act.

[2] Joel Kaplan, LinkedIn Post, July 18, 2025; see also Jess Weatherbed, "Meta Snubs the EU's Voluntary AI Guidelines," *The Verge,* July 21, 2025, https://www.theverge.com/news/710576/meta-eu-ai-act-code-of-practice-agreement.

"Transparency and Copyright" chapter[1]. On July 3, 2025, the "EU AI Champions Initiative," composed of European local technology enterprises including Airbus, ASML, Philips, Mistral AI, and Siemens Energy, jointly signed an open letter calling for a two-year deferral of the key obligation provisions of the AI Act. In the face of such large-scale "passive resistance" from both foreign and domestic enterprises, the EU has had no choice but to respond with strategic deferral of the Act's effective dates. A study published by the Carnegie Endowment for International Peace in December 2025 indicates that this "deregulatory turn" on the EU's part precisely undermines the foundation of rule credibility embedded in the "inelastic regulatory target" criterion of the Brussels Effect: when a governance actor is compelled to make practical trade-offs between "maintaining regulatory stringency" and "sustaining industrial competitiveness", its capacity for global rule diffusion inevitably weakens[2]. Thus, the interplay between corporate power and the traditional EU regulatory political forces has given rise to a new imbalance of power in the emerging field of AI, which in turn affects the strategic security dimension of the EU's pursuit of "strategic autonomy."

3. Disorientation amid Multipolar Geopolitics: The Failure of Norm Diffusion under Great-Power Technological Competition and Corporate Lobbying Pressures

It should be recognized that the EU Artificial Intelligence Act has encountered significant external obstruction from great-power technological competition in rule diffusion, making it difficult for the EU to translate its normative outputs into international consensus. Internally, coordinated lobbying by multinational corporations has diluted the demonstrative force of the EU's rules themselves. These dual pressures have caused the EU's AI governance model to drift away from its institutional anchor in the international standard-setting marketplace.

First, in the diffusion of the "Brussels Effect," U.S. enterprises have long treated compliance with European rules as a default, and U.S. legislation has never made substantial regulatory efforts. However, AI technological development has fundamentally reversed U.S. regulation from "rule-follower" to "rule-leader." On January 23, 2025, the Trump administration signed Executive Order 14179, *Removing Barriers to American Leadership* in Artificial Intelligence, which revoked the Biden administration's Executive

[1] TTMS, "EU AI Act Update 2025: Code of Practice, Enforcement, Industry Reactions," October 30, 2025, https://ttms.com/eu-ai-act-update-2025/.
[2] Carnegie Endowment for International Peace, "The EU's AI Power Play: Between Deregulation and Innovation," December 4, 2025, https://carnegieendowment.org/research/2025/05/.

Order 14110 on safe, secure, and trustworthy AI development and explicitly directed the U.S. federal government to identify, revise, or eliminate all regulatory rules impeding AI development and deployment[1]. In February 2025, U.S. Vice President J. D. Vance explicitly warned at the Paris AI Action Summit that overregulation "could kill AI[2]." On July 23, 2025, the United States released *America's AI Action Plan: Winning the AI Race*, establishing "accelerating AI innovation," "building AI infrastructure," and "leading international AI diplomacy and security" as its three pillars. The plan, established pursuant to Executive Order 14320, *Promoting the Export of the American AI Technology Stack*, explicitly targets the export of a "U.S. AI technology stack" encompassing hardware, models, software, applications, and standards to allies and partner countries as a strategic objective[3]. When the EU and the United States, as two major global jurisdictions, diverge fundamentally in AI governance models and underlying philosophies, "voluntary alignment with the EU" ceases to be the default compliance option for multinational technology enterprises. U.S. multinationals, with the connivance and acquiescence of their government, pursue "accelerated openness," disregarding European "rigidity and conservatism."

Second, the lobbying of multinational technology corporations has undermined the stability of EU rule-making. U.S. technology companies, for example, are strategically deploying innovation resources, converting their control over technical resources, investment decisions, and product deployment into visible and perceptible displays of power to influence EU AI policy-making processes. They signal their control over critical resources for European AI development to policymakers by threatening market withdrawal, delaying product deployment, or expanding investment commitments[4]. In the post-implementation phase of the AI Act, this competition has been further externalized as lobbying pressure from multinational technology corporations. According to the non-profit research and advocacy organization Corporate Europe Observatory, the final text of the *Digital Omnibus Regulation* adopted by the EU in November 2025 "highly reflects" the lobbying positions of large multinational technology companies[5]. On May 22, 2025, representatives of

---

[1] Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," January 23, 2025.
[2] "Vance Tells Europeans That Heavy Regulation Could Kill AI," Reuters, February 12, 2025, https://www.reuters.com/technology/artificial-intelligence/europe-looks-embrace-ai-paris-summits-2nd-day-2025-02-11/.
[3] The White House, "America's AI Action Plan: Winning the AI Race," July 23, 2025, https://www.whitehouse.gov/articles/2025/07/.
[4] Zhou Yijiang, "Beyond Lobbying: The Strategies and Logic of U.S. Technology Companies' Influence on EU Artificial Intelligence Legislation," *Germany Studies*, No. 6, 2025, p. 80.
[5] Corporate Europe Observatory, Study on Big Tech Lobbying in EU Digital Omnibus,

U.S. technology giants including IBM, Microsoft, Amazon, Meta, Google, and OpenAI met with European Commission officials to collectively pressure for streamlined compliance requirements [1]. Facing dual pressure from multinational capital and domestic industry, the European Commission announced the postponement of high-risk AI system compliance deadlines to December 2027 and the curtailment of the AI Office's regulatory authority. Thus, this demonstrates how the comprehensive expansion of multinational technology corporations across the AI value chain and industrial chain enables them to wield overwhelming technological advantages in lobbying and shape EU AI policy-making. This dynamic forces the bloc to continuously adjust its regulatory rules to ease corporate compliance burdens, since it cannot simultaneously pursue regulatory autonomy and innovative dynamism under its technological sovereignty agenda. This "weaponized strategic rivalry" by multinational technology corporations not only exposes the EU's critical weaknesses in AI technological capacity but also necessitates a new logic for analyzing the relationships between the state and the market, and between the state and strategic security, in the AI era.

V. Conclusion: Beyond the "Brussels Mirage"

From its official entry into force on August 1, 2024, the EU Artificial Intelligence Act had, by March 26, 2026, advanced to its inaugural round of legislative amendment. The European Parliament adopted its "first reading position" on the *Digital Omnibus Regulation*, signaling that the Act would demonstrate greater flexibility in implementation pace and procedural burdens, exercise restraint regarding fundamental rights protection thresholds, while retaining the impulse for expansion in response to emerging harmful risks. It is evident that the European Parliament and the Council of the EU have gradually accepted the deferral of the Act and the adoption of simplified measures as a form of self-constraint, endeavoring to confine such "regulatory relaxation" to procedural and implementation levels without readily compromising substantive protection standards. This trend toward "procedural burden reduction" represents an inevitable manifestation of the fundamental tension between the EU's dual identity as a "regulatory power" and a "technology follower". The "Brussels Mirage", as a more critical and pragmatic perspective on the relationship between institutional design and implementation conditions, reveals that the emergence of this phenomenon under the EU AI Act is not

---

quoted in Jacques Delors Centre, "The EU's Digital and AI Omnibus is Heading in the Wrong Direction," March 2026, https://www.delorscentre.eu/en/publications/.

[1] "Meetings of Cabinet Members of Executive Vice-President Henna Virkkunen with Interest Representatives," May 22, 2025, https://ec.europa.eu/transparencyinitiative/meetings/

attributable to any single factor, but rather constitutes a combined effect of the EU's technological dependency, institutional entrapment, and political retreat. For China, a thorough understanding of the experiences and dilemmas in the EU AI Act's legislative process carries significant strategic implications.

First, China should guard against treating the EU model as a ready-made template for wholesale legal transplantation. The EU's efforts to diffuse rules through unified legislation are encountering notable setbacks in the highly uncertain and rapidly evolving AI domain. The delays in implementation, lags in standard-setting, and contraction of industrial competitiveness suggest that unified horizontal legislation is neither the sole nor the inevitable pathway for AI governance[1]. On the contrary, when advancing its artificial intelligence legislation, China should fully recognize a core institutional advantage of its multi-layered, incremental governance framework. This system can effectively cope with technological uncertainties and rapid industrial iteration, so the country does not need to hastily push forward the enactment of comprehensive AI legislation. The framework is built upon foundational statutes including the Cybersecurity Law, the Data Security Law and the Personal Information Protection Law, and further supported by departmental regulatory documents such as the Interim Measures for the Management of Generative Artificial Intelligence Services, the Provisions on the Administration of Algorithmic Recommendation in Internet Information Services, and the Provisions on the Administration of Deep Synthesis in Internet Information Services.

Second, China should approach the spillover effects of the "Brussels Effect" with greater sobriety, avoiding path dependence on the EU regulatory model. From the *GDPR* to the AI Act, the diffusion effects of EU digital governance are shifting toward "selective adoption", meaning that most countries are selectively drawing on EU experience at the level of governance instruments rather than adopting its institutional framework wholesale. As a technologically leading country with a complete AI industrial chain, a vast domestic market, and rich application scenarios, China should ground its participation in global AI governance more firmly in its own governance wisdom, rather than passively adapting to the EU's regulatory framework. Moreover, from the perspective of national strategic security, China should reject the model of "technological hegemony" and instead commit to promoting open, fair, and effective governance mechanisms through Chinese initiatives such as the *Global AI Governance Initiative*, the *AI Capacity Building for All Initiative*, and the *Global AI Governance Action Plan*, within multilateral frameworks including the World Data Organization or the United

[1] Fu Xinhua, "Comprehensive AI Legislation Should Be Proceeded with Caution," *Oriental Law*, No. 3, 2025, p. 20.

Nations.

Third, the "Brussels Mirage" offers an important lesson: greater attention must be paid to technological capabilities and technical standards, with a "technology ecosystem" approach to achieving competitive advantage, and accelerated implementation and diffusion throughout the industrial chain. Scholars have recognized that in advancing AI-related legislation, China should emphasize the coordinated evolution of regulatory legislation and technical standards, fully leveraging the bridging role of national and industry standards within the regulatory system to ensure that legislative intent is effectively realized at the technical level[1]. Furthermore, targeted and precise institutional responses should be developed for specific issues already emerging in typical application scenarios such as generative AI, algorithmic recommendation, and deepfakes, rather than pursuing a grand "AI code[2]." It should be recognized that under current conditions, particularly amid accelerating AI technological iteration and before the boundaries of AI technology have been clearly defined and solidified, a guiding principle oriented toward national strategic security should be established, grounded in practical and industrial application-based approaches. China's AI regulatory model should be dynamically and progressively constructed with Chinese characteristics, rather than imposing development constraints in response to disciplinary pursuits of topical issues. It must be recognized that the core issues underlying AI technological development are great-power strategic interests, national strategic security, and future global technological dominance. In this strategic security competition, the Chinese approach to AI governance is to optimize governance through development rather than predetermining governance in advance.

Finally, the key to preventing technological disempowerment lies in building a multi-actor collaborative governance framework grounded in an autonomous technology foundation and centered on technological subjects. The EU's experience demonstrates that administrative regulation or legislative coercion alone cannot adequately address the complexity and uncertainty of AI technology. China should continue to refine its multi-actor governance mechanisms involving enterprises, government, the judiciary, and the public. At the corporate compliance level, market entities should be guided through laws, regulations, and standard systems to fulfill obligations including data

[1] Zhang Linghan, "What Kind of Artificial Intelligence Law Does China Need?—The Basic Logic and Institutional Architecture of China's AI Legislation," *Legal Science (Journal of Northwest University of Political Science and Law)*, No. 3, 2024, p. 10.
[2] Xue Lan, Jia Kai, and Zhao Jing, "Agile Governance Practice in Artificial Intelligence: Categorised Regulatory Approaches and the Construction of a Policy Toolbox," *Chinese Public Administration*, No. 3, 2024, p. 103.

security protection, algorithmic security management, content labeling for generated content, and ethical review. At the administrative regulatory level, government departments should exercise overall coordination and supervision, actively exploring flexible regulatory instruments adapted to technological characteristics. At the judicial level, rights boundaries and liability rules should be clarified through the adjudication of representative cases. At the societal level, broad participation from the public, academic institutions, industry organizations, and the media should be actively encouraged. In this manner, China's AI governance should transcend the “Brussels Mirage” and forge a governance path that not only accords with the developmental logic of AI technology but also aligns with China’s national conditions, safeguards national strategic security, and ultimately benefits the common well-being of all humanity. This path should be built on the basis of coordinating development and security, problem-oriented and collaborative governance, and the integration of domestic rule of law with foreign-related rule of law. This is not merely a theoretical response to transcending the “Brussels Effect,” but also a test of great-power governance wisdom for our era.

Normative Agency in Latecomer Domains: The Global South's Strategies in Artificial Intelligence Governance

*Cai Cuihong Guan Hang**

Abstract Norms of artificial intelligence governance are diffusing at an accelerating pace worldwide. The Global South has long been regarded as a passive norm-taker and a latecomer follower in emerging domains such as artificial intelligence. However, this assumption overlooks the strategic agency that Global South countries have demonstrated in norm selection, adaptation, and co-construction. Centering on the question of how actors exercise strategic agency in the process of norm-building within their latecomer domains, this article identifies four types of agentic strategies: developmentalization of issue framing, collectivization of agency, localization of normative content, and issue-linkage through multilateral action. Through an analysis of AI governance practices by African and Southeast Asian states and regional organizations, as well as India, the article validates the explanatory effectiveness of these four strategies in accounting for the Global South's participation in AI governance norm-building. The Global South is not a passive norm-taker in AI governance; its agency, while limited, is meaningful—conditioned by development needs, institutional capacity, and the constraints of international structures.

Keywords Global South; Artificial Intelligence Governance; Norm Diffusion; Agency

* Cai Cuihong, Professor at the Institute of International Studies, Fudan University, and Deputy Director of the Center for Global Artificial Intelligence Innovation Governance; Guan Hang, Ph.D. Candidate at the School of International Relations and Public Affairs, Fudan University.

I. Introduction

The rapid pace of AI technological iteration, the depth of its diffusion, and the extent of its impact have swiftly placed it at the center of the global governance agenda. A series of influential AI governance norms have emerged worldwide, including but not limited to the Organisation for Economic Co-operation and Development (OECD) *AI Principles*[1], the United Nations Educational, Scientific and Cultural Organization (UNESCO) *Recommendation on the Ethics of Artificial Intelligence*[2], the European Union (EU) Artificial Intelligence Act, and the United Nations (UN) global resolution on AI (A/RES/78/265)[3]. The intensive issuance of these normative texts signals that the diffusion of global AI governance norms has entered a new phase of accelerated competition and coordination.

However, both social and policy debates surrounding these norms remain markedly biased toward great-power perspectives. Whether examining normative practices such as U.S. executive orders, EU legislative pathways, or Chinese diplomatic initiatives, or reviewing mainstream research and public discourse, the focus has almost exclusively been on normative competition among technologically leading powers, while the circumstances and specific concerns of Global South countries receive only superficial attention. Global South countries, having long occupied the lower end of the industrial chain and followed rather than led innovation, find themselves in a latecomer position across many emerging domains. Moreover, objective conditions such as inadequate critical infrastructure and insufficient technological reserves exacerbate their dependence on external technologies, making it increasingly difficult to escape the structural predicament of "lagging from the start[4]." As Acharya has criticized, norm diffusion research has long suffered from a structural deficiency in neglecting the agency of non-Western actors[5]. This critique has yet to receive systematic response in the field of AI governance research.

This bias is not accidental; it reflects a more fundamental justice issue in

---

[1] OECD, "OECD AI Principles overview," https://oecd.ai/en/ai-principles.
[2] UNESCO, "Ethics of Artificial Intelligence," https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.
[3] United Nations, *Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development*, March 21, 2024, https://docs.un.org/zh/A/RES/78/265.
[4] Chen Zhimin, Wu Zhicheng, Sun Jisheng, et al., "Global Governance Initiatives and Major-Country Diplomacy with Chinese Characteristics," *Foreign Affairs Review (Journal of the Foreign Affairs College)*, No. 1, 2026, pp. 1-41.
[5] Amitav Acharya, "How Ideas Spread: Whose Norms Matter? Norm Localization and Institutional Change in Asian Regionalism," *International Organization*, Vol. 58, No. 2, 2004, p. 240.

the construction of global AI governance norms. Scholars who have sought to extend the theory of justice from the domestic to the global level argue that national borders do not constitute the external limits of social cooperation, and that global-level institutional arrangements are equally subject to the scrutiny of principles of justice[1]. Such scrutiny requires moral evaluation of the actual effects of different institutional designs, with particular attention to the consequences of rules being coercively applied to vulnerable groups[2]. In other words, sound international norms and orders inherently require the examination and realization of justice. Yet the norms of justice cannot merely stop at universal claims at the textual level, but rather, it must be actualized through the genuine and accessible participation and development rights for vulnerable actors. Moreover, such a framework must more fully incorporate the diversity of global civilizations and practices, as well as the dialectical relationships among them[3]. The modern international order is deeply entangled with the historical legacy of colonialism, and its embedded hierarchical structures are continuously reproduced under the guise of development discourse, locking certain states into the role of "passive recipients of order."[4] This colonial legacy has not dissolved in emerging domains such as AI governance; the logic of imposing standards and norms in the name of "civilizational standards" or "global best practices" continues to operate. In light of this, this article argues that, from the inherent requirements of global justice, the core criterion for evaluating the effectiveness of global AI governance norms should be whether Global South countries can genuinely gain development opportunities from this technological transformation and whether they can autonomously participate in the formulation and adaptation of rules.

With this in mind, this article seeks to examine the Global South, or more broadly, how actors exercise agency in their latecomer domains, from a more pragmatic and constructive perspective. Latecomer domains refer to the relational situation in which actors simultaneously face relative lag in technological capabilities, institutional experience, and norm-setting positions in specific emerging governance issues. More plainly, they are domains where actors are doubly pressured by structural disadvantage compounded by

[1] Charles Beitz, *Political Theory and International Relations*, Princeton: Princeton University Press, 1979, pp. 127-176.
[2] Thomas Pogge, *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*, Cambridge: Polity Press, 2002, pp. 169-195.
[3] Yin Zhiguang, "The Theoretical Poverty of Global Order Studies and the Practice of Modernization Diversity: A Global South Perspective," *Social Sciences*, No. 8, 2024, pp. 99-113.
[4] Owen R. Brown, "The Underside of Order: Race in the Constitution of International Order," *International Organization*, Vol. 78, No. 1, 2024, pp. 38-66.

technological weakness. Lateness is domain-specific and relative; for instance, the same country may possess strong institutional capacity on certain issues while remaining in a vulnerable position in AI foundation models, computing power, chip supply, or international standard-setting. Furthermore, this article does not treat the Global South as a homogeneous bloc, but rather understands it as a complex collection of actors that share certain developmental aspirations and unequal conditions within the global technological governance structure, yet are highly differentiated in internal capacities and interests.

On this conceptual basis, this article addresses the following question: How do actors exercise agency in the process of governance norm construction within their latecomer domains? After reviewing existing theories of norm diffusion and contestation, this article constructs a theoretical framework for how actors exercise agency in norm-building within latecomer domains, using it to analyze the normative agency practices of Global South countries in AI governance. The aim is to lay the groundwork for future explorations of what kind of global governance system would enable Global South countries to more fully exercise their agency in shaping AI governance norms, thereby allowing the Global South to autonomously safeguard its own interests in this monumental transformation of productive forces driven by intelligence.

II. Literature Review

This article centers on the exercise of agency by Global South countries in the construction of international norms. Since the end of the Cold War, international norm research has evolved from a focus on the unidirectional process of norm diffusion toward recognition of diverse actors' agency. However, the systematic mechanisms of Global South engagement in norm creation, adaptation, and resistance remain a frontier topic in academic discourse.

One strand of scholarship examines the linear process through which international norms spread from advocates to recipients. This theoretical tradition comprises two complementary approaches. The hegemonic socialization approach argues that great powers promote the acceptance of norms by secondary states through a dual mechanism of "external inducement" and "internal reconstruction," the core logic being the synergistic effect of material incentives and cognitive persuasion[1].

The concept of hegemonic socialization reveals two core mechanisms of great-power norm dissemination: "external inducement," which effects

[1] G. John Ikenberry and Charles A. Kupchan, "Socialization and Hegemonic Power," *International Organization*, Vol. 44, No. 3, 1990, pp. 283-315.

behavioral change in secondary states through material incentives and institutional design, and "internal reconstruction," the shaping of elite belief systems in secondary states through cognitive socialization and coercive power[1]. Furthermore, the deep structure of power asymmetry in international normative practice entails great-power norm violation in the name of upholding them, alongside rule-based constraint of weaker states' action space[2]. These theoretical foundations provide important reference points for understanding the power dimensions of AI governance norm diffusion. The U.S. approach of promoting norm adoption through export controls and technology alliances reflects the mechanism of external inducement, while the extraterritorial effects of the EU Artificial Intelligence Act exhibit characteristics of institutionalized socialization. However, the hegemonic socialization approach fundamentally reduces norm recipients to passive objects of socialization. Martha Finnemore and Kathryn Sikkink astutely observe that many widely diffused norms, including women's suffrage, originated with non-hegemonic advocates. In AI governance, practices such as the African Union's "Africa-centric, development-focused" framework, ASEAN's regional governance buffer layer, and India's AI summits and "MANAV" vision resist explanation by the unidirectional logic of hegemonic socialization.

Running parallel to this is the norm life cycle model, which partitions norm evolution into three stages: emergence, cascade, and internalization. It emphasizes norm entrepreneurs' use of organizational platforms and socialization effects to drive norms toward general acceptance[3]. The key conditions for norm diffusion success or failure include norm "salience," "clarity," and "proximity[4]." The model's core contributions are the identification of norm entrepreneurs' organizational-platform persuasion mechanisms and the "tipping point" concept, which offers strong explanatory power for shifts in norm diffusion pace. In the emergence stage, actors with strong convictions employ framing strategies to redefine issues and attract adoption by critical states. Once the number of adopting states surpasses the tipping point, socialization pressures and legitimacy concerns drive more states to follow suit, progressively facilitating norm cascade. Ultimately, norms become internalized as taken-for-granted standards of appropriate behavior. However, this model has two limitations in explaining Global South agency.

---

[1] G. John Ikenberry and Charles A. Kupchan, "Socialization and Hegemonic Power," *International Organization*, Vol. 44, No. 3, 1990, pp. 283-315.
[2] Stephen D. Krasner, *Sovereignty: Organized Hypocrisy*, Princeton: Princeton University Press, 1999.
[3] Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change," *International Organization*, Vol. 52, No. 4, 1998, pp. 887-917.
[4] Ibid.

First, the model's linear stage assumption presupposes a unidirectional norm evolution trajectory. The pathway from entrepreneurs to recipients is treated as the norm, while recipient adaptation, reconstruction, or even back-export of norms during diffusion remains insufficiently explored. Second, the model positions Global South countries primarily as followers in the "norm cascade" stage. Their adoption motives reduce to international legitimacy pursuit or bandwagoning effects, while selective adoption and reinterpretation of normative content based on domestic needs remain obscured. As Acharya has pointed out, local actors are not merely passive targets of external persuasion; they actively borrow and modify external norms to align with their own beliefs and practices[1].

These studies have contributed theoretically by revealing the basic mechanisms of norm transmission. However, both approaches presuppose Global South countries as passive recipients or objects awaiting socialization, paying insufficient attention to the agency of non-Western actors. Such research often treats norms as stable independent variables, avoiding examination of potential connotation shifts and meaning reconstruction during diffusion, while neglecting deep power asymmetries inherent in norm diffusion. In many cases, norm entrepreneurs themselves are the greatest selective enforcers of the norms they promote[2].

A second strand of scholarship focuses on local actors' active engagement with norms. Acharya identified three biases in the "moral universalism" perspective of social constructivism: a focus bias that overemphasizes "universal" norms at the expense of regional and local legitimacy; a normative hierarchy bias that places global "good" norms above local "bad" ones; and an actor bias that frames norm diffusion as a Western-led "civilizing" process, devaluing non-Western agency[3].

Building on this critique, Acharya proposed two core concepts: "norm localization" and "norm subsidiarity." Norm localization is the dynamic process whereby local actors borrow and adapt external ideas to local traditions and practices, rendering them congruent with existing local normative systems and cognitive priors. The key innovation of this concept is its emphasis on "congruence-building" agency. In norm transmission, local actors do not passively receive external norms but selectively borrow and creatively adapt

---

[1] Amitav Acharya, "How Ideas Spread: Whose Norms Matter? Norm Localization and Institutional Change in Asian Regionalism," *International Organization*, Vol. 58, No. 2, 2004, pp. 239-275.
[2] Stephen D. Krasner, *Sovereignty: Organized Hypocrisy*, Princeton: Princeton University Press, 1999.
[3] Amitav Acharya, *Whose Ideas Matter? Agency and Power in Asian Regionalism*, Ithaca: Cornell University Press, 2009.

them through framing, grafting, pruning, and cultural selection. Norm subsidiarity is the process whereby local actors create and establish norms to preserve their autonomy, thereby avoiding domination, neglect, infringement, or bullying by more powerful actors[1]. If norm localization concerns local actors' "reception" and "adaptation" of external norms, norm subsidiarity concerns their "creation" of norms to protect their own interests and autonomous space. Building on these two concepts, Acharya further proposed a "norm circulation" framework, wherein norm diffusion is not unidirectional from global to local but a bidirectional, multi-stage interactive process: transnational ideas are localized to fit local cognition and practices, while transformed local ideas feed back to modify and develop global norms. This framework fundamentally challenges the traditional understanding that the West creates norms for dissemination to the non-West, providing a systematic theoretical foundation for understanding Global South agency in norm construction.

While this theoretical turn is groundbreaking, its empirical analysis has concentrated mainly on Southeast Asian security norms. Systematic extension of this framework to emerging technology issues like AI governance faces three challenges. First, AI governance involves rapidly iterating technologies and highly specialized knowledge, necessitating adjustments to traditional security norm localization mechanisms. Second, the power of technology firms in AI governance far exceeds that in traditional security domains, yielding a more complex actor constellation. Third, AI governance involves tensions among multiple objectives, including development promotion, risk control, and ethical values, requiring Global South agency to navigate strategic choices within a more complex tension structure.

A third strand addresses the non-linear dynamics of norm evolution. Research on "norm regress" reveals how already-internalized norms decline due to competitive counter-narratives from challengers[2]. The concept of "norm antipreneurs" captures actors committed to maintaining the status quo and resisting normative change[3]. Norm contestation is increasingly understood as a constant feature of norm evolution rather than a sign of normative failure[4]. Chinese scholars have also contributed theoretically in this direction. The concept of "norm containment" describes actors in the ambiguous space

[1] Amitav Acharya, "Norm Subsidiarity and Regional Orders: Sovereignty, Regionalism, and Rule-making in the Third World," *International Studies Quarterly*, Vol. 55, No. 1, 2011, p. 97.
[2] Ryder McKeown, "Norm Regress: U.S. Revisionism and the Slow Death of the Torture Norm," *International Relations*, Vol. 23, No. 1, 2009, pp. 14-18.
[3] Alan Bloomfield, "Norm Antipreneurs and Theorising Resistance to Normative Change," *Review of International Studies*, Vol. 42, No. 2, 2016, pp. 310-333.
[4] Antje Wiener, *A Theory of Contestation*, Heidelberg: Springer, 2014.

between support and opposition, neither advancing nor rejecting new norms but adopting delaying and restrictive strategies to constrain norm development pace and scope[1]. Other research has examined the progressive mechanism from autonomy arguments to goal arguments, wherein emerging states, through successful practical experiences, promote substantive adjustments to normative content. For example, in international human rights norm construction, "peripheral great powers" have pursued positive identity to participate in norm-building, using norm adaptation and creation as strategies for social recognition, thereby enriching understanding of Global South normative agency motivations[2].

Despite the richness of existing research, several areas require further refinement, discussion, and clarification. First, existing theories have not systematically explored Global South normative agency in relatively weak or latecomer domains, particularly in emerging technology governance issues like AI and digital governance. The empirical foundations of existing agency research are concentrated in traditional areas like security, human rights, and the environment, whereas technology governance norms, with their rapid iteration, high specialization, and public-private actor interweaving, challenge existing framework applicability. Second, while existing studies have critiqued the unidirectional norm diffusion model, they have yet to provide operational typologies of observable strategies for weaker actors' modification of issue frameworks, organizational forms, and normative content. Third, how structural constraints on agency, including developmental needs, institutional capacity, and international structural constraints, jointly shape agency space has not been adequately theorized. This article therefore constructs a strategic agency framework that accommodates these empirical realities and theoretical challenges. On the one hand, it converts Global South agency from abstract "capacity" into comparable strategic types; on the other hand, it situates strategic choices within the constellation of constraints posed by developmental needs, institutional capacity, and international structure. Drawing on Global South agential practices in AI governance norm diffusion, this article explores how actors exercise agency in norm-building within disadvantaged latecomer domains, offering a new theoretical perspective on normative agency in issue areas where actors are relatively weak.

---

[1] Chen Zheng, "Norm Containment and Its Strategies: The Cases of China and Russia in the Evolution of the Responsibility to Protect," *World Economics and Politics*, No. 6, 2019, pp. 65-90.

[2] Chen Zheng, "Discursive Strategy Adaptation in Norm Contestation: The Case of the Responsibility to Protect," *World Economics and Politics*, No. 8, 2023, pp. 60-86.

III. Strategic Agency of Actors in Latecomer Domains

Drawing on the theoretical resources discussed above, this chapter constructs an analytical framework for explaining how actors exercise agency in the governance norm-building process within their latecomer domains. The framework takes Acharya's theories of norm localization and norm subsidiarity as its core pillars, while integrating Finnemore and Sikkink's norm life cycle model and Stephen Weymouth's analysis of power structures in digital governance.

The following analysis first defines the core concept of "strategic agency" and its three-tier connotations, then identifies four types of agentic strategies and specifies the framework's conditions of application, analytical boundaries, and its relation to the research question.

1. Defining Strategic Agency

This article defines "strategic agency" as the behavioral practice through which actors, in the governance norm-building and diffusion process within their relatively weak or latecomer domains, selectively adopt, creatively adapt, and proactively construct external norms in accordance with their own developmental needs and institutional conditions. The term "relatively weak" does not imply that actors lack all resources or can only follow passively; rather, it denotes their lack of a dominant position in rule-making, technological provision, or institutional output within specific issue areas. The term "strategic" emphasizes that actors do not respond randomly to external normative pressures but choose modes of action more conducive to preserving developmental autonomy and enhancing institutional legitimacy under varying constraints. This concept inherits Acharya's insights on the agency of weaker actors but shifts the analytical focus from traditional security norms to emerging technology governance domains, while highlighting the non-fixed and domain-specific nature of actors' weak positions and the strategic and conditional character of agentic behavior. Strategic agency comprises three levels: selective adoption, creative adaptation, and autonomous construction.

The first level is selective adoption. Global South countries do not accept all external governance norms in latecomer domains indiscriminately; rather, they adopt selectively according to their own stage of development, technological needs, and institutional environments. Agency at this level manifests as rational choice driven by consequentialist logic, yet "selection" itself transcends passive acceptance, involving active evaluation of normative content, comparison with domestic priority agendas, and strategic judgment regarding the timing of adoption. For the purposes of this article, African

countries' stance toward the UNESCO *Recommendation on the Ethics of Artificial Intelligence* exemplifies a typical pattern of selective adoption: they actively embrace provisions concerning capacity building, technology transfer, and development priorities while maintaining a cautious attitude toward clauses that might restrict data flows and technological applications[1]. ASEAN countries similarly draw selectively on OECD principles and the EU framework, endorsing abstract principles such as human-centeredness and transparency while refraining from adopting specific institutional designs like risk-based tiered regulation[2].

The second level is creative adaptation, which involves reinterpreting and reconstructing external norms to align them with local needs, values, and institutional contexts. In latecomer domain governance, this takes four specific forms. First, the framing mechanism, which redefines governance priorities, for instance, African countries place "development promotion" before "risk control" in AI governance. Second, the grafting mechanism, which transplants international principles onto local institutional traditions, as seen in ASEAN's embedding of consensus-based and non-binding principles into its AI governance framework. Third, the pruning mechanism, which eliminates normative elements inconsistent with local values, such as downplaying Western individualistic data rights discourse. Fourth, cultural selection, which incorporates local values into ethical frameworks, as exemplified by the Ubuntu philosophy's emphasis on collective rights dimensions in Africa[3]. The key feature of creative adaptation is that its outcomes are neither simple replicas of external norms nor outright rejections, but hybrid forms combining both global and local elements. This hybridity itself constitutes a marker of Global South agency, signifying local actors' active intervention and meaning-making in normative dynamics.

The third level is autonomous construction. Global South countries not only adapt external norms but also proactively create new governance norms and institutional arrangements, reflecting their active efforts to preserve autonomy and discursive power. This means they are not content with making choices within existing frameworks but seek to create new options. India's "MANAV" vision, launched at the 2026 AI Impact Summit, exemplifies this logic of autonomous construction and deliberate agency, aiming to propose an alternative framework distinct from those of the United States, the European

---

[1] Jake Okechukwu Effoduh, "Decolonizing the Governance of Artificial Intelligence in Africa: From Normative Mimicry to Epistemic Sovereignty," *Science and Public Policy*, Vol. 53, No. 2, 2026, pp. 245-257.
[2] Bama Andika Putra, "Governing AI in Southeast Asia: ASEAN's Way Forward," *Prometheus in Artificial Intelligence*, Vol. 7, 2024.
[3] Kinfe Yilma, "Ethics of AI in Africa: Interrogating the Role of Ubuntu and AI Governance Initiatives", *Ethics and Information Technology*, Vol. 27, No. 2, 2025.

Union, and China in global governance dialogue. While ASEAN's regional AI governance framework draws extensively on international elements, its choice of a "soft law" pathway itself constitutes an alternative institutional innovation to the hard law governance model.

The above three levels form a continuum from passive to active. The same actor may display different levels of agency across different issue areas or at different times. Advancement along this continuum typically entails a shift from the factual to the normative, that is, from purely instrumental choices toward the active construction of one's own normative identity.

2. Typology of Agentic Strategies

Under the strategic agency framework, the agentic strategies of the Global South in latecomer governance norm diffusion can be categorized into four ideal types, each corresponding to different constraints and behavioral mechanisms (see Table 1).

The first is the development-oriented of issue framing. This strategy presupposes that the Global South, for purposes of defending its behavior, seeking problem resolution, and integrating value judgments, redefines governance priorities and core frameworks at the discursive level[1]. In latecomer domains, this manifests as sustained pursuit of development-oriented framing and agenda-setting aligned with domestic developmental needs[2]. For instance, it emphasizes that AI governance should first serve development goals such as poverty reduction, education, and public health, rather than treating risk control as the primary or sole governance logic. In other words, actors change the direction of debate and the space for policy choice by redefining the nature and boundaries of the problem at hand[3]. The conditions for applying this strategy include: first, the actor has a clear development-priority agenda that is in tension with the priority ordering of external norms; second, the actor possesses sufficient discursive resources and organizational platforms to articulate alternative framings; and third, the international discursive environment provides space for the expression of developmental claims.

---

[1] Joy Wanjiku Njoroge, "Comparative Analysis of AI Governance in Africa Relative to Global Standards and Practices," *IOSR Journal of Computer Engineering*, Vol. 26, No. 5, 2024, pp. 19-25.
[2] Irene Poetranto, et al., "Look South: Challenges and Opportunities for the 'Rules of the Road' for Cyberspace in ASEAN and the AU," *Journal of Cyber Policy*, Vol. 6, No. 3, 2021, pp. 318-339.
[3] [Canda] Amitav Acharya, *Constructing Global Order: Agency and Change in World Politics*, translated by Yao Yuan and Ye Xiaojing, Shanghai People's Publishing House, 2021, pp. 39-50.

The second is the collectivization of agency. This strategy operates by pooling the strength of weaker actors through regional organizational platforms, transforming dispersed individual voices into a unified collective position, thereby gaining higher negotiating status and stronger norm-creation capacity in global governance dialogue. Although individual developing countries have limited influence, institutionalized coordination through regional organizations can aggregate strength and amplify voice, compensating to some extent for individual capacity deficiencies. The conditions for applying this strategy include: first, the regional organization possesses a certain level of institutionalization and collective action capacity; second, member states share sufficient consensus on core demands; and third, the regional organization enjoys a degree of discursive space and recognition in the international arena. Notably, regional collective agency is often constrained by internal divisions and collective action dilemmas, and the consensus reached typically exhibits a lowest-common-denominator character[1].

The third is the localization of normative content. This strategy operates by selectively drawing on certain elements of global governance frameworks and articulating them with local institutional traditions, cultural values, and developmental practices to form hybrid governance pathways with local characteristics. This strategy directly corresponds to the conceptual tools of framing, grafting, and pruning in Acharya's norm localization theory. Its essence lies in preserving the core features of local normative orders while selectively absorbing elements of external norms that enhance local institutional legitimacy and effectiveness. The conditions for applying this strategy include: first, the existence of local normative traditions or institutional arrangements with a degree of legitimacy and vitality; second, the presence of credible local actors not perceived as external agents to serve as drivers of the grafting process; and third, the existence of graftable space between external and local norms, meaning they are not entirely opposed but possess partial overlap or complementarity[2].

The fourth is issue-linkage through multilateral action. This strategy operates by connecting latecomer domain governance with issues of widespread concern to the Global South, such as the digital divide, technological equity, data sovereignty, and South-South cooperation, thereby gaining discursive power and agenda-setting authority through multilateral platforms. This strategy embeds specific governance issues in a broader

[1] Huang Yutao, "From Autonomy Arguments to Goal Arguments: How Emerging States Promote International Norm Change," *World Economics and Politics*, No. 4, 2023, pp. 62-95.
[2] Chen Dongxiao and Wang Tianchan, "The Global South Promotes a 'Path to Equal Rights' in Global Cyberspace Governance," *China Cyberspace*, No. 10, 2025, pp. 64-68.

discourse of developmental justice and global equity, enabling the Global South to leverage existing normative resources and moral advantages to support its positions in emerging governance domains. The conditions for applying this strategy include: first, the existence of linkable normative resources that have already gained a certain degree of international recognition; second, the actor has channels for voice and the possibility of agenda-setting in multilateral platforms; and third, there is a substantive connection between latecomer domain governance issues and existing Global South agendas rather than a tenuous analogy.

Table 1 Four Types of Agentic Strategies in Latecomer Domain Governance Norm Diffusion

Strategy Type	Core Mechanism	Conditions of Application
Developmentalization of issue framing	Redefining priority issue frameworks in domain governance	① Tension between domestic development agenda and priority ordering of external norms;
		② Availability of discursive resources and organizational platforms for articulating alternative framings;
		③ International discursive environment provides space for articulating claims
Collectivization of agency	Consolidating collective positions through regional organizational platforms	① Regional organization has a certain level of institutionalization and collective action capacity;
		② Member states share consensus on core demands;
		③ Regional organization has discursive space and is recognized
Localization of normative content	Forming hybrid norms with local characteristics	① Existence of legitimate and vital normative traditions or local institutions;
		② Credible local actors as drivers of grafting;
		③ Existence of graftable space between external and local norms
Issue-linkage through multilateral action	Connecting latecomer domains with existing issues, gaining discursive power and agenda-setting authority through multilateral platforms	① Availability of linkable, internationally recognized normative resources;
		② Channels for voice and possibility of agenda-setting in multilateral platforms;
		③ Substantive connection between domain governance issues and existing agendas

Source: Compiled by the author.

IV. Agentic Strategies and Practices of the Global South in AI Governance

Building on the theoretical framework developed above, this section provides empirical illustration and preliminary testing of the four types of agentic strategies in latecomer domain governance norm diffusion through four case clusters. This article does not claim that these cases exhaust all Global South practices in AI governance; rather, it treats them as typical scenarios for observing how different agentic strategies unfold. Case selection follows three principles: first, strategy correspondence, with each case cluster corresponding primarily to one type of agentic strategy; second, actor diversity, encompassing different levels of actors including African regional organizations, the ASEAN regional approach, the national level of India, and Global South multilateral platforms; and third, temporal relevance, with all cases drawn from the critical period of accelerated AI governance norm diffusion between 2024 and 2026.

1. Development-oriented of Issue Framing: Reframing from AI Risk Control to Development Promotion

The core mechanism of the development-oriented strategy lies in redefining the priority issue framework of domain governance. Its conditions of application include tension between the actor's development agenda and the priority ordering of external norms, the availability of discursive resources and organizational platforms to articulate alternative framings, and international discursive space for articulating developmental claims. The practices of Global South countries in AI governance provide the most typical empirical evidence for this strategy type.

In August 2024, the African Union released the *Continental Artificial Intelligence Strategy*. The defining feature of this strategy is its substantive reframing of AI governance priorities. In contrast to the EU Artificial Intelligence Act, which centers on "risk control" issues such as algorithmic bias, privacy protection, and job displacement, the African strategy explicitly articulates an "Africa-centric, development-focused approach to AI," closely aligning AI governance with the development vision of Agenda 2063 and emphasizing that AI technologies should serve development goals including poverty reduction, education, public health, and agricultural modernization. This framing is not merely rhetorical adjustment but substantive reconstruction of the normative core, positioning the right to development, rather than merely civil and political rights, at the center of the technology governance system[1].

[1] African Union, "Continental Artificial Intelligence Strategy," July 2024, https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

Most African countries are still in the early stages of AI technology application. Simple replication of Western risk control frameworks could produce two negative consequences: excessively stringent regulatory standards may stifle AI innovation, while the prioritization of risk control frameworks may crowd out limited governance resources. The tension between domestic development agendas and the priority ordering of external norms thus constitutes the fundamental driver of issue reframing. The African Union, as a regional organization with 55 member states, provides ample discursive resources and organizational platforms, enabling African countries to articulate alternative norms and visions with a collective voice.

Moreover, the development-oriented strategy is not limited to principled declarations at the regional organizational level but is deepening into national legislation. Nigeria, as a major African economy, is advancing the continent's most comprehensive AI legislation. Its *National Digital Economy* and *e-Governance Bill*, scheduled for passage in the first half of 2026, introduces a risk-tiered model similar to the EU's but with the fundamental objective of promoting digital economy development rather than pure risk control. The bill introduces regulatory sandboxes to encourage innovation, providing start-ups with space to test AI systems under supervised conditions. Kashifu Abdullahi, Director-General of the National Information Technology Development Agency, stated in a Bloomberg interview that Nigeria aims to adopt a "proactive rather than reactive" stance in AI regulation. This case exemplifies the autonomous construction dimension of strategic agency: Nigeria not only creatively adapts external norms but also proactively creates legislative arrangements with African characteristics[1], reflecting the growing importance of technological factors in modern state image and national identity construction[2].

At the national level, India similarly demonstrates the strategic logic of developmentalization. At the 2026 India AI Impact Summit, Prime Minister Modi introduced the "MANAV" vision, encompassing M for Moral and Ethical Systems, A for Accountable Governance, N for National Sovereignty, A for Accessible and Inclusive AI, and V for Valid and Legitimate Systems, placing "accessibility and inclusivity" on equal footing with "morality and

[1] Bloomberg, "Nigeria Set to Pass Sweeping AI Rules for Digital Economy," January 13, 2026, https://www.bloomberg.com/news/articles/2026-01-13/nigeria-set-to-pass-sweeping-ai-rules-for-digital-economy; TechAfrica News, "Nigeria Advances Landmark Digital Economy and e-Governance Bill at National Assembly Hearing," November 14, 2025, https://techafricanews.com/2025/11/14/nigeria-advances-landmark-digital-economy-and-e-governance-bill-at-national-assembly-hearing/.
[2] Su Shitian, "Technological Nationalism from a Global South Perspective: The Case of Laos' State Building and Development," *Southeast Asian Studies*, No. 5, 2025, pp. 114-136.

ethics." India's incremental strategy of "develop first, regulate later" positions AI as a tool for driving inclusive economic growth rather than a mere object of regulation, representing a significant practice through which actors change the direction of debate and the space for policy choice by redefining the nature and boundaries of the problem at hand[1].

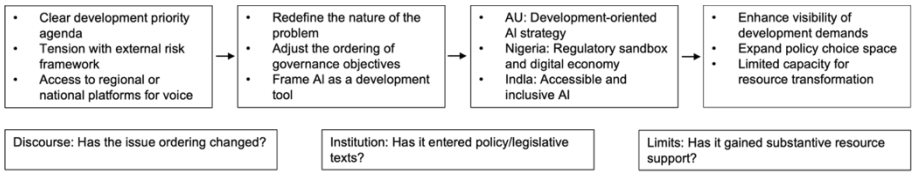


Figure 1 Analytical Diagram of the Developmentalization of Issue Framing Strategy

The above three cases collectively demonstrate that when Global South countries experience tension between their own development agendas and Western-dominated normative priority orderings, they neither simply reject nor passively accept external norms, but exercise agency by redefining governance priority issue frameworks. Development needs constitute both the impetus for accepting external norms and the fundamental basis for localized adaptation.

2. Collectivization of Agency: Regional Organizations as AI Governance Platforms

The core mechanism of the collectivization strategy lies in consolidating collective positions through regional organizational platforms, transforming dispersed individual voices into unified collective stances. Its conditions of application include a certain level of institutionalization and collective action capacity of the regional organization, consensus among member states on core demands, and discursive space and recognition for the regional organization in the international arena. The practices of the African Union and ASEAN in AI governance clearly demonstrate the operational mechanisms of this strategy.

In June 2024, over 130 African ministerial-level officials and experts participated in the second special session of the specialized technical committee on communications and ICT, unanimously adopting the *Continental Artificial Intelligence Strategy*. Individual African countries have extremely limited voice and influence in global AI governance. However, through the African Union, which has 55 member states representing over 1.4 billion people, they can form unified positions and establish collective bargaining status. This illustrates that regional organizations perform dual

[1] Government of India, "PM Inaugurates India AI Impact Summit 2026," https://www.pmindia.gov.in/en/news_updates/pm-inaugurates-india-ai-impact-summit-2026/.

functions in Global South agency, serving as both organizational carriers for collective agency and incubators for institutional capacity building. By 2025, 44 African countries had adopted data protection laws and 38 had established enforcement agencies, and this accelerated institution-building is inseparable from the African Union's collective coordination role[1]. It must be noted, however, that collective agency is simultaneously constrained by internal tensions. African countries exhibit vast disparities in digital economic development levels, political systems, and governance capacity. Kenya has 27 percent of its population using ChatGPT daily, while least developed countries such as South Sudan and the Central African Republic cannot even guarantee basic electricity access. This internal diversity results in a "lowest-common-denominator effect," whereby the final position adopted tends to reflect minimal consensus rather than the most ambitious goals.

ASEAN's collective agency displays a different institutional form. In February 2024, ASEAN released the *ASEAN Guide on AI Governance and Ethics*, and on January 17, 2025, the Fifth ASEAN Digital Ministers Meeting in Thailand issued the *Expanded ASEAN Guide on AI Governance and Ethics for Generative AI*. More significantly, ASEAN formally established the AI Governance Working Group in Phnom Penh in March 2024, chaired by Singapore's Infocomm Media Development Authority, responsible for coordinating AI policies among its 10 member states. The ASEAN Digital Economy Framework Agreement aims to double the regional digital economy to $2 trillion by 2030. This collective action gives ASEAN, with its 680 million people, a negotiating position in global governance dialogue that cannot be ignored[2].

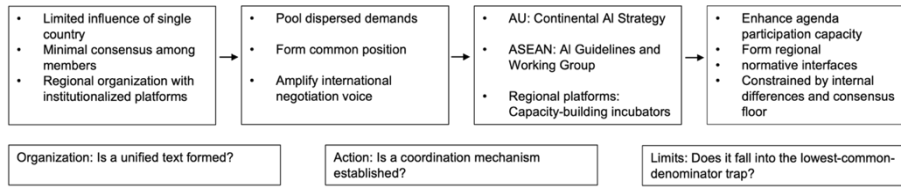


Figure 2 Analytical Diagram of the Collectivization of Agency Strategy

Comparing the collective agency of the African Union and ASEAN yields two findings. First, the form of collective agency is strongly shaped by institutional capacity constraints. The African Union's collective agency

[1] Code for Africa, "Africa's data protection laws began to bite in 2025," May 5, 2026, https://medium.com/code-for-africa/africas-data-protection-laws-began-to-bite-in-2025-322285625315.
[2] ASEAN, "ASEAN Guide on AI Governance and Ethics," February 2024, https://asean.org/book/asean-guide-on-ai-governance-and-ethics/; IMDA, "ASEAN Working Group on AI Governance," https://www.imda.gov.sg/about-imda/international-relations/asean-working-group-on-ai-governance.

focuses on principled strategic coordination, while ASEAN's leans more toward operational guideline development and working group coordination, directly related to ASEAN's higher institutionalization level and Singapore's normative intermediary role. Second, both organizations face constraints on collective action from internal development disparities, but respond differently. The African Union forms unified positions through "lowest-common-denominator" consensus, while ASEAN preserves ample policy flexibility for member states through "soft law" pathways.

3. Localization of Normative Content: Localizing AI Governance Models

The core mechanism of the localization strategy lies in grafting elements of global governance frameworks onto local institutional traditions, cultural values, and developmental practices, thereby forming hybrid norms with local characteristics. Its conditions of application include the existence of legitimate and vital normative traditions or institutional arrangements at the local level, the presence of credible local actors as drivers of the grafting process who are not perceived as external agents, and the existence of graftable space between external and local norms. The practices of ASEAN and Africa provide the most typical empirical evidence for this strategy.

ASEAN's AI governance pathway is a classic case of norm localization. First, it redefines AI governance priorities from "risk prevention" to "innovation promotion and risk balancing." Second, it systematically embeds core elements of the "ASEAN Way," including consensus-building, non-interference, and gradualism, into the AI governance framework. Third, it prunes institutional designs of mandatory compliance and external review characteristic of the EU model, replacing them with voluntary compliance and self-assessment. Fourth, it incorporates the Southeast Asian core value of respecting member state sovereignty and policy autonomy into the governance framework.

The key function of this localization process lies in constructing a regional buffer layer between global norms and national practices. Internally, it buffers normative pressures from external actors such as the EU and the United States, enabling member states to advance domestic AI governance without fully adopting international standards. Externally, it provides discursive resources for ASEAN to participate in global dialogue as a collective. This finding resonates with existing research on proactive norm localization under the IPEF framework, where whether the United States conducts prior localization adaptation of its advocated international norms directly affects

whether those norms are accepted or resisted by ASEAN countries[1]. ASEAN, as a norm recipient, has in fact actively localized external norms, demonstrating the agency of Global South actors[2].

In the process of norm localization, the presence of credible local actors as drivers of grafting constitutes an important condition for localization to take effect. Singapore's role in ASEAN AI governance fully illustrates this point. As the most technologically advanced and institutionally capable ASEAN member state, Singapore's *AI Governance Framework*, first released in 2019, became an important reference template for regional governance, with its core principles later localized in ASEAN-level documents. Singapore's IMDA chairs the ASEAN Working Group on AI Governance and provides operational technical support for regional governance through open-source tools such as AI Verify. This normative intermediary role translates global governance experiences into policy solutions adapted to the ASEAN context, without being perceived as an external agent, thereby ensuring the credibility of norm transmission.

On the African side, norm localization similarly exhibits distinctive cultural selection characteristics. The African Union's *Continental Artificial Intelligence Strategy* emphasizes that "AI systems must adapt to Africa's diversity, languages, cultural heritage, and unique contexts." It should be noted that concepts such as Ubuntu should not be understood as fixed "African essences," but rather as normative resources actively mobilized and reinterpreted by African actors in concrete governance practices. Incorporating the Ubuntu philosophy's emphasis on individual-community interdependence into the AI ethics framework entails greater consideration of community consensus, collective well-being, and intergenerational responsibility in specific issues such as algorithm design, data use, and decision-making transparency, rather than focusing solely on individual privacy and autonomous choice. This practice demonstrates how the "cultural selection" mechanism operates in norm localization, creating hybrid norms within the region that combine international consensus with local cultural values.

[1] Xing Ruili, "Norm Leadership, Proactive Localization, and the Diffusion of U.S.-Style Digital Governance Norms in ASEAN Countries", Contemporary Asia-Pacific, No. 6, 2024.
[2] ASEAN, "Expanded ASEAN Guide on AI Governance and Ethics—Generative AI," January 2025, https://asean.org/wp-content/uploads/2025/01/Expanded-ASEAN-Guide-on-AI-Governance-and-Ethics-Generative-AI.pdf.

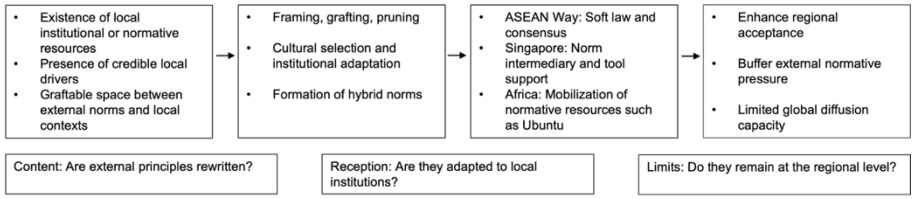


Figure 3 Analytical Diagram of the Localization of Normative Content Strategy

4. Issue-Linkage through Multilateral Action: Connecting AI Governance with Development Agendas

The core mechanism of the multilateral action linkage strategy lies in connecting latecomer domains with existing issues, gaining discursive power and agenda-setting authority through multilateral platforms. Its conditions of application include the availability of linkable, internationally recognized normative resources, channels for voice and the possibility of agenda-setting in multilateral platforms, and a substantive connection between AI governance issues and existing agendas. UN General Assembly resolutions, the India AI Summit, and the BRICS cooperation mechanism provide ample empirical evidence for this strategy.

On July 1, 2024, the 78th session of the UN General Assembly unanimously adopted the resolution on "International Cooperation on Strengthening Artificial Intelligence Capacity-Building," proposed by China and co-sponsored by over 140 countries. The resolution connects AI governance, as an emerging technology issue, with the capacity-building needs of developing countries, calling for the bridging of AI divides both between and within countries, advocating for an open, fair, and non-discriminatory business environment, and supporting the central role of the United Nations in international cooperation[1]. For this resolution, existing normative resources such as the SDGs and South-South cooperation provide an ample normative basis for issue-linkage, while the UN General Assembly, as the most representative multilateral platform, provides institutionalized channels for developing countries' discursive expression. The most fundamental condition, of course, is the genuinely important and substantive connection between AI governance and issues such as the digital divide and technological equity. With approximately 43 percent of the global population lacking internet access, the linkage between AI capacity-building and development agendas has ample empirical grounding.

Furthermore, the 2026 India AI Impact Summit provides another

[1] JURIST News, "UN General Assembly Adopts Draft Resolution to Bridge AI Gap for Developing Countries," July 3, 2024, https://www.jurist.org/news/2024/07/un-adopts-draft-resolution-to-bridge-artificial-intelligence-gap-for-developing-countries/.

important empirical case for the multilateral linkage strategy. During the summit, representatives from Global South countries including Indonesia, Uganda, and Ghana actively connected AI governance with broader development issues such as the digital divide, technological equity, and South-South cooperation. The Africa-Asia AI Policymaker Network held an open session on South-South AI policy cooperation, exploring concrete cooperation roadmaps including infrastructure and data sharing, regulatory experimentation, and mutual recognition frameworks. India also established a tripartite partnership with Italy and Kenya to jointly design scalable sovereign AI pathways, and positioned AI infrastructure and datasets as "global public goods" in the New Delhi Declaration, thereby connecting the demand for technological democratization with existing development justice discourse[1].

On broader multilateral platforms, China's 2025 *Global AI Governance Action Plan* proposes 13 specific measures, connecting AI governance with existing agendas including South-South cooperation, development assistance, and technology transfer through bilateral and multilateral platforms such as the Digital Silk Road, the BRICS cooperation mechanism, and the Forum on China-Africa Cooperation.

The above cases collectively demonstrate that the effectiveness of the multilateral action linkage strategy lies in embedding specific governance issues in a broader discourse of development justice and global equity, enabling the Global South to leverage existing normative resources and moral advantages to support its positions in emerging governance domains. It must be noted, however, that the limitations of this strategy are equally clear: issue-linkage has yet to be fully translated into substantive institutional arrangements and resource allocation mechanisms. Under the structural condition where a few technology companies control the core elements of the AI value chain, Global South agency manifests more as strategic responses and partial adjustments than as fundamental challenges to the power structure itself.

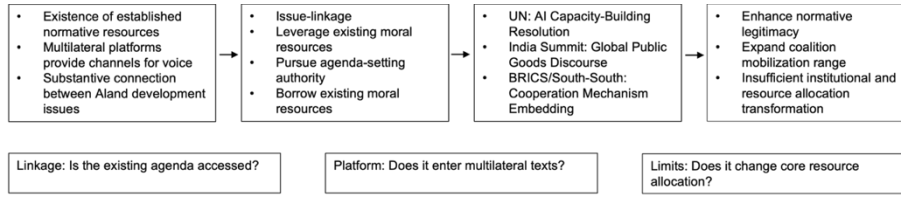


Figure 4 Analytical Diagram of the Issue-Linkage through Multilateral Action Strategy

[1] Government of India, "India AI Impact Summit 2026," https://impact.indiaai.gov.in/; ARC Advisory Group, "From Dialog to Action: India's Landmark AI Impact Summit 2026," https://www.arcweb.com/blog/dialog-action-indias-landmark-ai-impact-summit-2026.

V. Effectiveness and Structural Constraints of Latecomer Domain Agency

The four case clusters analyzed above demonstrate that Global South agency in AI governance norm diffusion exhibits strategic, conditional, and multi-pathway composite characteristics. The application of each strategy is shaped by specific conditions, and different strategies are not mutually exclusive but often overlap. For instance, the African Union's practice simultaneously embodies the development-oriented issue framing and the collectivization of agency strategies, while ASEAN's practice simultaneously embodies the collectivization of agency and the localization of normative content strategies. This strategy overlap suggests that the same actor may display different levels and types of agency across different issue areas or at different times.

Furthermore, the influence and actual effects of different strategies vary considerably. This article observes agency effects at three levels: the discursive level, namely whether actors can change the priority ordering and problem definition of AI governance issues; the institutional level, namely whether these discourses can be translated into regional guidelines, national legislation, working groups, or capacity-building mechanisms; and the resource level, namely whether normative claims can further influence actual allocation in technology transfer, funding support, computing power acquisition, and standard-setting. By this measure, the development-oriented strategy can draw international attention to the special needs and concerns of developing countries, but remains limited in translating discursive influence into concrete institutional arrangements and resource allocation. The collectivization strategy proves relatively effective in enhancing negotiating status and forming unified positions, but is constrained by internal divisions and collective action dilemmas, often yielding only minimal consensus. Localization of normative content can achieve certain institutional outcomes at the regional level, but its demonstration effects and radiation capacity at the global level remain limited.

It follows that the space, forms, and effects of agency in latecomer domains are shaped and constrained by both internal and external structural conditions. Understanding agentic behavior requires simultaneous attention to actors' strategic choices and the structural environments in which they are embedded, thereby avoiding two theoretical biases: neglecting structural constraints or denying the possibility of agency. These constraints not only delimit the possible space for agentic behavior but also, to a considerable extent, determine the selection of agentic strategies and the variation in their effects.

First, development needs play a dual role. In the context of this article, Global South countries commonly face practical challenges typical of latecomer domains, including weak digital infrastructure, low internet penetration rates, shortages of technical talent, and insufficient R&D investment. These development needs produce a dual effect on agentic behavior. On the one hand, development needs constitute the impetus for accepting external technological norms, namely acquiring technology transfer, capacity building, and funding support through participation in global governance frameworks. On the other hand, development needs also constitute the fundamental basis for localized adaptation, ensuring that governance norms serve development objectives rather than becoming development obstacles or new technical barriers. African countries' refusal to simply replicate the EU-style risk control framework stems from the core concern that excessive regulation might stifle the nascent AI innovation ecosystem. The analytical significance of the development needs constraint lies in revealing that Global South countries' responses to external norms, whether selective adoption, creative adaptation, or autonomous construction, depend to a large extent on the degree of alignment between those norms and their development priority agendas.

Second, institutional capacity generates internal differentiation. Actors' institutional capacities in norm formulation, implementation, and oversight vary considerably, directly affecting the form and effectiveness of agentic behavior. Countries with stronger institutional capacity can independently formulate relatively comprehensive AI governance norms and play active roles in global governance dialogue. Countries with medium institutional capacity typically develop domestic policies with reference to international frameworks while participating in global governance through regional coordination mechanisms. Countries with weaker institutional capacity tend to rely on regional organizational platforms for incremental capacity building and collective agency. The analytical significance of the institutional capacity constraint lies in revealing the dual function of regional organizations in Global South agency: they serve as both organizational carriers for collective agency and incubators for institutional capacity building. The African Union and ASEAN, by integrating dispersed demands of individual countries into unified regional positions, compensate to some extent for the institutional capacity deficiencies of individual member states, enabling them to participate in global governance dialogue as collective actors. This intermediary function is a key variable for understanding the diversity of Global South agency pathways.

Third, the international structure solidifies domain weakness. This constitutes the most powerful constraint facing actors exercising agency in

latecomer domains. The existence and expansion of the "intelligence divide" not only locks Global South countries at the bottom of the industrial chain but also further weakens their discursive power in AI agenda-setting and standard-setting games, while giving rise to new forms of governance inequality such as "algorithmic dependency" and "digital colonialism[1]." Weymouth's research reveals the emerging trend of "digital disintegration" in the AI field, whereby a few major technological powers and multinational corporations control the core elements of the AI value chain, including advanced chips, computing resources, and foundation models. These "digital warlords" exercise power beyond the jurisdictional reach of domestic regulation and multilateral oversight, controlling the discourse on who can access advanced AI technologies and under what conditions[2]. Strategic competition among states over these core elements has given rise to competing "techno-blocs," and the Global South faces a structural risk of marginalization, lacking autonomous AI technological capacity while being highly dependent on a few major powers and corporations for technology supply. This structural dependency constitutes the external boundary of agentic behavior. Agency in latecomer domains manifests more as strategic responses and partial adjustments within existing power structures than as fundamental challenges to those structures themselves[3]. Even when the Global South demonstrates proactive agency, such agency always operates under conditions of significant power asymmetry.

Consequently, agentic strategies face three systemic risks. First, co-optation: Global South agency is incorporated into existing dominant frameworks by major powers or international organizations, diluting its independence and critical edge. Second, symbolization: Global South participation is treated as decoration or embellishment for multilateral governance legitimacy rather than substantive normative co-construction. Third, marginalization: amid intensifying techno-bloc competition, the Global South continues to be objectively excluded from core, widely followed and applied norm-formulation and technical standard-setting processes. The above case analysis also fully validates the analysis of the threefold structural constraints in this article's theoretical framework. The development needs constraint shapes the fundamental driver of the development-oriented strategy; the institutional capacity constraint affects the specific forms of collective

---

[1] Chen Zhimin, Wu Zhicheng, Sun Jisheng, et al., "Global Governance Initiatives and Major-Country Diplomacy with Chinese Characteristics," pp. 1-41.
[2] Stephen Weymouth, "Digital Disintegration: Techno-Blocs and Strategic Sovereignty in the AI Era," *International Organization*, Vol. 79, Supplement, 2025, pp. S57-S70.
[3] Que Tianshu and Zheng Zhaochen, "Challenges for the 'Global South' in Participating in International AI Governance and China's Options," *Contemporary China and World*, No. 3, 2025, pp. 14-23.

agency and norm localization; and the international structural constraint sets the external boundary of agency. Global South agency should therefore be understood as a limited but meaningful form of agency. It does not mean that the Global South can dominate norm construction, but it nevertheless represents the Global South's ongoing efforts to gain discursive space, improve institutional environments, and preserve development autonomy within structural constraints.

VI. Conclusion

Centering on the question of how actors exercise strategic agency in the process of norm-building within their latecomer domains, and building on existing norm diffusion research while incorporating Acharya's theories of norm localization and norm subsidiarity, this article has constructed a conditional agency analytical framework that identifies four types of agentic strategies: development-oriented issue framing, collectivization of agency, localization of normative content, and issue-linkage through multilateral action. Theoretically, this article extends norm diffusion research from traditional security and human rights domains into emerging technology governance domains at the issue level, while providing an analytical pathway that balances agency and structure at the methodological level. Empirically, through case analyses of African countries, ASEAN, India, and Global South multilateral cooperation, this article yields three core findings.

First, the Global South demonstrates diverse and strategic agentic practices in AI governance norm diffusion. This agency unfolds through multiple pathways including development-oriented framing, regional collective agency, normative localization and grafting, and multilateral issue advocacy, reflecting Global South countries' active efforts to preserve development autonomy and participate in global norm construction. The implicit assumption in existing norm diffusion research that the Global South constitutes passive norm recipients requires revision.

Second, Global South agency exhibits strategic, conditional, and multi-pathway composite characteristics. The specific forms and actual effects of agentic behavior are jointly shaped by three conditions: development needs, institutional capacity, and international structure. While constrained by structural conditions, agency nonetheless always contains actors' strategic choices and creative adaptations. Agency should therefore be understood as a limited but meaningful form of agency, rather than a reversal of normative dominance or a fundamental transformation of power structures.

Third, understanding Global South agency is of great significance for

constructing a more just and effective global AI governance order. Norm circulation and even norm contestation better capture the actual dynamics of AI governance norms at the global level than traditional norm diffusion models. The sustainability and legitimacy of global AI governance depend on liberating Global South agency from marginalization or co-optation, making it a formal and substantive participant in normative co-construction.

From this perspective, the goal and ideal of inclusive AI cannot be achieved through technology transfer alone[1]. Technology transfer is certainly important, but if the Global South lacks substantive participation and influence in the formulation of governance norms, technology transfer may paradoxically reinforce existing power asymmetries. Developing countries may acquire technology while being compelled to accept attendant governance standards and institutional arrangements, thereby losing the capacity to choose and adapt governance pathways according to their own needs. Truly inclusive AI requires recognition and institutionalization of Global South agency space at the governance norm level, enabling developing countries to participate in and shape AI governance orders from the very source of norm formulation. Furthermore, global AI governance needs to shift from a unidirectional norm transmission model toward a bidirectional norm co-construction model. While the existing global AI governance framework has conducted global consultations to some extent during its formulation, in terms of core concepts, priority issues, and institutional design, it remains dominated by the experiences and demands of developed countries. The future global AI governance framework should provide institutionalized channels for expression and influence mechanisms for the Global South's diverse agentic strategies, incorporate Global South experiences of agency into the formal agenda of global governance, and enable the Global South to truly exert its strategic influence in promoting the reform and construction of the global governance system[2].

As an important member of the Global South and a major participant in the AI field, China occupies a distinctive position and enjoys unique advantages in promoting more just and inclusive global AI governance. On the one hand, China's own AI governance practices provide a reference pathway for the Global South that differs from Western models. On the other hand, capacity building, technology sharing, and joint norm-shaping under the South-South cooperation framework can not only enhance the collective

[1] Xue Lan and Liang Zheng, "Building an Inclusive and Equitable Global AI Governance System: Real Challenges and Policy Recommendations," *People's Forum·Academic Frontiers*, No. 9, 2025, pp. 23-32.
[2] Zhao Kejin, "The 'Global South' in the Centurial Transformation," *People's Forum·Academic Frontiers*, No. 23, 2023, pp. 5-12.

agency capacity of the Global South in AI governance but also provide a technological governance pathway for the vision of "a community with a shared future for mankind[1]." In advancing these cooperative endeavors, China should fully respect its partners' development autonomy and agency space, avoid replicating the Western "top-down" model of norm export, and explore more equal, mutually beneficial, and inclusive pathways of normative co-construction.

[1] Liang Zheng, Tian Guiping, and Song Yuxin, "Multi-dimensional Dilemmas and China's Solutions in Global AI Governance," *Tsinghua Financial Review*, No. 9, 2025, pp. 20-22.

Issues and Norms

Institutional Construction of Artificial Intelligence Data Security Governance from the Perspective of the Global South

*Xu Duoqi Ying Yue**

Abstract In the era of artificial intelligence, practices of data flow and model training have transformed data colonialism into AI colonialism, a more intense and more concealed form of domination. Countries in the Global South are structurally disadvantaged in relation to data exploitation and national security, while existing solutions remain inadequate. Despite this unfavorable position, the Global South can still leverage its own resources to engage in asymmetric bargaining. Vertically, it can strategically choose among the services and rules offered by multiple AI innovation hubs, thereby generating competitive pressure. Horizontally, it can expand South-South cooperation to strengthen its collective bargaining leverage. The implementation of this strategy requires institutional design along three dimensions. In the dimension of security-oriented competition, institutions should reward AI service providers that respect data sovereignty and promote a turn toward open-source models. In the dimension of security baselines, institutions should prevent critical AI infrastructure from becoming excessively dependent on any single source of service provision. In the dimension of security cooperation, countries in the Global South should strengthen solidarity and pursue international collaboration through favorable mechanisms such as the United Nations and the Belt and Road Initiative.

* Xu Duoqi, Professor at the School of Law, Fudan University; Director of the Research Center for Digital Economic Rule of Law and the Smart Rule of Law Laboratory, Fudan University. Ying Yue, PhD Candidate at the School of Law, Fudan University; Research Fellow at the Research Center for Digital Economic Rule of Law and the Smart Rule of Law Laboratory, Fudan University; Visiting Scholar at Cornell Law School, Cornell University.

This paper represents a phased research output of the Major Project of the National Social Science Fund of China, titled A Comparative Study on Data Security Governance Strategies, Laws and Regulations of Major Developed Countries (Project No. 25&ZD244).

Keywords Global South; Data Security; Artificial Intelligence Colonialism; Asymmetric Bargaining

I. Data Security and Colonialism

The past never fades away, nor has it ever truly passed. Within a relatively peaceful international landscape, traditional colonialism centered on territorial seizure and resource plunder has transformed into a new form defined by data exploitation. This shift receives systematic elaboration under the concept of "data colonialism," whose core logic constructs a new "colonial order." Its focus departs from spatial occupation, and instead captures human life through data extraction. By converting human experiences and behaviors into computable data for monopolized exploitation, actors extract value unilaterally, replicating the colonial expansion dynamic under which colonizers seized land and labor from colonized populations[1]. Parallel to this framework, the notion of "digital colonialism" places greater emphasis on asymmetric power structures across global systems, particularly the reproduction of "core-periphery" hierarchies within digital spaces[2]. The digital infrastructure centered on major U.S. technology companies has imposed a form of structural colonial domination over the Global South. These firms control foundational pillars of global digital ecosystems including software, hardware and network connectivity. Closed-source code generates intellectual property rents, while massive datasets power analytical, inferential and predictive capabilities to sustain market dominance. Locked platform network effects consolidate monopolies, and all resultant profits flow back to home states[3].

Building on these two frameworks, scholars have coined the term "AI colonialism" to describe how AI-centric technological systems restructure global power architectures and value distribution mechanisms[4]. Data forms the foundational input for AI model training, which draws extensively from

[1] Jim Thatcher, David O'Sullivan and Dillon Mahmoudi, "Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data," *Environment and Planning D: Society and Space*, Vol. 34, No. 6, 2016, pp. 990–1000; Nick Couldry and Ulises A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media*, Vol. 20, No. 4, 2019, pp. 336–349.
[2] Raúl Prebisch, *The Economic Development of Latin America and its Principal Problems*, United Nations/ECLA, 1950, pp. 1–10.
[3] Michael Kwet, "Digital Colonialism: US Empire and the New Imperialism in the Global South," *Race & Class*, Vol. 60, No. 4, 2019, pp. 3–26.
[4] James Muldoon and Boxi A. Wu, "Artificial Intelligence in the Colonial Matrix of Power," *Philosophy & Technology*, Vol. 36, Article 80, 2023, pp. 1–24; Salvador Santino F. Regilme, "Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South," *SAIS Review of International Affairs*, Vol. 44, No. 2, 2024, pp. 75–92.

daily user behaviors, linguistic expressions and social interactions. The collection and utilization of such data perpetuate the cycle of social life's continuous datafication into extractable resources. Advanced AI systems ingest massive volumes of publicly accessible online materials[1], yet Northern domestic data and copyright holders only maintain partial access to domestic legal remedies for compensation, regardless of remedy adequacy[2]. "AI colonialism," by contrast, permanently positions the Global South as a low-cost raw data source for Northern AI model development. This sustains an asymmetric cycle where the Global South, as the periphery, supplies low-return raw data, while the Global North, as the core, develops high-value models and retains control over governance mechanisms, platform rules and underlying infrastructure[3].

"Data colonialism" and "digital colonialism" frame platform ecosystems anchored in local commerce and consumption, such as advertising services that capture a fixed proportion of the Global South's economic output with inherent revenue ceilings. The technical architecture underpinning large language models, the new generation of AI systems, compresses and encodes knowledge captured from massive datasets once training concludes. Trained models replicate, deploy and execute reasoning at superior efficiency and price competitiveness. This widens the steep value disparity between peripheral raw data inputs and core AI product outputs, generating an accelerating feedback loop. Greater data contributions from the Global South enable the Global North to train higher-value AI systems, which the Global South then relies on increasingly heavily. Each cycle of AI service consumption generates further data inflows to Northern developers. Rising AI value from core economies depreciates the relative worth of Southern data in exchange, widening the "price scissors between Southern raw data and Northern AI outputs." Continuous value transfers flow from the Global South to the Global North via API calling fees and subscription charges, driven by the data flywheel, constrained less by the scale of Southern economic output and more by the operational force of "scaling laws." Monopolization further enables exorbitant pricing levied exclusively on Southern end users. AI deployment also triggers a prisoner's dilemma among Southern enterprises.

---

[1] OpenAI, "GPT-5.5 System Card," published 23 April 2026, https://deploymentsafety.openai.com/gpt-5-5/model-data-and-training. Anthropic, "Anthropic's Transparency Hub," last visited 26 April 2026, https://www.anthropic.com/transparency.

[2] *Bartz v. Anthropic PBC*, No. 3:24-cv-05417-WHA, 2025 WL 1741691 (N.D. Cal. June 23, 2025); *The New York Times Co. v. Microsoft Corp.*, No. 23-cv-11195 (SHS), 2025 WL 1009179 (S.D.N.Y. Apr. 4, 2025).

[3] Selena Nemorin and Beatrice Bonami, "AI Ethics through a Decolonial Lens: What Does AI Ethics Look Like if We Take Seriously the Push to Decolonise It," *AI & Society*, 2026, pp. 1–11.

While AI automation boosts operational efficiency and displaces human labor, resultant aggregate demand contraction carries negative externalities borne by all market participants. Individual firms capture all cost savings from automation, creating incentives for over-reliance on AI. Enterprises adopt AI en masse to secure exclusive residual gains and avoid competitive displacement, even with full awareness that excessive automation erodes shared consumer demand. Cumulative over-adoption ultimately triggers aggregate economic contraction across Southern economies[1]. When mass unemployment emerges, the AI service providers accumulating the largest revenues lack tax residency within Global South jurisdictions. Southern states possess limited corrective tools beyond VAT or sales tax collection, or coordinated cross-border efforts to integrate these corporations into the BEPS 2.0 Pillar One framework. AI expansion exposes the Global South to unprecedented crises in data factor security concerning the preservation and utilization of value embedded within data assets.

Shifts from peaceful coexistence to confrontation or armed conflict escalate data security risks from economic value leakage into existential national security threats. In terms of cyberattack capacity, independent of AI service supply disruptions, frontier AI models trained on Southern data deliver formidable cyber operational capabilities that pose severe threats to Southern financial networks, power grids, communication systems, airports, hospitals, water conservancy works and other critical national infrastructure[2]. For intelligence, surveillance and strike planning, AI systems such as Palantir fuse multi-source datasets to generate precise strategic and tactical targeting decisions, as demonstrated by extraordinary information processing efficiency during the Gaza and Iran conflicts[3]. Data flows from the Global South to Northern jurisdictions now serve dual purposes, spanning commercial model training and military strike intelligence development.

Scholarly analysis identifies concentrated design, production, deployment and export of "autonomous weapons systems" within a small cohort of militarized tech powers. The Global South becomes a dual-purpose space functioning as raw data supply bases, live testing grounds and operational theaters for such weapons, with state sovereignty subjugated to

---

[1] Brett Hemenway Falk and Gerry Tsoukalas, "The AI Layoff Trap," arXiv:2603.20617 [econ.TH], Mar. 21, 2026, pp. 1–32.
[2] Liv McMahon and Joe Tidy, "What Is Claude Mythos and What Risks Does It Pose?" *BBC News*, 18 April 2026, https://www.bbc.com/news/articles/crk1py1jgzko; UK AI Security Institute (AISI), "Our Evaluation of Claude Mythos Preview's Cyber Capabilities", 13 April 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.
[3] Du Zhihang, "War Opens AI's 'Pandora's Box'," *Caixin Weekly*, No. 11, 23 March 2026, https://weekly.caixin.com/2026-03-21/102425684.html.

these military AI applications. This structural dynamic mirrors historical colonial hierarchies that positioned colonies as resource mines, military experimentation zones and subjugated populations, replacing Maxim machine guns with contemporary military AI[1]. Escalating geopolitical tensions drive qualitative shifts in data security risk manifestations, marked by civilian-to-military AI capability conversion[2]. Data factor security during peacetime primarily manifests as value erosion, economic exploitation and deepening technological dependency. Under confrontation or wartime conditions, security risks spill over from asset protection to cross-domain information network vulnerabilities. Threat vectors include AI-powered cyberattacks targeting critical infrastructure, alongside precision intelligence strikes and autonomous weapons decision-making pipelines built on fused multi-source data that undermine territorial sovereignty and national survival. The Global South occupies a structurally disadvantaged position within both economic and security value chains governing global data flows.

II. Existing Responses of the Global South and Their Limitations

The Global South has taken certain countermeasures against AI colonialism, yet all prevailing solutions display tangible defects in practical implementation. This section delivers a review of these response strategies.

1. Existing Countermeasures against AI Colonialism

Data cross-border flows constitute the source of risks to data factor security, which renders data localization the most intuitive response measure. Even if data is physically stored within the country, AI service providers can still extract core data value via cloud service interfaces for model training and feature extraction, thereby effecting the cross-border transfer of the value embodied in that data. Other security regulatory models also carry drawbacks of excessive implementation costs and insufficient regulatory efficacy. Early scholarly research points out that *General Data Protection Regulation* (GDPR) interprets data flows under the paradigm of traditional personal data protection law. The core value of big data, however, originates from large-scale data aggregation, multi-source data correlation, secondary exploitation

[1] Olivia Zhang and Rujuta Karekar, "Autonomous Weapons Systems as Digital Colonialism," *Stop Killer Robots*, 18 August 2025, https://stopkillerrobots.medium.com/autonomous-weapons-systems-as-digital-colonialism-21fc9a29ce8d.

[2] Qiao-Franco, G. and M. Javadi, "Bridging the Gap Between the North and South in the Governance of Dual-Use Artificial Intelligence Technologies," in L. Belli and W. B. Gaspar, eds., *AI from the Global Majority*, Cham: Springer, 2026, pp. 147–155.

for subsequent purposes, and machine-driven correlation discovery. GDPR rules including collection under predefined specific purposes, data minimization, segregation of ordinary and sensitive data, and obligation to provide a logical explanation for automated decision-making generate varying degrees of tension with the openness, reusability and predictability inherent to big data processing[1]. The limitations of GDPR rules grow more prominent within the AI era. The right to erasure serves as an illustrative example. Under the self-attention mechanism of Transformer architecture, every token computes attention weights alongside all other tokens within a sequence to form global contextual correlations, which update all parameter layers of the model via backpropagation gradients. After iterative training on massive datasets, the influence of any single piece of personal data no longer exists in isolation within a small set of parameters. Instead, such influence integrates into weight matrices of billions or even trillions of parameters in a highly dispersed and coupled manner. This technical reality creates substantial barriers to the precise localization and subsequent erasure of contributions made by specific personal data to model parameters, and poses severe challenges to effective exercise of the right to erasure as a remedial instrument for data security[2].

2. Practical Status of AI Service Providers

User agreements, privacy policies, data processing rules, model training policies and security management commitments formulated by AI service providers do not automatically constitute formal legal sources under domestic law, yet they form the practical expression of data security norms within AI service scenarios. The Global South faces a disadvantageous position amid such institutional arrangements.

(1) Data Localization

Data flows within AI services feature stronger dynamism, compound attributes and derivative characteristics compared with conventional internet services. User inputs, outputs, logs, cache and other data may reside within distinct technical links and geographical boundaries. Two tiers of data localization therefore demand differentiation: the first tier refers to data residency at rest, and the second tier stands for inference residency.

---

[1] Tal Z. Zarsky, "Incompatible: The GDPR in the Age of Big Data," *Seton Hall Law Review*, Vol. 47, No. 4, 2017, pp. 995–1020.

[2] Ludovica Robustelli, "Le droit à l'oubli à l'ère de ChatGPT: un droit 'effacé'," *Cahiers Droit, Sciences & Technologies* [En ligne], 21 | 2026, mis en ligne le 01 mars 2026, consulté le 08 mars 2026, http://journals.openedition.org/cdst/13450. See also Nicholas Carlini et al., "Quantifying Memorization Across Neural Language Models," arXiv:2202.07646, 2022.

Data residency at rest centers on the geographic storage location of persistent customer data. OpenAI's public documentation defines data residency as the storage of static customer content within designated geographic regions, covering dialogues, uploaded files, image generation inputs and outputs, Canvas content and other materials. This capability primarily applies to qualified API clients and newly enrolled ChatGPT Enterprise/Edu users rather than all consumer-grade chat product users, with current supported regions limited to Australia, Canada, Europe, India, Japan, Singapore, South Korea, the United Arab Emirates, the United Kingdom and the United States[1]. Anthropic's public disclosures indicate that, for commercial products such as Claude Enterprise and the Anthropic API, their default settings may route customer traffic to be processed in several regions across the United States, Europe, Asia, and Australia, while static data storage occurs exclusively within the United States[2].

Inference residency governs geographic control over model inference processing. User inputs, segments of uploaded files, context from retrieval-augmented generation, system prompts, model outputs and embedding vectors generated from content may enter graphics processing units (GPUs) or alternative computing resources during inference, generating KV cache, document slices and other temporary cached data. Multiple factors determine data localization effectiveness: whether local execution completes all inference tasks, whether GPU operations are confined to designated territories, and whether cross-regional load balancing, disaster recovery scheduling or global routing mechanisms exist. ChatGPT's inference residency currently supports Europe, the United States and the United Arab Emirates only. The platform additionally clarifies that non-GPU workflows such as authentication, authorization, routing, orchestration, analytics and third-party service integration may operate outside selected regions[3]. Anthropic restricts inference geographic options to "Global" and "United States" solely[4]. While

[1] OpenAI, "Data residency and inference residency for ChatGPT," *OpenAI Help Center*, last updated 24 April 2026, https://help.openai.com/en/articles/9903489-data-residency-and-inference-residency-for-chatgpt.

[2] Anthropic, "Where are your servers located? Do you host your models on EU servers?" *Anthropic Privacy Center*, last visited 25 April 2026, https://privacy.claude.com/en/articles/7996890-where-are-your-servers-located-do-you-host-your-models-on-eu-servers; Anthropic, "Data residency," *Claude API Docs*, last visited 25 April 2026, https://platform.claude.com/docs/en/build-with-claude/data-residency.

[3] OpenAI, "Data residency and inference residency for ChatGPT," *OpenAI Help Center*, last updated 12 May 2026, https://help.openai.com/en/articles/9903489-data-residency-and-inference-residency-for-chatgpt.

[4] Anthropic, "Where are your servers located? Do you host your models on EU servers?" *Anthropic Privacy Center*, last visited 25 April 2026, https://privacy.claude.com/en/articles/7996890-where-are-your-servers-located-do-you-

Microsoft Foundry and Google Vertex AI offer broader inference residency coverage beyond the US and EU, their service regions remain concentrated in the Global North[1].

Data localization within AI services constitutes a composite issue shaped by cloud infrastructure network topology, model inference architecture, service agreements and product tiering. The Global South must remain vigilant as to whether the value of its data is being progressively extracted, transformed, and accumulated into the model capabilities of Northern enterprises through successive stages, including storage, inference, caching, logging, embedding, tool calling, and model optimization. Enterprise clients with strong bargaining power and sufficient compliance budgets can purchase high-level data control services, while ordinary individual users and small and medium-sized organizations across the Global South are forced to accept default training rules, global routing and low-transparency processing arrangements. Furthermore, mainstream data residency service regions are inherently centered on the Global North. OpenAI's published residency lists are almost entirely absent from Latin American and African states, with the United Arab Emirates as the sole Middle Eastern jurisdiction on the roster. Google Vertex AI extends regional coverage to Brazil yet excludes all other Latin American nations, the whole African continent, other Middle Eastern countries and most South Asian countries. Even if governments or enterprises in Peru, Kenya, Bangladesh and comparable jurisdictions explicitly stipulate localization requirements in their contracts, their data may still be "localized" to a Northern node thousands of kilometers away.

(2) Data Deletion

Data deletion mechanisms are of limited efficacy for consumer-facing users. OpenAI stipulates that deletion requests lose binding force once data enters model improvement workflows under de-identified forms. Google Gemini Apps establishes a rule that once a user's active behavior is selected for sampling and enters manual review, the relevant content may persist on Google's servers for up to three years even after the user subsequently deletes all activity records, and disabling activity records does not prevent anonymized processing of the user's data[2]. In contrast, deletion commitments on the commercial and API side are relatively robust. For instance,

---

host-your-models-on-eu-servers; Anthropic, "Data residency," *Claude API Docs*, last visited 25 April 2026, https://platform.claude.com/docs/en/build-with-claude/data-residency.

[1] Google Cloud, "Data residency," *Generative AI on Vertex AI*, last updated 22 April 2026, https://docs.cloud.google.com/vertex-ai/generative-ai/docs/learn/data-residency.

[2] Google, "Gemini Apps Privacy Hub," *Google Support*, last visited on 18 April 2026, https://support.google.com/gemini/answer/13594961?hl=en.

Anthropic's enterprise edition offers a "Zero Data Retention" (ZDR) arrangement, under which "Anthropic does not retain your inputs or outputs except as required to comply with law or combat abuse," while "Anthropic still retains user safety classifier results to enforce our usage policies." The ZDR applies only to "eligible Anthropic API" and "Anthropic products using commercial organization API keys," not to consumer-grade subscription plans such as Claude Free, Pro, or Max[1].

Three structural disadvantages confront the Global South. First, consumer users operate under long-standing institutional exemptions for data de-identification. Second, national legislation establishing the right to erasure or data deletion loses normative force due to the technical reality of dispersed data embedding within Transformer model weight parameters. This weakness overlaps with contractual exemptions for de-identification and extended retention of audit data set by all AI service providers. Third, all contractual provisions include compliance carve-outs. If jurisdictions with substantial security concerns enforce mandatory data seizure via legislation, the Global South holds no effective countermeasures.

(3) Input Content and Retraining Authorization

AI service providers have generally established a dual-track system featuring broad authorization on the consumer side and default no-training on the commercial and API side. The consumer side may be further divided into two subtypes.

The first type, represented by xAI, adopts broad authorization in its consumer terms, which require users to "grant xAI an irrevocable, perpetual, transferable, sublicensable, royalty-free, worldwide right to use, reproduce, store, modify, distribute, reprint, publish, display publicly in public forums, list information related to, make derivative works from, and aggregate your User Content and derivative works thereof for any purpose, including but not limited to: (a) maintaining and providing the Services; (b) improving our products and services and for our other commercial purposes such as data analysis, customer and market research, developing new products or features, or identifying or demonstrating usage or User Content trends; and (c) taking other measures to enforce these Terms, comply with our Privacy Policy, comply with applicable law, or ensure the security of our Services."[2]

---

[1] Anthropic, "I have a zero retention agreement with Anthropic—what products does it apply to?" *Anthropic Privacy Center*, last visited on 18 April 2026, https://privacy.claude.com/en/articles/8956058-i-have-a-zero-retention-agreement-with-anthropic-what-products-does-it-apply-to; OpenAI, "Services Agreement," *OpenAI Policies*, last visited on 18 April 2026, https://openai.com/policies/services-agreement/.

[2] xAI, "Terms of Service," *xAI Legal*, last visited on 18 April 2026,

The second type, represented by Anthropic's consumer terms, offers an opt-out mechanism subject to exceptions: "We may use materials to provide, maintain, and improve the Services and to develop other products and services, including training our models, unless you opt out of training in your account settings," while simultaneously providing that "even if you opt out, we will still use materials for model training in the following circumstances: (a) you provide feedback to us on any materials, or (b) your materials are flagged in connection with a safety review, to improve our ability to detect harmful content, enforce our policies, or advance safety research."[1] Google Gemini API[2] and Mistral's EU consumer terms[3] adopt similar models. Many AI service providers also offer "temporary chat" or "private chat" modes on the consumer side, stipulating that such chats do not appear in chat history and are not used for model training.

The commercial and API side presents a contrasting picture. Anthropic's commercial terms explicitly provide: "Anthropic will not train models on Customer Content from the Services."[4] xAI's enterprise terms similarly state: "xAI shall not use any User Content for its internal AI or other training purposes." [5] Microsoft 365 Copilot documentation likewise specifies: "Prompts and responses are not used to train foundation models."[6]

For the Global South, the dual-track structure of training authorization carries structural significance. Broad authorization on the consumer side means that ordinary users of free or consumer-grade products face a substantially higher probability of their data being used for model training compared with enterprise customers accessing services through APIs, a dynamic that disadvantages countries with limited corporate customer bases. Northern enterprise customers can secure default opt-out protection through commercial contracts, while individual users in both the North and South generally use products as consumer-grade users, rendering their interaction data training materials. In the South, however, where commercial AI

---

https://x.ai/legal/terms-of-service.

[1] Anthropic, "Consumer Terms of Service," *Anthropic Legal*, last updated on 8 October 2025, https://www.anthropic.com/legal/consumer-terms.

[2] Google, "Gemini API Additional Terms of Service," *Google AI Developers*, last visited on 18 April 2026, https://ai.google.dev/gemini-api/terms.

[3] Mistral AI, "EU Consumers Terms of Service," *Mistral AI Legal*, last visited on 18 April 2026, https://legal.mistral.ai/terms/eu-consumers-terms-of-service.

[4] Anthropic, "Commercial Terms of Service," *Anthropic Legal*, last visited on 18 April 2026, https://www.anthropic.com/legal/commercial-terms.

[5] xAI, "Enterprise Terms of Service," last visited on 18 April 2026, https://x.ai/legal/terms-of-service-enterprise.

[6] Microsoft, "Microsoft 365 Copilot Chat Privacy and Protections," *Microsoft Learn*, last visited on 18 April 2026, https://learn.microsoft.com/en-us/copilot/privacy-and-protections.

deployment remains relatively thin, consumer-grade usage likely accounts for a higher proportion, and the total volume of data collectively contributed to Northern model training may, relative to economic scale, exceed the returns the South can capture from the global AI value chain.

(4) Output Rights and Model Distillation

Distillation refers to a training methodology in which powerful "teacher models" generate responses, which are then used to fine-tune smaller "student models." Leading Northern AI service providers adopt rigid prohibitive stances against external competitive distillation. OpenAI forbids "any automated or programmatic means to extract data or outputs from the Services" and the use of "outputs to develop models that compete with OpenAI."[1] Google Gemini API supplementary terms similarly provide: "You may not use the Services to develop models that compete with the Services (e.g., the Gemini API or Google AI Studio). You also may not attempt to reverse engineer, extract, or copy any component of the Services, including underlying data or models such as parameter weights."[2] Microsoft's Services Agreement goes further in its AI services provisions, extending the prohibition beyond competitive models to include directly or indirectly creating, training, or improving any AI technology.[3]

Model distillation magnifies structural North–South conflicts. Southern enterprises attempting technological catch-up via distillation of frontier Northern models may be publicly named and characterized as violating terms of service or endangering national security.[4] Distillation thereby evolves into an institutional tool targeting Southern industrial upgrading attempts. Notably, laboratories accused by Anthropic have delivered large language models with comparable performance to ChatGPT and Claude across Chinese and multilingual scenarios, which constitute viable diffusion channels for cutting-edge model technologies toward the Global South. Once this pathway is completely blocked, the Global South will find itself not only at a disadvantage on the data supply side, but also facing even greater difficulties in achieving technological leaps on the model capability side.

---

[1] OpenAI, "Terms of Use (Non-European Regions)," *OpenAI Policies*, last updated on 1 January 2026, https://openai.com/policies/row-terms-of-use/.
[2] Google, "Gemini API Additional Terms of Service," *Google AI Developers*, last visited on 18 April 2026, https://ai.google.dev/gemini-api/terms.
[3] Microsoft, "Microsoft Services Agreement," §14.s.iv (AI Services), *Microsoft Legal*, last visited on 18 April 2026, https://www.microsoft.com/en-us/servicesagreement.
[4] Anthropic, "Detecting and preventing distillation attacks," *Anthropic News*, 23 February 2026, https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks.

III. Asymmetric Games and Institutional Construction for the Global South

The foregoing analysis demonstrates the systematic defects of existing countermeasures adopted by the Global South. Nevertheless, superior strength does not guarantee inherent dominance, nor does weakness equate to complete powerlessness. Asymmetric interactions between powerful and weak actors fundamentally shape final outcomes.

1. Analytical Framework of Asymmetric Interaction Theory

Asymmetric interaction theory illustrates the dynamism and malleability of power disparities between strong and weak actors. Institutional strategies of weaker states manifest in two forms. The first is relational power, which enables weak actors to secure more favorable outcomes, redistribute benefits and alter rival conduct within established international regimes. The second is meta power, which reconstructs negotiation frameworks and benefit distribution logics by reshaping the core principles, rules and decision-making procedures of institutional structures themselves.[1]

In terms of relational power, the Global South, as disadvantaged actors, is not confined to passive compliance with the international system. Instead, it may strategically arbitrage within such a system. Weak states may maintain parallel, limited ties with two or more competing major powers and refrain from premature unilateral alignment, reserving policy flexibility amid shifting geopolitical conditions to expand strategic maneuver space amid great power rivalries[2]. Furthermore, disadvantaged parties may translate specific stances into binding commitments visible to rivals through legal constraints, institutional arrangements, public statements and the invocation of established principles and precedents. Such self-binding mechanisms enhance negotiation credibility and narrow opponents' room to extract concessions through coercion[3]. Once domestic legislation frames data security as non-negotiable bottom lines, governments gain legitimate grounds to reject compromises in international negotiations. Thus, the leverage of weak states is not limited to military strength, capital, or market size, but also includes what international

---

[1] Stephen D. Krasner, *Structural Conflict: The Third World Against Global Liberalism*, Berkeley: University of California Press, 1985, pp. 6–60.
[2] Kuik Cheng-Chwee, "The Essence of Hedging: Malaysia and Singapore's Response to a Rising China," *Contemporary Southeast Asia*, Vol. 30, No. 2, 2008, pp. 159–181; Evelyn Goh, "Great Powers and Hierarchical Order in Southeast Asia," *International Security*, Vol. 32, No. 3, 2008, pp. 121–154.
[3] Thomas C. Schelling, *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, 1960, pp. 34–49.

legal resources, moral narratives, and legitimate discourse they can mobilize[1].

Beyond arbitrage within the existing system, the Global South may still exercise meta power over great-power decision-making and systemic operations, even if it lacks the capacity to unilaterally overhaul global institutional structures. Such leverage may be realized through coalitions and blocs, embedded participation within international organizations for norm formulation, and the exploitation of differentiated interest intensities and political leverage on targeted policy domains[2]. International political economy research verifies that international organizations provide platforms for weak-state solidarity and institutional collective action. By forging unified ideologies and worldviews to coordinate joint initiatives, the Global South may reshape global rule-making. The solidarity strategies pursued by Third World nations in the 1970s, alongside the restructuring of North–South agendas driven by the Group of 77 and OPEC, serve as illustrative cases[3]. In essence, relational power operates vertically to generate competitive pressures across multiple AI innovation hubs, while meta power functions horizontally to advance South–South coalitions and collective bargaining. These two approaches reinforce one another and jointly constitute the strategic space available to the Global South within AI data security governance.

2. Specific Paths for Vertical and Horizontal Strategic Interaction

Currently, the global AI landscape features multiple AI innovation hubs rather than a single dominant center, with the United States and China functioning as two primary suppliers of AI technologies. The European Union seeks to secure discursive power over AI governance rules via the "Brussels Effect."[4] This multipolar structure delivers critical strategic opportunities to the Global South. For dominant powers, the coexistence of rival suppliers eliminates the feasibility of imposing unilateral standards on the weaker side, and also makes it difficult to form a united front on pricing and contractual terms. For weaker states, multipolarity enables strategic selection among competing service providers and governance models to generate market pressure as consumers. In other words, vertical strategic interaction is essentially a strategy of exchanging choice for bargaining leverage.

---

[1] Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change," *International Organization*, Vol. 52, No. 4, 1998, pp. 893–903.
[2] Robert O. Keohane, "Lilliputians' Dilemmas: Small States in International Politics," *International Organization*, Vol. 23, No. 2, 1969, pp. 291–310.
[3] Stephen D. Krasner, *Structural Conflict: The Third World Against Global Liberalism*, Berkeley: University of California Press, 1985, pp. 6–60; Robert O. Keohane and Joseph S. Nye, *Power and Interdependence*, 4th ed., Boston: Longman, 2012, pp. 1–33.
[4] Yang Siyuan, "The International Law Implications of the 'Brussels Effect' and China's Responses," *German Studies*, Issue 1, 2026, pp. 94–98.

Specifically, nations of the Global South must avoid overreliance on AI services sourced from a single jurisdiction and instead balance supplies from the United States, China and other economies. As the world's primary supplier of open-source AI, China's open-source models deliver transparency and local deployment advantages, rendering them a vital option for the Global South to diversify supplier exposure and foster market competition. In particular, in areas such as government AI services across the Global South, providing preferential public procurement access to suppliers that diversify AI services can help achieve this objective. Southern nations need not fully adopt any single governance framework, and may selectively adopt institutional components aligned with their developmental stages and institutional capacities. For instance, Chinese frameworks may inform data exit evaluation and tiered supervision, while European Union structures may guide human rights protection standards. Flexible switching among multiple AI suppliers and regulatory frameworks substantially elevates Southern bargaining power.

Vertical strategic interaction relies on horizontal coalitions for support. Isolated weak states lack market scale and negotiation leverage when confronting cross-border platforms and technological powers, yet coalition formation transforms the bargaining dynamic. This coalition logic applies equally to the AI sector, supported by unique material foundations rooted in data factors. The Global South hosts approximately 80 percent of the global population, and future increments of AI training data will originate from Southern territories, which hold exclusive linguistic, cultural and behavioral datasets. Southern nations must recognize their collective bargaining weight as primary suppliers of raw AI training data. Regional AI governance blocs, coordinated positions among developing states within the UN framework, and collective negotiations over data benefit distribution convert fragmented, low-intensity resistance into consolidated influence. The *Global Digital Compact* and *Global AI Governance Initiative* establish practical institutional platforms for such collaboration. The *Global Action Plan on Artificial Intelligence Governance* advocates for "an inclusive, open, sustainable, equitable, secure and reliable digital and intelligent future for all." It prioritizes "the United Nations as the primary channel," targets "bridging digital divides and advancing equitable, inclusive development for developing countries," and promotes "the construction of an inclusive, equitable multilateral global digital governance system grounded in international law, state sovereignty and differentiated developmental realities." The plan also endorses two UN-backed mechanisms: an international AI scientific panel and a global AI governance dialogue. Within such institutional frameworks, the mission of countering AI colonialism and jointly advancing a community with a shared future for mankind in AI and data security domains should serve as the shared

narrative and normative foundation for China and all Global South states to safeguard sovereign independence and pursue equity and justice[1].

3. Institutional Implementation of Asymmetric Games Across Three Dimensions

Stable bargaining outcomes from vertical and horizontal gaming require robust institutional support, structured along three interconnected dimensions: security competition, security baselines and security cooperation.

First is the security competition dimension, which operationalizes vertical market-side leverage by translating supplier selection power into market incentive structures comprehensible and responsive to private actors. If an AI service provider leverages data access restrictions as leverage, Global South states shall pivot toward open-source models and multi-source AI vendors, rewarding actors that uphold data sovereignty even at the cost of marginal reductions in model performance. To this end, the Global South may establish a data sovereignty rating system for AI service providers. Drawing on credit rating methodologies, authorities conduct tiered assessments covering open-source adoption, data transparency, localization compliance and benefit-sharing mechanisms, with rating outcomes tied to market entry eligibility and government procurement preferences. This system generates competitive advantages for sovereignty-respecting vendors while imposing market penalties on non-compliant operators. DeepSeek's terms of use exemplify inclusive benefit-sharing frameworks for the Global South. The document stipulates that users retain any rights they may have in their input content, while the platform assigns to end users any proprietary rights it may have in the generated outputs. It explicitly permits deployment of inputs and outputs across broad use cases including personal applications, academic research, derivative product development and model distillation[2]. The Global South may enact open-source model priority policies. When proprietary AI vendors disregard data sovereignty standards, public resources shall be prioritized for localized fine-tuning and ecosystem development of open-source alternatives. As a global leader in open-source AI and digital infrastructure construction, China must advance institutional frameworks to facilitate open-source AI deployment and infrastructure rollout across the

[1] Zhang Hui, "A Community with a Shared Future for Mankind: Contemporary Advances in the Social Foundation Theory of International Law," *Social Sciences in China*, Issue 5, 2018, pp. 53–68; Wang Shumin, "Bridging the Global Digital Divide: Where International Law Should Proceed," *Journal of Political Science and Law*, Issue 6, 2021, pp. 11–13.

[2] DeepSeek, *DeepSeek Terms of Use*, last updated on 27 March 2026, https://cdn.deepseek.com/policies/en-US/deepseek-terms-of-use.

Global South, advancing intelligent equalization to counter colonial dynamics[1]. Supporting policy instruments include tax incentives for Chinese AI enterprises expanding overseas to deploy open-source models and regional infrastructure, such as additional super-deductions for R&D expenses, VAT exemption, credit and refund and reduced corporate income tax brackets. Dedicated white lists may also be instituted within domestic data compliance and export control regimes.

Second is the security baselines dimension. When market incentives fail to constrain vendor misconduct, institutional safeguards must prevent critical infrastructure and core data from being weaponized through single-supplier dependency. Sovereign control over domestic core AI infrastructure is mandatory, while outsourced AI services must be diversified across multiple suppliers to avoid excessive concentration of exposure to any single vendor. Regulators may adopt large exposure limits modeled on financial supervision frameworks, imposing caps on the proportion of critical national infrastructure services sourced from any individual AI provider. Cross-border data governance must also be extended to AI models themselves, particularly to model parameters that, after domestic training, are transmitted back to Northern jurisdictions through cloud architectures. Northern regulators have already initiated oversight of cross-border model flows. As previously established, AI models embed information extracted from training data within parameters and weight matrices during training cycles. Data governance therefore encompasses both raw training datasets and the AI models themselves. Existing technical research confirms large language models can memorize and reproduce training samples extractable via targeted model queries or adversarial attacks[2]. The European Data Protection Board's Opinion 28/2024 explicitly addresses three core questions: criteria for anonymous AI model classification, legitimate legal bases for model development using personal data, and downstream deployment risks stemming from unlawfully sourced training data. The opinion underscores that personal information within training datasets may become "absorbed" into model parameters, even when models are not intentionally engineered to generate identifiable natural persons[3]. The French National Commission on

---

[1] Liu Yanhong, "Intelligent Equalization Realized through Open-Source AI Legislation," *Journal of Comparative Law*, Issue 1, 2026, pp. 27–29; Yao Xu and Li Chenyao, "Global AI Security Governance: Transition Pathways Catering to Global South Needs," *China Information Security*, Issue 4, 2025, pp. 40–41.

[2] Nicholas Carlini et al., "Extracting Training Data from Large Language Models," *Proceedings of the 30th USENIX Security Symposium*, 2021, pp. 2633–2645; Milad Nasr et al., "Scalable Extraction of Training Data from Aligned, Production Language Models," *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025, pp. 1–10.

[3] European Data Protection Board, *Opinion 28/2024 on Certain Data Protection Aspects*

Informatics and Libertés released parallel guidance in January 2026[1], while U.S. regulatory rules have laid preliminary groundwork for formal cross-border model control frameworks under 28 C.F.R. §202.301(b)(1)[2]. Drawing on these precedents, Global South states may integrate review protocols for residual data and extractable training data risks within existing cross-border data security assessment regimes, preventing exfiltration of national data characteristics, social structural metrics, industrial operational intelligence and strategic resources via model parameter transfers.

Third is the security cooperation dimension. Unified mobilization of collective bargaining power as primary raw data suppliers empowers the Global South to jointly counter AI colonial hegemony under the principles of multilateralism and consultative collaboration, while developing institutional public goods reflecting Southern perspectives on AI data security governance[3]. Regional blocs may formulate shared AI data governance model laws to reduce redundant domestic legislative costs for individual nations. States lacking capacity to operate independent AI technical evaluation systems may establish regional joint technical assessment bodies to capture economies of scale through shared personnel and infrastructure. Bilateral and regional trade agreements may embed capacity-building clauses for AI data security, framing technical assistance and knowledge transfer as reciprocal preconditions for market access[4]. At the international level, Southern states must fully leverage institutional windows created by the *Global Digital Compact* and *Global Action Plan on Artificial Intelligence Governance* to shape global agendas, translating normative consensus into substantive institutional arrangements via regional coordination. Initiatives such as the Belt and Road Initiative advocate balanced global technological progress and promote the establishment of win-win trade regimes among countries, with existing cooperation mechanisms that limit the unequal transfer of excess surplus value across nations[5]. Such cooperation platforms shall actively absorb technical collaboration, digital trade and rule-setting demands arising

---

Related to the Processing of Personal Data in the Context of AI Models, adopted on 17 December 2024, paras.17–40.

[1] Commission Nationale de l'Informatique et des Libertés, "Analysing the Status of an AI Model with Regard to the GDPR," 5 January 2026, https://www.cnil.fr/en/analysing-status-ai-model-regard-gdpr.

[2] 28 C.F.R. §202.301(b)(1).

[3] Liao Fan, "Multilateralism and International Rule of Law," *Social Sciences in China*, Issue 8, 2023, pp. 75–77.

[4] Susan Aaronson, "Data Is Different, and That's Why the World Needs a New Approach to Governing Cross-Border Data Flows," *Digital Policy*, Regulation and Governance, Vol. 21, No. 5, 2019, pp. 441–460.

[5] Ma Yan, Li Jun and Wang Lin, "The Anti-Inequality Nature of the Belt and Road Initiative: A Rebuttal to the New Colonialism Allegations," *World Economy*, Issue 2, 2020, pp. 1–20.

from AI data security governance.

IV. Conclusion

Data security governance within the AI era extends beyond technical regulation, touching upon the global order of value distribution and the integrity of state sovereignty. Massive data demand for large model training, paired with capability advancement driven by scaling laws, transforms data colonialism into AI colonialism of higher intensity and more concealed forms under the AI technological paradigm. The Global South falls into an accelerating loop marked by widening the price scissors between Southern raw data and Northern AI outputs between its supply of raw training data and consumption of AI services. When the international environment shifts from peace to confrontation, it may also face direct threats to critical infrastructure and sovereign integrity from AI-enabled cyberattacks and precision strikes supported by multi-source intelligence fusion. Nevertheless, asymmetric interaction theory identifies dual strategic avenues for disadvantaged actors via relational power and meta power. Vertically, the coexistence of multiple AI innovation hubs creates strategic leeway for the Global South to gain bargaining power through selective service adoption and rule reference. Horizontally, the collective identity as primary suppliers of AI raw data alongside institutional frameworks for South–South cooperation supplies normative resources for the remaking of international governance rules. Realization of such strategic room relies on supporting institutional construction. The security competition dimension demands incentives for service providers that uphold data sovereignty and prioritized development of open-source models. The security baseline dimension mandates caps on exposure to single service suppliers for critical infrastructure, and expands regulatory coverage from raw data to model parameters. The security cooperation dimension consolidates collective bargaining power through institutions including regional model laws and joint assessment bodies, while institutional windows such as the *Global Digital Compact*, the *Global Action Plan on Artificial Intelligence Governance*, and the *Belt and Road Initiative* deliver rare opportunities to the Global South. Amid continuous restructuring of global power by AI, only Global South legal scholarship equipped with theoretical self-awareness and institutional imagination can contribute its due intellectual power to breaking the core-periphery digital dependency structure and building a more just and equitable global AI governance order.

The Authentication and Governance of AI Generated Content: A Global Perspective

*Li Wenlong Luo Jiakun**

Abstract As generative AI reshapes the global information environment, AI-generated content authentication has moved from a transparency device to a broader infrastructure for trust, traceability and responsibility. This article unpacks the technical foundations of content labelling and examines their limits in real-world circulation. It shows that labelling systems can easily break down when content moves across platforms, is re-edited, generated through open-source models, or subjected to privacy and verification constraints. The article then maps the emerging regulatory landscape across China, the European Union, the United States, South Korea, Vietnam, Brazil and Japan. It shows that these regimes are working from different assumptions about informational risk, platform responsibility and the proper place of technical governance, and a genuinely interoperable global labelling framework is unlikely to come together easily. The article further contends that a range of factors—industrial and policy divisions, weak China–West dialogue, the limited convening power of international organisations, and the shrinking reach of the Brussels Effect—all hold back meaningful coordination. China could potentially move beyond rule-taking and play a more active role in shaping global AI content governance by building on its domestic standards, plugging them into multilateral forums, scaling them through major platforms, and pursuing mutual recognition through regional and international standard-setting processes.

Keywords AI-generated Content; Digital Watermarking; Content Provenance; Global AI Governance; AI Authenticity

* Li Wenlong, Research Fellow at Guanghua Law School and Institute of Digital Rule of Law, Zhejiang University; Luo Jiakun, Research Fellow of the AI and Law Programme, Faculty of Law, the University of Hong Kong

I. Statement of the Research Problem

Human information ecosystems are undergoing an unprecedented transformation driven by advances in artificial intelligence capabilities. Synthetic texts, images, audio and video materials have exhibited exponential growth across all platforms in recent years. According to the *2026 AI Index Report*, over half of newly released online content after 2025 originates directly or indirectly from AI systems. News articles, short videos, images and comment section dialogues encountered in daily life are largely algorithm-generated[1]. Media outlets and academic circles have issued repeated warnings that AI collaborative outputs will surpass all human creative works in volume around 2026 and become the primary source of information production. The Nieman Lab has discussed this trend and pointed out that the traditional human-centred communication order is being reshaped, with sharply blurred boundaries between human-authored and machine-generated information within digital spaces[2].

To a certain extent, the realism of AI-generated content has crossed classic thresholds once used to distinguish human creation from machine output. The Turing Test carries numerous methodological and philosophical controversies[3]. Within an environment where large language models automate most interactive exchanges, ordinary users without professional tools can no longer reliably differentiate authentic content from synthetic alternatives. More problematic outcomes emerge when synthetic texts, images and videos infiltrate official documents, judicial filings and government routine paperwork through daily circulation channels. Verification mechanisms reliant on human visual inspection and professional intuition become extremely fragile. Recent media disclosures reveal that multiple basic-level courts cited a non-existent statutory provision within official announcements. The error stemmed from staff copying unvalidated online templates during drafting, with corrections only implemented following reminders from superior courts[4]. Misstatements of fundamental legal provisions by judicial institutions obligated to understand and apply law inflict severe damage on judicial credibility, administrative legitimacy and the overall rule-of-law order.

---

[1] Sha Sajadieh et al., "The AI Index 2026 Annual Report," Institute for Human-Centered AI, Stanford University, April 2026, https://hai.stanford.edu/ai-index/2026-ai-index-report.

[2] Davey Alba, "Predictions for Journalism 2026," NiemanLab, https://www.niemanlab.org/2025/12/in-2026-ai-will-outwrite-humans/.

[3] A. Pinar Saygin, I. Cicekli and V. Akman, "Turing Test: 50 Years Later," *Minds and Machines*, Vol. 10, pp. 463–493.

[4] Wang Hui and Zhang Pan, "Why Multiple Courts Cite a Non-existent Statute," Qiushi Theory Network, 24 April 2026, https://www.qstheory.cn/20260424/a1d7ca7fb6dc443080a2c80c0a6390d6/c.html.

From a macro analytical perspective, the technological pursuit of Turing-style indistinguishability stands in mirror-image opposition to contemporary governance rationales. For decades, AI research guided by the Turing paradigm prioritised machine production of human-like or superhuman outputs across linguistic and visual domains as core markers of technological progress. Regulatory and governance frameworks, by contrast, treat this identical simulation capacity as a latent source of informational confusion and rely on institutional and instrumental design to rebuild human discriminatory capacity. The more successfully technology mimics humans, the more governance needs to render machine-generated traces visible. Against this backdrop, labelling for AI-generated content emerges as a core instrument for order reconstruction within AI-infused information ecosystems.

This paper defines AI-generated content labelling as the embedding or attachment either in the AI-generated content itself or in its accompanying metadata, of critical information such as generation pathways, model origins and revision chains, in a format that combines machine-readability with human perceptibility. Such embedded data enables downstream platforms, regulators and ordinary users to identify synthetic materials and conduct reasonable evaluative judgements[1]. AI-generated content labelling carries inherent cross-border and global attributes. Short videos and social media materials circulate across platforms, cultures and linguistic boundaries by nature. Chinese platforms including Douyin, WeChat and Xiaohongshu regularly redistribute reprocessed content originating from overseas services such as YouTube and TikTok. Conversely, synthetic videos and images produced by Chinese creators continuously enter foreign platforms as components of cultural outreach initiatives[2]. EU platforms frequently fail to recognise legitimate labels embedded by Chinese or non-EU tools, and the reverse scenario also holds. This mismatch generates mass unmarked or misclassified synthetic content in cross-border circulation[3]. On the other hand, although there is still a certain degree of national specificity at the model level, with China currently relying primarily on domestically developed large models, the country is also actively promoting the overseas expansion of related technologies. At the same time, it is not impossible that models from

[1] Olivia Burrus, Amanda Curtis and Laura Herman, "Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content," *Interactions*, Vol. 31, No. 4, p. 38.
[2] Jufang Wang, "'Is this AI Generated?' The Challenges of Regulating and Implementing AI-Generated Content Labelling in the EU and China," *Oxford Global Society*, 18 April 2026, https://oxgs.org/2026/04/18/is-this-ai-generated-the-challenges-of-regulating-and-implementing-ai-generated-content-labelling-in-the-eu-and-china/.
[3] Martina J. Block, "Transparency in human-AI interaction—An analysis of Article 50(1) AI Act," *Computer Law & Security Review*, Vol. 61, p. 106327.

regions such as the EU and US may enter the Chinese market on a larger scale in the future. This highly interconnected technical and content ecosystem faces systemic risks if jurisdictions adopt incompatible technical pathways and legal obligations for AI labelling. Cross-border content identification and liability tracing become unworkable, imposing excessive adaptation burdens on platforms and developers confronted with fragmented jurisdictional rules. Widespread open-source model deployment further complicates unified pre-output labelling and raises overall governance uncertainty.

Regulators across multiple jurisdictions have acknowledged the importance of AI content labelling and developed distinct legislative frameworks and technical tools within domestic laws and policies. The global regulatory landscape nevertheless remains highly fragmented, lacking consensus-based interoperable universal frameworks for implementation. China has constructed a supervisory framework for AI-generated content labelling following the release of the *Provisions on the Administration of Deep Synthesis in Internet Information Services* in 2022[1], and after the *Measures for the Labelling of AI-Generated Synthetic Content*[2] and the supporting national standard GB 45438-2025[3] in 2025, China became the world's first jurisdiction to impose mandatory labelling requirements for AI-generated synthetic outputs. The United States adopted soft incentives for watermarking and labelling via executive orders during the Biden administration, while state-level statutes introduced disclosure mandates for political advertisements and commercial promotions. California will launch "pilot" compliance rules for AI content labelling starting August 2026[4]. Although the European Union pioneered comprehensive AI regulatory frameworks with the initial 2021 draft of the AI Act containing transparency provisions[5], but the specific compliance requirements and standards have mainly been advanced through EU codes of practice, beginning to take shape

[1] Cyberspace Administration of China et al., *Provisions on the Administration of Deep Synthesis in Internet Information Services*, issued 25 November 2022; see also Wenlong Li, "Understanding China's AI-generated Content Labelling Mandate towards a New Mode of Algorithmic Stewardship," *The Paris Conference on AI & Digital Ethics*, 2026, https://paris-conference.com/event/understanding-chinas-ai-generated-content-labelling-mandate-towards-a-new-mode-of-algorithmic-stewardship/ (last visited on 1 May 2026); See also Nicolaj Feltes, "Article 50 AI Act: Do the Transparency Provisions Improve Upon the Commission's Draft?" *JIPITEC*, Vol. 16, No. 2, p. 222.

[2] Cyberspace Administration of China et al., *Measures for the Labelling of AI-Generated Synthetic Content*, Document No. CAC〔2025〕2.

[3] State Administration for Market Regulation and Standardization Administration of China, GB 45438-2025 Cybersecurity Technology—Specification for Labelling AI-Generated Synthetic Content.

[4] California AI Transparency Act, Senate Bill 942 (SB-942).

[5] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence.

in early 2026. Separate differentiated labelling regimes have also developed in South Korea[1], Vietnam[2], Brazil[3] and Japan[4]. Such multi-centred, divergent institutional development demonstrates regulatory innovation vitality yet creates technical and legal incompatibilities for cross-border content, with international coordination lagging far behind technological diffusion speeds.

Table 1 Current Legislative Frameworks for AI-Generated Content Labelling in Major Jurisdictions

Jurisdiction	Legislative / Policy Document	Legislative Status	Effective Date	Core Features
China	*Measures for the Labelling of AI-Generated Synthetic Content*	In force	1 September 2025	Full-chain liability; dual-layer human-perceptible and machine-readable labelling
European Union	Article 50 of the EU AI Act and Draft Second Version of the Code of Practice	Published, not yet effective	August 2026	Dual liability of providers and deployers; multi-tier labelling; proportionality principle; artistic exemption
United States (Federal)	White House Executive Order on Safe, Secure and Trustworthy Development and Use of Artificial Intelligence	Revoked (formally repealed by Executive Order 14148 issued by President Trump on 20 January 2025)	Signed 30 October 2023, terminated 20 January 2025	Mandatory watermark embedding for AI outputs produced by federal agencies; no direct binding force on private operators
United States (California)	California Senate Bill 942 / Assembly Bill 853	Published, not yet effective	2 August 2026	Mandatory obligations for large platforms; dual visible and invisible labelling
United States (New York)	New York Senate Bill 8420-A / Assembly Bill 8887-B	Published, not yet effective	9 June 2026	Disclosure requirement for synthetic performers in advertisements
South Korea	AI Basic Act	In force	22 January 2026	General disclosure obligation; exemption for artistic works
Vietnam	Law on Artificial Intelligence	In force	1 March 2026	Dual liability of content providers and implementers
Brazil	AI Legislative Bill No. 2338/2023	Pending final Chamber vote	—	Scenario-based labelling; protection mechanisms for vulnerable groups under PBIA

[1] Republic of Korea, *Framework Act on the Development of Artificial Intelligence and Establishment of Trust* (AI Basic Act), Act No. 20676, adopted 26 December 2024, promulgated 21 January 2025, effective 22 January 2026.
[2] Socialist Republic of Vietnam, *Law on Artificial Intelligence*, Law No. 134/2025/QH15, adopted 10 December 2025, effective 1 March 2026.
[3] Brazil, Bill No. 2338/2023 (*Marco Legal da Inteligência Artificial*), submitted 3 May 2023, Senate approval 10 December 2024, pending Chamber final vote, Article 7.
[4] Japan, *Act on Promoting R&D and Application of AI-Related Technologies* (Act No. 53 of Reiwa 7), promulgated 4 June 2025, effective 1 September 2025.

Jurisdiction	Legislative / Policy Document	Legislative Status	Effective Date	Core Features
Japan	Act on Promoting R&D and Application of AI-Related Technologies	Fully effective	1 September 2025	Voluntary transparency principle

Notes:

1. No confirmed effective date has been specified for the Brazilian bill, marked as "—" within the table.

2. All legislative status assessments take April 2026 as the benchmark timeline.

II. Technical Principles and Practical Dilemmas of AI Labelling

Prior to discussions on AI content labelling frameworks, clear definitions of watermarking and associated concepts are required. The term watermark is widely used in loose discourse and frequently conflated with metadata-based labelling. Strictly speaking, metadata labelling does not embed itself into the pixel or signal structure of the content itself, but rather attaches to the file structure externally, thus constituting an independent approach within covert labelling. The watermarks examined in this paper refer to technical mechanisms embedded inside AI-generated content for identification and traceability, distinguishable from "model watermarks" for intellectual property protection and "data watermarks" designed for training data compliance and ownership confirmation.

Watermarking technology can be split into overt and covert categories. Traditional "overt labels" overlay text, logos or graphic marks onto images and videos for direct human perception, serving copyright notification and anti-removal functions while remaining vulnerable to deliberate editing and erasure. By contrast, "covert labels" encode subtle patterned signals within pixel distributions, frequency domains or generation probability spaces invisible to human observers, identifiable only via dedicated detection algorithms with inherent margins of detection error.

Table 2 Comparison Between Overt and Covert Labels

Dimension	Overt Labels	Covert Labels
Display Position	Visible markers placed appropriately within content	Model IDs, generation timestamps stored in file metadata
Target Audience	General human users	Platforms, regulatory authorities, and forensic agencies
Core Functions	Risk alert and confusion prevention, safeguarding public	Traceability and forensic evidence collection, compliance

Dimension	Overt Labels	Covert Labels
	right to know	auditing
Perceptibility	Instantly recognizable to human eyes or ears	Decoding via dedicated tools required, no disruption to user experience
Resistance to Erasure	Partially removable through screenshots and cropping	Retain integrity across editing if properly engineered

From a technical developmental perspective, covert labelling originates from "steganography," an ancient information concealment technique[1]. The core logic of steganography lies in hiding the very existence of transmitted information rather than encrypting its content to realise covert communication. Both "concealing writing on wax tablets" and "tattooing characters on scalps," as documented by the ancient Greek historian Herodotus, were methods of embedding information beneath visible surfaces to achieve covert communication. In the digital era, "digital watermarking" constitutes standardised steganographic practice with a functional shift from concealment to traceability, rendering it a vital technical instrument for transparency and liability allocation within generative AI ecosystems. Digital watermarking is not a single algorithm but an integrated technical spectrum including "statistical watermarking" and machine learning-based variants. A Brookings Institute report distinguishes two high-potential watermarking approaches. "Statistical watermarking" introduces systematic bias into sampling distributions during generation. For instance, the "red-green vocabulary list" mechanism biases large language model sampling toward preselected authorised word sets, creating statistical fingerprints identifiable exclusively by detection algorithms within lengthy texts. Machine learning-based watermarking leverages neural networks to embed latent perceptual patterns into images, audio and video outputs, maintaining natural visual and auditory appearances while embedding machine-decodable fingerprints within latent model spaces[2].

1. Technical Pathways

Four mainstream technical pathways enable identification and labelling of AI-generated content[3]. First is "post-hoc pattern recognition." This

[1] Ingemar J. Cox et al., *Digital Watermarking and Steganography*, 2nd Edition, Morgan Kaufmann, 2008, pp. 61–103.
[2] Siddarth Srinivasan, "Detecting AI Fingerprints: A Guide to Watermarking and Beyond," *Brookings*, 4 January 2024, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.
[3] Siddarth Srinivasan, "Detecting AI Fingerprints: A Guide to Watermarking and Beyond,"

approach deploys machine learning models to capture subtle statistical disparities between machine and human-authored text, images, audio and video after content generation, outputting probabilistic judgements on synthetic origins. Representative tools include OpenAI's retired AI Classifier and GPTZero. This pathway demands no modifications to generative models and delivers universal applicability for any suspicious content. Reliance on output statistical signatures creates critical vulnerabilities: minor rewriting, translation, sentence rearrangement, image cropping or filter application may erase or attenuate statistical traces, damaging overall accuracy and robustness. Second is retrieval via dedicated content databases. Model developers systematically archive all generated text, imagery and audiovisual material within proprietary repositories. Semantic retrieval or similarity matching algorithms determine whether the submitted content significantly overlaps with or is semantically similar to any AI-generated sample in the database. Database retrieval outperforms post-hoc classifiers for unaltered or lightly revised textual content. Analogous "hash database" matching frameworks already operate across non-AIGC scenarios, such as the Global Internet Forum to Counter Terrorism (GIFCT) terrorist content database[1], StopNCII non-consensual intimate image matching systems[2], and Microsoft PhotoDNA for child exploitation material detection[3]. These deployments supply actionable precedents for extending retrieval architectures to AI content governance. The core prerequisite for this mechanism is sustained, comprehensive storage of all generative outputs, imposing massive operational costs and triggering privacy risks around personal data processing and data minimisation obligations. Third is "metadata-based labelling," which attaches a structured set of "identity credentials" to content files or transmission protocols, recording information such as generation time, generating entity, model and parameters used, and content serial number. Unlike the two reverse-inference pathways above, metadata labelling relies on proactive marking by generators and full-chain security maintenance to deliver high interpretability and verifiability. China's *Measures for the Labelling of AI-Generated Synthetic Content* and the Coalition for Content Provenance and Authenticity (C2PA) technical alliance[4] both prioritise

Brookings, 4 January 2024, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/.

[1] Global Internet Forum to Counter Terrorism, https://gifct.org/ (last visited on 1 May 2026).

[2] StopNCII, https://stopncii.org/ (last visited on 1 May 2026).

[3] Microsoft PhotoDNA, https://www.microsoft.com/en-us/photodna (last visited on 1 May 2026).

[4] Coalition for Content Provenance and Authenticity, "The C2PA Standard: History, Promises and Structural Limitations," TrueScreen, https://truescreen.io/articles/c2pa-standard-history-limitations/.

metadata embedding as the primary covert labelling solution, as long as the labels are authentic and the full chain of custody remains intact, the identification results have a high degree of certainty and auditability. Standardised universal fields also facilitate cross-platform mutual recognition. Limitations include susceptibility to loss during transcoding, screenshotting, or platform migration, alongside heavy dependence on voluntary compliance by upstream developers. Widespread open-source local deployment undermines the capacity of single-jurisdiction regulation to cover all generative activities. Fourth is "embedded digital watermarking." Systematic perturbations to pixel layouts, frequency signals and text sampling probabilities during generation embed identifiable latent features while preserving natural perceptual quality. Embedded traces deliver superior cross-platform robustness relative to standalone metadata yet it is often unrealistic to apply the same detection logic to text and multimedia content; differentiated approaches and reasonable expectations are needed for different media.

2. Practical Challenges

Each technical pathway exhibits distinct strengths in accuracy, interpretability and robustness with partial complementary effects, yet all confront systemic real-world barriers determined by cross-platform compatibility, developer compliance, open-source loophole risks, privacy constraints and the maturity of trusted third-party infrastructure[1]. First, cross-media cross-platform transmission complexity. AI-generated content circulates across platforms with repeated reposting, editing, transcoding, screenshotting, splicing and offline screen recording. Labelling mechanisms limited to single platforms or specific file formats lose most governance value and in actual distribution chains and cannot function as universal digital infrastructure. Recent solutions including Seedance 2.0[2] and GCmark[3] claim extractable watermark information survives multiple editing iterations. Fragmentation persists without unified embedding and detection interfaces across vendors, creating siloed ecosystems where labels recognised by one platform remain undetectable to others. Some scholars even bluntly state that

[1] Jon Dykstra, "AI Watermark Rules Diverge Despite Efforts to Create Global Standards," *Jurvantis.AI*, 26 November 2025, https://jurvantis.ai/ai-watermark-rules-diverge-despite-efforts-to-create-global-standards/.
[2] Volcengine AIGC Lab, "An Analysis of Seedance 2.0 Watermark Technology and Compliant Removal Solutions," *Volcengine*, 13 April 2026, https://www.volcengine.com/article/41913.
[3] Zhejiang University College of Cyber Security, "Release of GCmark Multi-Modal Synthetic Content Security Label Platform," *Zhejiang University*, 22 April 2024, https://icsr.zju.edu.cn/2024/0422/c70144a2906086/page.htm.

watermarking without standards is not AI governance, pointing out that in the absence of unified norms, watermarks serve more as decoration than as genuine governance tools[1]. Second, heavy reliance on voluntary cooperation from model developers and service providers. Metadata architectures demand automatic structured metadata writing during output and sustained integrity across distribution pipelines, imposing stringent engineering, cybersecurity and compliance requirements. Only large resource-rich platforms sensitive to regulatory oversight consistently implement mandatory labelling. Small-scale independent developers frequently operate without labelling due to limited technical capacity and weak compliance incentives. Third, the open-source loophole effect amplifies governance failures. Mandatory labelling rules confined to closed-source proprietary models drive malicious actors toward freely configurable open-source alternatives functioning as regulatory safe harbours. Regulators may enforce metadata and watermark embedding as market access prerequisites for closed-source large models via administrative supervision, industrial norms and market entry thresholds. Local self-hosted open-source instances evade uniform oversight entirely. Even if the upstream open-source model has built-in metadata or watermarking mechanisms by default, downstream developers can still disable them by modifying the code, or bypass the original labels through retraining or fine-tuning. Fourth, inherent privacy trade-offs in embedded traceability. Both metadata and embedded watermarks attach supplementary data to content. If the additional information is limited to the model and platform level, such as recording the model type, service provider name, content serial number and signature, the governance value is relatively high while privacy risks remain manageable. Granular tags recording specific users, geographic coordinates and device fingerprints risk transforming labelling systems into pervasive cross-platform tracking tools. Fifth, the absence of credible third-party verification infrastructure. Without credible, transparent and accountable third-party mechanisms, no matter how sophisticated the identification technology itself is, it may be reduced to a symbolic compliance tool. Many metadata and watermark schemes implicitly rely on some form of neutral arbiter in the verification process, such as certificate authorities, unified signature verification services and cross-platform detection portals. This issue directly foreshadows the key challenge to be discussed in the next chapter: in building global AI labelling obligations and technical standards, how to allocate infrastructure governance authority among state regulators, platform self-regulation and multilateral standard-setting bodies in a way that avoids both fragmentation and the emergence of new technological hegemony.

---

[1] Alexander Nemecek, Yuzhou Jiang and Erman Ayday, "Watermarking Without Standards Is Not AI Governance," *arXiv:2505.23814*, 2026, p. 23814.

III. A Review of Global Legislative and Practical Frameworks for AI Labelling

Institutional construction around labelling rules for generative AI content remains in rapid development, without unified normative frameworks. This section analyses structural divergences and corresponding governance logics behind such regimes across four dimensions: obligation thresholds, labelling formats, normative adjustment tools and liability architectures.

1. Triggering Thresholds for Labelling Obligations

Jurisdictions adopt regulatory approaches ranging from comprehensive prevention to scenario-specific restrictions when defining the conditions triggering mandatory labelling, a divide rooted in divergent understandings over whether AI-generated informational risks constitute inherent or context-specific hazards[1].

China adopts a comprehensive prevention model by framing confusion risks stemming from AI outputs as inherent hazards. Its Labelling Measures take "potential public confusion" as the core triggering standard, making labelling mandatory for all AI-generated content that carries risks of misleading the public[2]. The European Union implements a scenario-restricted model. Article 50 of the AI Act sets separate transparency obligations for two categories of actors. Providers must embed machine-readable markers in all AI outputs as a universal technical requirement, while deployers bear the duty of visible disclosure to end users only for deepfakes or texts involving public interests. Such tiered design reflects the EU's stance on the universal necessity of technical traceability alongside targeted protection of the public's right to know under high-risk scenarios. The United States splits AI labelling into discrete policy domains including federal accountability, platform obligations, consumer protection and electoral security, rather than adopting unified cross-cutting technical regulation. South Korea and Vietnam apply a hybrid model combining scenario triggers and confusion risk tests, while Brazil concentrates labelling mandates on high-risk interactive scenarios such as emotion recognition and biometric classification targeting vulnerable groups.

---

[1] Mei Xiaying, "Fundamental Issues in AI Legislation: Algorithmic Regulation and Human-Machine Relations in Specific Scenarios," *The Jurist*, No. 6, 2025, p. 16.

[2] Liang Zheng and Tian Guiping, "People-Centered, Agile Coordination: The Concept and Practice of China's AI Labelling Scheme," Center for AI Governance, Tsinghua University, January 2026, https://aiig.tsinghua.edu.cn/local/B/9C/B9/6963FE0068377D36D0AD067431B_827C45FA_122B7F.pdf.

2. Labelling Formats

Human-perceptible labels deliver direct notifications to end users regarding AI generation origins. China mandates visible markers, including text, audio prompts and graphic symbols, for all materials likely to confuse the public, with clear presentation embedded within outputs. The draft EU Code of Practice proposes a unified AI icon required to appear upon initial user contact. South Korea imposes comparable disclosure duties without specifying formats, while Vietnam allows flexible "appropriate marking" for creative works. Brazil mandates "easily identifiable icons or symbols."

Machine-readable labels underpin technical traceability and detection, a prerequisite for elevating labelling from simple informational prompts to functional governance instruments[1]. China adopts metadata embedding as the core covert labelling method, with supporting national standards detailing technical specifications. The draft EU Code of Practice establishes a sophisticated multi-layer marking architecture. The first layer consists of embedded metadata and digital signatures; the second layer features interleaved watermarks directly inscribed into content pixels, resistant to compression and cropping; the third layer relies on fingerprint identification and log retrieval for source verification where metadata and watermarks are stripped. The EU also requires universal detection tools covering both marked and unmarked synthetic content. Vietnam obliges AI outputs to carry machine-readable tags in formats stipulated by state authorities.

3. Normative Adjustment via Exemptions and Proportionality

Multiple jurisdictions deploy exemption clauses and the proportionality principle to balance labelling obligations against competing values including freedom of expression, industrial innovation and vulnerable group protection. Artistic exemptions represent a widely adopted adjustment mechanism. China's Labelling Measures are relatively limited in this regard and have not yet established dedicated provisions for artistic works. The draft EU Code of Practice permits "non-intrusive" markers for artistic and satirical works without disrupting exhibition or appreciation. Paragraph 3 of Article 31 of South Korea's AI Basic Act explicitly provides that where AI-generated content constitutes or forms part of artistic or creative works, the disclosure methods "shall not interfere with the display or appreciation" of such works.

---

[1] China Academy of Information and Communications Technology, *2025 Research Report on AI Governance*, January 2026, https://www.caict.ac.cn/kxyj/qwfb/ztbg/202602/P020260213607027089305.pdf; Zeng Xiping, "AI Labelling as a Core Governance Tool for AIGC," *People's Posts and Telecommunications News*, 26 September 2024, p. 3.

Vietnam's AI Law requires "appropriate marking" that avoids impairing the presentation of films and artworks. Brazil's rules omit artistic exemptions while establishing enhanced safeguards for vulnerable populations. New York's statute exempts advertisements featuring synthetic performers where such figures appear consistently within creative works. Brazilian regulations impose tailored design requirements, including clear labelling, on AI systems serving children, adolescents, the elderly and persons with disabilities to guarantee accessibility. The draft EU Code of Practice similarly mandates perceptible markers for disabled users via alternative text, audio prompts and high visual contrast.

4. Liability Architectures

Liability allocation constitutes the foundational functional objective of AI content labelling regimes. China emphasizes a full-chain liability framework covering service providers, distribution platforms and end users. Providers must embed markers at generation; platforms conduct compliance inspections on circulated materials; individual users bear limited disclosure duties when deploying AI tools. Full-chain coverage eliminates liability loopholes yet raises overall compliance costs and limits replicability. The EU establishes a dual-liable subject mechanism split between providers and deployers. Providers must implement technical labelling for all AI-generated content and explicitly prohibit the removal or alteration of markers through contractual terms, while deployers assume visible disclosure duties solely for deepfake and public-interest content. The EU further recommends embedded structural markers during training for open-weight models to facilitate downstream compliance. U.S. liability rules remain fragmented and jurisdiction-specific. California Senate Bill 942 targets large generative AI platforms, while New York's legislation regulates advertisement creators. Federal executive orders primarily bind the AI usage of federal government agencies, reflecting U.S. reliance on industry self-regulation and state-level pilot rules with inconsistent cross-state liability standards. South Korea defines liable parties as all operators offering high-impact or generative AI products and services. Vietnam distinguishes content providers responsible for technical labelling from implementers obligated to notify the public during creation and editing. Brazil assigns sole liability to corporate entities without further subdivision. Japan merely sets forth general transparency duties with unspecified corresponding liability rules, consistent with its voluntary regulatory orientation.

IV. Difficulties and Barriers to the Formation of an International Framework

Against the current geopolitical and industrial landscape, forming truly binding and interoperable joint efforts at the international level faces formidable obstacles. The following section turns to the international governance level and attempts to analyze, from four aspects, the structural challenges facing the current effort to build a global AI labelling framework. This section elaborates four systemic hurdles hindering unified international rule-making for AI content labelling.

1. Dual Divisions within Western Industry and Policy Circles

Even among states conventionally regarded as a unified bloc, European and American parties hold divergent governance approaches to AI labelling, with clear rifts across industrial and legislative domains.

From an industrial perspective, Silicon Valley technology enterprises maintain sharp disagreements over watermarking and labelling mechanisms. Corporations including Adobe, Google, and Sony participate in the C2PA and the Content Authenticity Initiative (CAI) to promote metadata and digital signature-based content credential systems[1]. Nevertheless, some key platforms and hardware manufacturers choose to keep their distance or conduct only limited trials. For example, Apple has long refrained from formally joining the C2PA alliance, with its support at the device and system level lagging behind[2], while X, after launching its proprietary Grok model, withdrew from several self-regulatory authenticity initiatives[3]. Strategic fragmentation among leading enterprises erodes prospects for industry-driven unified labelling infrastructure.

On the legal and policy front, U.S. federal governance of AI labelling relies predominantly on soft law and executive guidance. The executive order issued under the Biden administration merely encouraged watermarking for

[1] Sofia Yan, "Content Authenticity Initiative Part I: Conception and Advocacy," Medium, 25 February 2021, https://medium.com/numbers-%E4%B8%BB%E5%BC%B5%E6%95%B8%E6%93%9A/cai-content-authenticity-initiative-%E5%85%A7%E5%AE%B9%E7%9C%9F%E5%AE%9F%E6%80%A7%E8%A8%88%E7%95%AB-%E4%B8%8A-%E5%8B%95%E6%A9%9F%E8%88%87%E7%B7%A3%E8%B5%B7-ec0effaadf00.

[2] Blake Montgomery, "Apple Delays Launch of AI-powered Features in Europe, Blaming EU Rules," *The Guardian*, 21 June 2024, https://www.theguardian.com/technology/article/2024/jun/21/apple-ai-europe-regulation.

[3] Engadget, "X Won't Train Grok on EU Users' Public Posts," 4 September 2024, https://www.engadget.com/big-tech/x-wont-train-grok-on-eu-users-public-posts-155438606.html (last visited on 25 April 2026).

federal AI outputs without establishing mandatory federal legislation[1]. Only a small number of state statutes impose labelling and disclosure requirements for narrow scenarios such as political advertisements and deceptive commercial promotions, creating a fragmented, scenario-based regulatory landscape. Under the second Trump administration, AI labelling has slipped down the federal policy agenda, further weakening U.S. capacity and willingness to export unified rules. The EU institutionalized transparency obligations for generative AI and deepfakes under Article 50 of the AI Act and initiated drafting work for supporting codes of practice in 2025 to deliver operational guidance via soft law tools.

At the standard-setting level, C2PA was once expected to transition from technical specifications to formal international standards around 2025[2]. Its 2026 State of Content Authenticity Report, however, shifts focus to internal compliance schemes and member ecosystem expansion instead of near-term timelines for formal international standardization[3]. Multiple analyses confirm that while C2PA specifications serve as de facto references for numerous institutions, their penetration rate remains low across the Global South and open-source communities[4].

2. Insufficient China-West Dialogue Mechanisms

Separated from specific technical pathways, communication channels and alignment incentives between China and Western jurisdictions on AI labelling remain extremely limited. Historically, globally consensus-based rules usually originate from institutional frameworks forged within Western regions before spreading via bilateral and multilateral channels to form universal standards. The AI labelling field reverses this trajectory. Despite China's early and sophisticated legislative design for mandatory AI content labelling, Western jurisdictions lack mature docking mechanisms, leading to parallel rule-making on both sides with limited cross-system communication.

Weak dialogue also manifests in Track II and multi-stakeholder forums. On platforms dominated by Western bodies including UNESCO, the OECD, and G7 summits, Chinese delegates possess restricted speaking space, with

---

[1] "Voluntary Commitments from Leading Artificial Intelligence Companies on July 21, 2023," *Harvard Law Review*, Vol. 137, p. 1284.
[2] "The C2PA and Embassy of France Collaborate to Advance Authenticity and Transparency in Digital Content," *Coalition for Content Provenance and Authenticity*, 2 October 2024, https://spec.c2pa.org/post/embasssyoffrance_pr/.
[3] Andy Parsons, "The State of Content Authenticity in 2026," *Content Authenticity Initiative*, 18 January 2026, https://contentauthenticity.org/blog/the-state-of-content-authenticity-in-2026.
[4] "C2PA Standard: History, Promises and Structural Limitations," *TrueScreen*, https://truescreen.io/articles/c2pa-standard-history-limitations/.

most discussions framing Chinese governance models as external risks rather than engaging in substantive technical and institutional exchanges[1]. Major China-hosted summits such as the World Artificial Intelligence Conference (WAIC) and the World Internet Conference prioritize domestic governance experience demonstrations, while Western participants mostly attend at academic and corporate levels without systematic institutional coordination[2]. Agenda-setting further marginalizes AI content labelling, which is treated as a minor subtopic under information integrity and electoral interference rather than an independent governance pillar alongside AI safety and foundation model risk assessment. This two-way parallel communication pattern impedes complementary progress on shared standards and cross-border verification infrastructure.

3. Diminished Influence of Traditional International Organizations

Traditional bodies mandated with cross-jurisdictional standard coordination, including the EU, UNESCO, the G7, and the International Telecommunication Union (ITU), face shrinking mobilizing capacity amid fragmented geopolitics.

The UN and its specialized agencies retain partial leadership on AI ethical frameworks, exemplified by the UNESCO *Recommendation on the Ethics of Artificial Intelligence*[3]. However, when it comes to specific and actionable technical standards and mandatory rules, these organizations often face deep divergences among member states, making it difficult to reach consensus on core issues such as the scope of labelling obligations, technical baselines, and enforcement mechanisms, thus hindering substantial progress.

---

[1] Fu Ying, "Remarks at the Paris AI Action Summit," Center for International Security and Strategy, Tsinghua University, 21 February 2025, https://ciss.tsinghua.edu.cn/info/zlyaq/7957; AI Data Press-News Team, "When Global AI Governance Stalls, Scientists and Civil Society Take the Lead," *AI Data Press*, 10 March 2026, https://www.aidatapress.com/news/ai-governance-fragmentation-fadi-daou-globethics.

[2] Elonnai Hickok et al., "Multistakeholder Promises and Power Gaps in Global AI Summits," *Tech Policy Press*, 17 March 2026, https://www.techpolicy.press/multistakeholder-promises-and-power-gaps-in-global-ai-summits/; Gregory C. Allen and Issac Goldston, "Understanding U.S. Allies' Current Legal Authority to Implement AI and Semiconductor Export Controls," *CSIS*, 14 March 2025, https://www.csis.org/analysis/understanding-us-allies-current-legal-authority-implement-ai-and-semiconductor-export; Qiheng Chen, "China's Emerging Approach to Regulating General-Purpose Artificial Intelligence: Balancing Innovation and Control," *Asia Society Policy Institute*, 7 February 2024, https://asiasociety.org/policy-institute/chinas-emerging-approach-regulating-general-purpose-artificial-intelligence-balancing-innovation-and.

[3] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, adopted 24 November 2021; UNESCO, *Guidance for Generative AI in Education and Research*, published November 2023.

Against the macro backdrop of stalled Security Council reform and challenges to the multilateral trading system, the UN system's authority and enforcement capacity in emerging technology governance have generally declined, making it even more difficult to impose substantive constraints on highly politicized AI standard-setting[1]. The ITU has made incremental progress, such as the December 2025 adoption of Recommendation X.2210 covering generative AI watermark implementation guidelines[2] and the January 2026 launch of the AI and Multimedia Authenticity Standards Collaboration (AMAS) working group[3]. Nonetheless, ITU standards remain non-binding voluntary documents, with insufficient political will for cross-jurisdictional alignment under geopolitical tensions.

By contrast, quasi-standard alliances such as C2PA accumulate de facto soft authority. With thousands of member organizations by early 2026, its specifications act as de facto benchmarks in media and cultural sectors[4]. Yet its governance structure prioritizes founding corporations and core members without equal representation for Global South states and small-scale platforms, failing to deliver universally inclusive public standard governance.

4. The Diminishing Reach of the Brussels Effect: The EU's Limited Capacity to Shape Global Labelling Order Unilaterally

The EU's pace and technical choices for AI labelling prevent a GDPR-style unilateral reshaping of global rules via the Brussels Effect.

In chronological terms, the EU's operationalization of the transparency obligations under Article 50 significantly lags behind the hard law practices of some countries. China established, at an early stage, a multi-tiered and technically refined system of labelling obligations through regulatory documents such as the Provisions on the Administration of Deep Synthesis in Internet Information Services and the Measures for the Identification of AI-Generated Synthetic Content. Several U.S. states have also taken the lead

---

[1] Isabella Wilkinson, "Can the UN's New AI Governance Efforts Weather the AI Race?" *Chatham House*, 18 September 2025, https://www.chathamhouse.org/2025/09/can-uns-new-ai-governance-efforts-weather-ai-race; Huang Jiexin, "US Withdraws, China Steps In: Where Is the Great-Power Competition Behind AI Governance Headed?" Tongji University, National Institute of Innovation and Development, 23 December 2025, https://niid.tongji.edu.cn/8c/d1/c36481a363729/page.htm.

[2] International Telecommunication Union (ITU), *Implementation Guidelines for Digital Watermarking* (ITU-T X.2210), finalized 11 December 2025, https://www.itu.int/itu-t/workprog/wp_item.aspx?isn=21686.

[3] International Telecommunication Union, ITU-T, "AI and Multimedia Authenticity Standards Collaboration" (AMAS).

[4] Andy Parsons, "The State of Content Authenticity in 2026," *Content Authenticity Initiative*, 18 January 2026, https://contentauthenticity.org/blog/the-state-of-content-authenticity-in-2026.

in introducing mandatory labelling rules for areas including political advertisements. In contrast, the EU's code of practice is still at the drafting and consultation stage, and its innovation lies mainly in procedural and governance structures, rather than in proposing breakthrough solutions in specific technical methods. Geopolitical and internal socioeconomic pressures further undermine the EU's credibility and enforcement capacity as a global digital standard-setter. The *Digital Omnibus* proposal put forward by the European Commission in November 2025 plans to postpone compliance obligations for high-risk AI systems from August 2026 to the end of 2027[1]. Internal and external disputes, alongside EU concessions to the U.S. on data and platform regulation, signal softened rule export ambitions[2]. Multiple observers note a reversed Brussels Effect: rather than EU standards spreading outward, the bloc contracts inward amid heavy regulatory burdens[3].

V. Conclusion

AI-generated content labelling is not a standalone technical measure but an integrated governance mechanism that tightly couples technical pathways with responsibility structures. Its core function lies in reconstructing informational risks and responsibility boundaries by rendering AI participation visible. Divergences between China, the European Union, the United States, and other major jurisdictions in triggering thresholds, labelling formats, and responsible subjects stem from distinct fundamental judgments over informational uncertainty and the role of technical governance, resulting in a highly fragmented global regulatory landscape. Against a backdrop of divided Western governance routes and weakened mobilization capacity of traditional international organizations, China has taken the lead in establishing relatively complete domestic labelling and traceability systems, and now has the practical conditions to shift from being a rule-taker to a rule-provider on certain issues. In the future, a feasible Chinese pathway for global AI content labelling governance may take shape by embedding domestic standards into multilateral platforms, fostering de facto standards via platform scale advantages, and advancing mutual recognition and coordination through bilateral and regional agreements alongside international standardization processes, thereby lowering cross-border governance frictions.

With the shrinking reach of the "Brussels Effect," weakened

[1] European Commission, *Digital Omnibus legislative proposal*, submitted 19 November 2025, not yet in force.
[2] Bertin Martens, "The European Union Needs More than the Digital Omnibus to Make Digital Services Competitive," *Bruegel Analysis*, Vol. 38, p. 1.
[3] Li Xing et al., "The 'Brussels Effect' Turns Against the EU," *Global Times*, 21 October 2025, https://finance.sina.com.cn/jjxw/2025-10-21/doc-infurpnt9089299.shtml.

mobilization capacity of traditional international organizations, and the absence of unified U.S. domestic frameworks, the global AI labelling governance system suffers evident institutional gaps[1]. Within such a landscape, China may avoid starting from abstract geopolitical narratives and take AI-generated content labelling as a core entry point. Drawing on its institutional and technical first-mover advantages, China can proactively engage in and shape international AI governance agendas. Instead of attempting to rewrite overarching global frameworks, it shall accumulate credibility and discourse power through progressive rule export, standard alignment, and joint infrastructure construction, focusing on concrete issues with mature domestic practices and tight coupling between technology and institutions.

China's relatively comprehensive institutional and standards system for AI content labelling lays a solid foundation for external rule export. Domestic implementation has seen the rollout of labelling functions across short-video, live streaming, e-commerce, and other platforms, alongside exposed challenges including easy removal of labels and inconsistent platform enforcement, providing practical feedback for subsequent technical and regulatory optimization[2]. China's AIGC labelling system carries several externally interpretable and replicable characteristics: first, the integration of human-perceptible and machine-readable labels into a unified framework that balances user awareness and technical traceability; second, the specification of legal obligations through mandatory national standards, turning abstract regulatory principles into auditable engineering requirements; third, a full-chain responsibility structure covering both service providers and content disseminators. These features enable China to put forward systematic technical and institutional solutions in international venues.

Built on this domestic foundation, China may systematically incorporate AI content labelling into global governance discussions via multilateral mechanisms, without simply transplanting domestic rules or passively adopting EU or U.S. frameworks. On one hand, under frameworks led by the United Nations and its specialized agencies, China may promote the inclusion of metadata labelling and trusted traceability into national implementation roadmaps. On the other hand, it shall fully leverage the World Artificial Intelligence Conference (WAIC), a major China-led high-level global AI

[1] AI Data Press-News Team, "When Global AI Governance Stalls, Scientists and Civil Society Take the Lead," *AI Data Press*, 10 March 2026, https://www.aidatapress.com/news/ai-governance-fragmentation-fadi-daou-globethics.
[2] Cyberspace Administration of China, "Special Campaign of 'Clear the Internet – Rectify Chaos in AI Applications' Launched," *CAC.gov.cn*, 30 April 2026, https://www.cac.gov.cn/2026-04/30/c_1779289298718765.htm.

governance platform, to set AI content labelling as a standing agenda item. By releasing joint action initiatives, best practice guidelines, and regional coordination frameworks, China may translate domestic AIGC policy experience into shareable institutional public goods. Multiple global AI governance initiatives were announced at WAIC 2025, and dedicated action lines around content labelling and traceability may be added under this framework in the future.

Beyond formal legal and standard-setting channels, China may capitalize on its dominant market and technological strengths in short-video, e-commerce, and social media platforms to form influential de facto standards through industrial practice. Since the implementation of the *Measures for the Labelling of AI-Generated Synthetic Content* and supporting national standards, Douyin, Kuaishou, Xiaohongshu, and WeChat Channels have launched diverse labelling functions, combining overt user prompts with backend covert identification despite uneven execution levels. For overseas products and services, Chinese platforms may proactively adopt and even upgrade domestic multi-tier labelling mechanisms: first, fully deploy integrated human-perceptible and covert labelling plus risk classification experience refined from domestic business scenarios into overseas product lines; second, moderately disclose technically compatible metadata field structures and open labelling verification APIs to offer detection and mutual recognition foundations for overseas regulators, research institutes, and cooperative platforms.

In terms of formal external rule export, mutual recognition clauses for AI-generated content labelling may be inserted into bilateral and regional agreements. Through trade and data cooperation arrangements, domestic institutions can be translated into enforceable minimum international standards. On one hand, digital trade agreements, e-commerce chapters, and cross-border data flow provisions may incorporate clauses on AIGC labelling and traceability cooperation, requiring all parties to adopt minimum consensus standards for high-risk fields such as electoral information, financial advertising, and minor-related content, and mutually recognize labels generated via compliant technical pathways. Such arrangements cut repeated adaptation costs for Chinese overseas service providers and embed core elements of China's domestic regime into regional rules. On the other hand, memorandums of understanding and capacity-building cooperation projects signed with Asia-Pacific and Global South nations may take joint construction of labelling infrastructure as a key component, including co-establishment of regional labelling verification nodes, shared open-source detection tools, and joint formulation of regional metadata specifications.

On the international standardization front, China shall evolve from an active participant into a leading proposer and designer of agendas. For concrete implementation, domestic bodies including the National Information Technology Standardization Technical Committee AI Subcommittee and the National Cybersecurity Standardization Technical Committee may coordinate enterprises, research institutions, and regulators to draft international standard proposals for AI-generated content labelling, transferring mature technical clauses from GB 45438-2025 into international draft documents in appropriate forms.

Abstracts and Key Words

Global Governance of Artificial Intelligence: Current Landscape, Inherent Paradoxes, and Future Pathways

Li Yan

Abstract Against the backdrop of an accelerating evolution of intelligent society intertwined with profound changes unseen in a century, artificial intelligence (AI) has risen from a technological application to a comprehensive strategic engine that drives great power competition, reshapes security rules, changes the nature of warfare, and even restructures international power dynamics. Driven by the United Nations framework, multilateral mechanisms, regional platforms, and multi-actor coordinated efforts, the current global governance process of AI is progressing. However, due to factors such as the technological characteristics of AI, the law of technology governance, and the complex effects of geopolitics, there remains a significant structural gap between governance supply and demand. Paradoxes encountered in the governance process—including security control versus innovation vitality, transnational coordination versus divergence of interests, national security versus public interest, and ethical consensus versus cultural diversity—pose deep constraints on advancing global governance. Closing these gaps requires new thoughts and strategies. On the premise of maintaining the overall strategic direction of AI for good, practical stage-based goals should be clearly defined, and initiative pragmatic strategies focused on the largest governance commons and leveraging "small cooperation" to drive "large mechanisms," innovative agile governance models embedded in technology, scientifically defined risk thresholds and safety "red lines," and the promotion of a diverse governance system that is resilient, inclusive, and flexible. This will guide AI development along a path that remains directionally correct and provides a solid foundation for sustainable AI development and international strategic stability.

Keywords Artificial Intelligence; Technological Evolution; Social Applications; Geopolitics; Global Governance

The Dual Tracks of International AI Governance Debates under the UN Framework: The Cases of LAWS Arms Control Negotiations and the Global Digital Compact

Xu Peixi

Abstract The processes involved in AI international governance are complex and interwoven. There are many related aspects with the existing Internet governance, digital governance, and cybers disarmament

negotiation processes. The AI governance dialogue within the UN framework has not been limited to a single track but follows military and non-military threads. This article takes the LAWS disarmament process and the Global Digital Compact negotiations as examples. The article observes that the LAWS disarmament negotiations have in recent years attracted heavy attention and achieved many results, but the views of all countries are seriously polarized on issues such as whether to sign a treaty, how to apply international laws such as International Humanitarian Law, whether to expand the scope of negotiations, and whether to start a new negotiation track. In the area of non-military AI governance, the United Nations has established legitimacy by negotiating and agreeing on the Global Digital Compact, which requires the United Nations to play an important role in AI governance. China's many propositions in the field of international governance of non-military AI governance are inextricably linked to texts such as the Global Digital Compact under the United Nations framework. The future rule-making processes need to take into consideration AI capabilities of different countries, distinguish between military and non-military dimensions, analyze the similarities and differences between Internet governance and AI governance, coordinate domestic governance and international governance, clarify the roles of various stakeholders of governments, businesses and civil society, and formulate governance strategies in a cross-ministry and multistakeholder manner. This article records the LAWS disarmament process and the Global Digital Compact negotiations as two specific scenarios about military and non-military AI governance under the United Nations framework, and how China articulates international positions linked to these processes.

Keywords UN Framework; AI International Governance; LAWS Disarmament; *Global Digital Compact*

Research Trends of Artificial Intelligence Governance by Think Tanks in the West

Lang Ping Zhang Mengxi

Abstract Nowadays, artificial intelligence (AI) is profoundly reshaping national security, international competition, and the global landscape. Based on an analysis of AI-related research publications by seven influential think tanks in the West between January 2023 and April 2026, this study delineates a multidimensional analytical framework, encompassing social impact and domestic regulation, national strategy and great power competition, military application and security risk, industrial transformation and market restructuring, as well as international cooperation and global governance. Overall, think tanks in the West are adopting an increasingly pragmatic and policy-driven approach to integrating AI into national and global agendas. Their research trends can reveal the evolution of policy thinking patterns, while providing a critical lens for deciphering the future dynamics of technological competition and global governance.

Keywords Think Tank; Artificial Intelligence; Great-power Competition;

National Security; Global AI Governance

Research on the Role and Strategies of International Organizations in the AI Governance Process

Feng Shuai *Zhang Tianyu*

Abstract In the current global AI governance process, international organizations play an important role. There are three types of international organizations participating in AI governance, universal intergovernmental organizations represented by the United Nations; multilateral national mechanisms represented by the G20; and many functional technical organizations. International organizations play an important role as rule makers, platform builders, resource coordinators, and supervision promoters in the AI governance process. They use four major strategies: policy guidance, project cooperation, capacity building, and publicity and promotion in the governance process. Play a role. International organizations mainly adopt three methods to participate in governance activities, namely, building platforms and forums, relying on neutrality and appeal to build consensus; formulating standards and ethical frameworks, relying on authority and professionalism to establish global rules; and promoting capacity building and project piloting, based on inclusive and development missions to bridge the digital divide. These three methods are connected to each other and form a full-chain governance system of "consensus agglomeration-rule establishment-implementation." Although international organizations have played an irreplaceable role in promoting global AI governance, they still face structural limitations such as insufficient execution, differences of interests, and concentration of resources. There is still a need to strengthen institutional integration and innovation, further leverage the role of international organizations, and promote the global AI governance process.

Keywords AI Governance; International Organization; Role; Strategy; Full-chain Governance System

The "Brussels Mirage" under the EU's Digital Sovereignty Strategy: A Case Study of the Technological Disempowerment Dilemma in the Governance of the EU AI Act

Yu Nanping *Shu Zhongsheng*

Abstract Driven by its digital sovereignty strategy, the European Union has taken the lead in regulatory legislation across three dimensions—data, technology, and markets—while extending its reach through extraterritorial jurisdiction rules. The EU Artificial Intelligence Act, the world's first systematic AI regulation, leverages the scale of the EU single market to replicate the regulatory pathway of the General Data Protection Regulation (GDPR). It seeks to generate a "Brussels Effect" that establishes the "EU model" as the global paradigm for AI governance. In practice, however, the EU trails significantly behind China and the United States in AI computing capacity, industrial investment, and engineering capability. Consequently, within the governance framework of the AI Act, technical standard-setting, regulatory capacity, and

corporate willingness to comply manifest pronounced fragmentation and heterogeneity. When technological strength cannot sustain the global diffusion of rules, even a textually rigorous regime may run idle, and the intended Brussels Effect risks degenerating into a "Brussels Illusion" stemming from technological disempowerment. Against the current backdrop of accelerating technological competition and diverging regulatory regimes in global digital governance, China should therefore examine the spillover effects of the EU's AI legislative model with sober judgment. Only when technological supply, industrial diffusion, and national regulatory capacity reinforce one another can regulatory frameworks avoid devolving into a form of technological disempowerment marked by "norms without technology, order without foundation," thereby safeguarding national strategic security.

Keywords European Union; Artificial Intelligence Act; Brussels Effect; Technological Disempowerment; National Strategic Security

Normative Agency in Latecomer Domains: The Global South's Strategies in Artificial Intelligence Governance

Cai Cuihong Guan Hang

Abstract Norms of artificial intelligence governance are diffusing at an accelerating pace worldwide. The Global South has long been regarded as a passive norm-taker and a latecomer follower in emerging domains such as artificial intelligence. However, this assumption overlooks the strategic agency that Global South countries have demonstrated in norm selection, adaptation, and co-construction. Centering on the question of how actors exercise strategic agency in the process of norm-building within their latecomer domains, this article identifies four types of agentic strategies: developmentalization of issue framing, collectivization of agency, localization of normative content, and issue-linkage through multilateral action. Through an analysis of AI governance practices by African and Southeast Asian states and regional organizations, as well as India, the article validates the explanatory effectiveness of these four strategies in accounting for the Global South's participation in AI governance norm-building. The Global South is not a passive norm-taker in AI governance; its agency, while limited, is meaningful—conditioned by development needs, institutional capacity, and the constraints of international structures.

Keywords Global South; Artificial Intelligence Governance; Norm Diffusion; Agency

Institutional Construction of Artificial Intelligence Data Security Governance from the Perspective of the Global South

Xu Duoqi Ying Yue

Abstract In the era of artificial intelligence, practices of data flow and model training have transformed data colonialism into AI colonialism, a more intense and more concealed form of domination. Countries in the Global South are structurally disadvantaged in relation to data

exploitation and national security, while existing solutions remain inadequate. Despite this unfavorable position, the Global South can still leverage its own resources to engage in asymmetric bargaining. Vertically, it can strategically choose among the services and rules offered by multiple AI innovation hubs, thereby generating competitive pressure. Horizontally, it can expand South-South cooperation to strengthen its collective bargaining leverage. The implementation of this strategy requires institutional design along three dimensions. In the dimension of security-oriented competition, institutions should reward AI service providers that respect data sovereignty and promote a turn toward open-source models. In the dimension of security baselines, institutions should prevent critical AI infrastructure from becoming excessively dependent on any single source of service provision. In the dimension of security cooperation, countries in the Global South should strengthen solidarity and pursue international collaboration through favorable mechanisms such as the United Nations and the Belt and Road Initiative.

Keywords Global South; Data Security; Artificial Intelligence Colonialism; Asymmetric Bargaining

The Authentication and Governance of AI Generated Content: A Global Perspective

Li Wenlong *Luo Jiakun*

Abstract As generative AI reshapes the global information environment, AI-generated content authentication has moved from a transparency device to a broader infrastructure for trust, traceability and responsibility. This article unpacks the technical foundations of content labelling and examines their limits in real-world circulation. It shows that labelling systems can easily break down when content moves across platforms, is re-edited, generated through open-source models, or subjected to privacy and verification constraints. The article then maps the emerging regulatory landscape across China, the European Union, the United States, South Korea, Vietnam, Brazil and Japan. It shows that these regimes are working from different assumptions about informational risk, platform responsibility and the proper place of technical governance, and a genuinely interoperable global labelling framework is unlikely to come together easily. The article further contends that a range of factors—industrial and policy divisions, weak China—West dialogue, the limited convening power of international organisations, and the shrinking reach of the Brussels Effect—all hold back meaningful coordination. China could potentially move beyond rule-taking and play a more active role in shaping global AI content governance by building on its domestic standards, plugging them into multilateral forums, scaling them through major platforms, and pursuing mutual recognition through regional and international standard-setting processes.

Keywords AI-generated Content; Digital Watermarking; Content Provenance; Global AI Governance; AI Authenticity

Introduction of the Center for Global AI Innovative Governance (CGAIG)

The Center for Global AI Innovative Governance (CGAIG) was established at Fudan University with strong support from the Shanghai municipal government, and serves as an international platform dedicated to research, dialogue, and collaboration on global AI innovative governance. Leveraging Fudan University's strengths in interdisciplinary research and international cooperation networks, CGAIG brings together scholars, policymakers, and industry experts from around the world to advance research and policy innovation in AI governance.

The Center's Secretariat is located at the Fudan Development Institute. CGAIG promotes talent development, youth exchanges, professional training, AI for Science (AI4S), and research translation through organizing international forums, thematic seminars, and joint research programs. The Center seeks to build a collaborative platform linking Shanghai with the global community, facilitating dialogue among academia, industry, and policymakers. Through these activities, CGAIG contributes to the development of governance frameworks, standards, and policy guidelines for AI.

Website: https://cgaig.fudan.edu.cn/
Email: cgaig@fudan.edu.cn
WeChat Official Account: cgaigfudan
Address: Think Tank Building, 220 Handan Road, Shanghai, China
Post code: 200433
Tel: 021-65641067

Call for Papers

Global AI Innovative Governance (Monographic Series)

The Center for Global AI Innovative Governance (CGAIG) serves as a specialized and international collaborative platform focused on global AI innovation governance. In partnership with Shanghai People's Publishing House, we are launching the book series Global AI Innovative Governance (published bi-annually as a monographic series), an initiative dedicated to establishing a sustained and influential series for academic and policy research on AI governance. We cordially invite experts and scholars from around the world to submit their original research.

The series adopts a global perspective by integrating Chinese experience with international comparative studies, focusing on the critical dimensions of artificial intelligence governance. We welcome research submissions that explore the development of regulatory frameworks and international public goods within Governance Systems and International Rules, as well as the dynamics of infrastructure and technological sovereignty under Factor Allocation and Industrial Transformation. Furthermore, the series seeks to address Social Impact and Risk Control through capacity-building and mitigation strategies, alongside the promotion of Ethics and Civilizational Dialogue to foster value-based ethics and mutual learning across global contexts.

Submissions must closely address the theme and be original, unpublished, and characterized by clear argumentation and rigorous reasoning. While the series welcomes manuscripts in both Chinese and English, English submissions may be translated into Chinese for publication upon successful peer review. In terms of length, manuscripts should generally range from 8, 000 to 10, 000 words for English or 12, 000 to 15, 000 characters for Chinese. All citations must be presented as footnotes in accordance with the Shanghai People's Publishing House Style Guide appended below. Furthermore, each submission must include the author's name, institutional affiliation, contact information, an abstract, keywords, and a brief biographical note.

Submission Process
Email: cgaig@fudan.edu.cn
Subject Line: "CGAIG Submission + Author Name + Article Title"
Timeline: This call for papers will remain open. Manuscripts are accepted and reviewed on a rolling basis.
Link: https://cgaig.fudan.edu.cn/cgaigchinese/wqqrgznzlw/list.htm

Center for Global AI Innovative Governance