CGAIG 全球人工智能创新治理中心
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

Fudan Think Tank Building,
No. 220 Handan Road, Shanghai, China, 200433
Tel: +86-21-65641067
Email: cgaig@fudan.edu.cn
Website: https://cgaig.fudan.edu.cn

FUDAN
REPORT SERIES

WAIC 2026 Forum
AI Global Governance and Sustainable Development: Challenges and Countermeasures

Global Artificial Intelligence Governance Trends: Observations Since the *Global AI Governance Action Plan*

Yao Xu et al.

Center for Global AI Innovative Governance
Fudan Development Institute, Fudan University

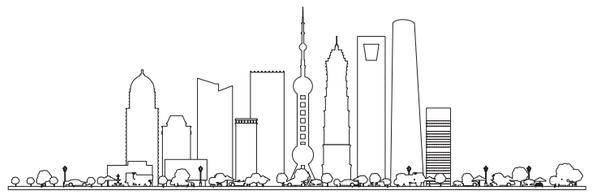
FUDAN REPORT SERIES

Global Artificial Intelligence Governance Trends: Observations Since the *Global AI Governance Action Plan*

Yao Xu et al.

Center for Global AI Innovative Governance
Fudan Development Institute, Fudan University

July 2026



Project team leader:

Yao Xu, Secretary General of Center for Global AI Innovative Governance, Associate Research Fellow at Fudan Development Institute

Project team members:

Jiang Tianjiao, Researcher at the Center for Global AI Innovative Governance, Associate Prof at Fudan Development Institute

Xin Yanyan, Deputy Secretary-general of the Center for Global AI Innovative Governance, Assistant Prof at Fudan Development Institute

Shen Qi, Project Director of Center for Global AI Innovative Governance

Luo Haoyue, Project Director of Center for Global AI Innovative Governance

Zhang Shuyan, Research Assistant, Center for Global AI Innovative Governance

Yang Zhao, Research Assistant, Center for Global AI Innovative Governance

Zhang Ao, Research Assistant, Center for Global AI Innovative Governance

Wang Yibo, Research Assistant, Center for Global AI Innovative Governance

Yuan Luming, Research Assistant, Center for Global AI Innovative Governance

Hu Xiangyu, Research Assistant, Center for Global AI Innovative Governance

Wang Yexu, Research Assistant, Center for Global AI Innovative Governance

Jiang Junji, Research Assistant, Center for Global AI Innovative Governance

Contents

Core content

Since the *Global AI Governance Action Plan* was proposed at the 2025 World Artificial Intelligence Conference (WAIC), the global governance of artificial intelligence is moving from a principled initiative to a new stage of mechanism undertaking, capacity building and project implementation. At present, the global development of artificial intelligence still shows a significant imbalance: a few countries and leading enterprises continue to concentrate advantages in computing power, data, models, platforms, and standard setting, while a large number of developing countries generally face constraints such as insufficient infrastructure, high costs of technology access, weak local language support, limited governance capabilities, and inadequate participation in international rule-making. Whether artificial intelligence can truly become a technological resource serving shared development depends not only on the capabilities of the models themselves, but also on whether countries can participate equitably in governance, obtain affordable technological capabilities, develop local application ecosystems, and establish corresponding foundations for safety governance.

Global artificial intelligence governance is entering a new stage characterized by equal emphasis on development and security.

Security risks remain a bottom-line issue that the international community must safeguard, but capacity building, inclusive development, industrial applications, and technological accessibility are becoming increasingly prominent items on the agenda.Global AI governance is entering a new stage of "equal emphasis on development and security". Security risks are still a bottom-line issue that the international community must adhere to, but capacity building, inclusive development, industrial application and technology accessibility are becoming more prominent on the agenda. Future governance will no longer just discuss "how to prevent risks", but also answer "how to make AI truly accessible, affordable and well-used to more countries.

The "intelligence divide" is emerging as a new form of structural inequality following the digital divide.

Imbalances in computing power, data, algorithms, languages, talent, and application scenarios have led countries in the Global South to face deeper technological dependency and passive rule-taking in the era of artificial intelligence. Without effective cooperation, developing countries may remain for a long time in positions of data providers, market consumers, and rule adopters, making it difficult for them to move into higher value-added areas such as model development, platform governance, and standard-setting.

Great power competition is exacerbating the fragmentation of global governance.

Competition over semiconductors, computing power, foundation models, critical minerals, data flows, and technological standards is reshaping the institutional environment for artificial intelligence governance. On the one hand, such competition accelerates technological iteration; on the other hand, export controls, technology alliances, and the divergence of standards may also constrain the policy space of smaller and medium-sized countries, increase the costs of cross-border cooperation, and weaken the efficiency of global public goods provision.

The focus of North–South cooperation should shift from "technology transfer" to "capacity co-building."

Truly effective artificial intelligence cooperation should not remain at the level of equipment procurement, platform access, or short-term training, but should be oriented toward long-term development in areas such as local computing infrastructure, data governance capacity, talent development systems, low-resource language models, industry-specific application scenarios, and safety evaluation tools. Its goal is not to create new forms of technological dependency, but to help developing countries build sustainable and autonomous capabilities for artificial intelligence development.

China's provision of international public goods in artificial intelligence is becoming an important variable in global governance.

Centered on the *Action Plan for Global AI Governance*, China is advancing alignment with multilateral mechanisms, the supply of open-source models, capacity-building initiatives, cooperation in digital infrastructure, and industrial

application scenarios, thereby promoting the transformation of AI technologies from competitive resources controlled by a small number of countries and enterprises into development tools that are more widely accessible, participatory, and transferable. This process is not only related to China's own role positioning in global AI governance, but also to whether artificial intelligence can truly serve shared development.

Looking ahead, the key to global governance of artificial intelligence does not lie in establishing a closed rule system defined by a small number of countries, but in forming a more open, inclusive, and sustainable network of cooperation. If China continues to base its efforts on genuine multilateralism, lower technological barriers through open-source and openness, strengthen development foundations through capacity building, and respond to people's livelihood needs through scenario-based cooperation, it may be able to provide more practically meaningful international public goods in narrowing the global intelligence divide, enhancing governance capacity in developing countries, and promoting the beneficial development of artificial intelligence.

Part I The Global Governance Landscape of Artificial Intelligence

Artificial intelligence represents an unprecedented technological paradigm. Turing Award laureate and "father of reinforcement learning" Richard S. Sutton has argued that "artificial intelligence is the inevitable next step in the evolution of the universe, and humanity should embrace it with courage, pride, and a spirit of adventure." At present, AI has been applied across all industries, with application scenarios spanning healthcare, finance, transportation, agriculture, manufacturing, and retail. The rapid iteration and wide-ranging applications of AI have also placed higher demands on global governance. This section reviews

global AI governance developments since 2025 and finds that significant new progress has been made in the global AI governance system. Countries have introduced policies encouraging technological development, while security-related issues have relatively weakened in emphasis. International cooperation mechanisms centered on the United Nations as the main channel are gradually being improved, forming a governance pattern of coordination between the UN and regional mechanisms. Governance approaches are undergoing transformation and upgrading, moving toward application-oriented, front-loaded, and more scientific development. However, global AI governance also faces challenges: great-power competition and cooperation have become the norm, and multiple governance contradictions coexist. Looking ahead, advancing the global AI governance system requires seeking common ground while reserving differences on development issues, drawing clear red lines on security issues, actively leveraging the role of the Global South, and promoting international cooperation to reach a new level.

(I) New Developments in the Global AI Governance System

Since 2025, three major developments have emerged in global AI governance.At the national level, promoting development has become the central theme of AI policymaking. At the international level, the United Nations has continued to strengthen its governance mechanisms and is increasingly serving as the primary channel for global AI governance.In terms of governance approaches, increasing emphasis has been placed on application-oriented governance, upstream governance, and evidence-based governance.

a) Shifts in National AI Policies

Artificial intelligence has increasingly become a strategic engine of national development, prompting governments around the world to elevate AI to the level of national strategy. Against the broader backdrop of balancing innovation and regulation, development-oriented policies have become increasingly prominent as countries seek to secure a competitive advantage in the next wave of industrial transformation.

China has promoted AI capacity building on multiple fronts. At the conceptual level, it has emphasized the coordinated advancement of development and security by issuing the *Guiding Opinions on the In-depth Implementation of the "AI Plus" Initiative*, promoting the deep integration of AI across industries while strengthening the legal and security foundations for AI development. At the technological level, the *15th Five-Year Plan Outline* proposes an integrated innovation strategy spanning models, chips, cloud infrastructure, and applications, fostering coordinated innovation across the entire AI value chain. Institutionally, China released the *AI Safety Governance Framework (Version 2.0)*, which builds upon practical experience and incorporates emerging technologies and governance concepts to comprehensively refine governance principles, risk classification, technical

measures, governance mechanisms, and implementation guidelines.

The United States and several other Western countries have undergone a marked shift in their AI policies, with their strategic focus moving further from risk prevention toward technological competitiveness, industrial development, and infrastructure expansion. Domestically, the policy orientation has become increasingly deregulatory. Following the start of President Trump's second term, the United States issued Executive Order 14179, Removing Barriers to American Leadership in Artificial Intelligence, rescinded Executive Order 14110, and directed the development of a new national AI action plan. In July 2025, the White House released America's AI Action Plan, which sets out policy actions around accelerating innovation, building AI infrastructure, and leading in international diplomacy and security. In November 2025, the administration further launched the Genesis Mission, seeking to mobilize national laboratories, scientific data, computing resources, and public–private collaboration to accelerate AI-enabled scientific discovery.

Countries across the Global South are pursuing sovereign AI strategies to achieve secure and autonomous development. India, Brazil, Malaysia, and several other countries have introduced a range of national AI strategies. The *India AI Mission* identifies seven strategic pillars for building sovereign AI and is developing datasets covering more than twenty-two Indian languages through the Bhashini initiative while supporting the development of the indigenous full-stack AI platform Sarvam AI. Brazilian President Luiz Inácio Lula da Silva has repeatedly criticized the monopolization of AI by major technology companies. Under the *Brazilian Artificial Intelligence Investment Plan (PBIA 2024–2028)*, Brazil plans to invest BRL 23 billion (approximately USD 4 billion) in AI development over a four-year period. Malaysia has become the first country in Southeast Asia to establish a sovereign full-stack AI ecosystem. Malaysian company Skyvast Cloud operates AI infrastructure based on Chinese AI chips and DeepSeek's open-source models. Meanwhile, Malaysia's national MaaS platform and the national sovereign AI talent development laboratory, Z·UM AI Lab, have locally optimized Z.ai's open-source foundation models to maintain a sovereignty-centered data security architecture and ensure the autonomy of the country's national AI platform.

Countries have adopted increasingly differentiated approaches to international AI policy. China has promoted global AI cooperation on multiple fronts. At the conceptual level, it advocates open innovation, inclusive development, and safe and trustworthy AI. In 2025, China successively released the *Action Plan for Global AI Governance and the International Cooperation Initiative on "AI Plus"*, promoting AI as an international public good that benefits all humanity. At the technological level, China launched the *International Open-Source AI Cooperation Initiative*, aiming to foster a global open-source ecosystem through international collaboration. The DeepSeek-R1 model has also demonstrated a new technological pathway through its open-source, low-cost, and high-performance development model. At the institutional level, China has proposed establishing a World AI Cooperation Organization as a key international public cooperation platform for the Global South, with the goal of promoting a more equitable and open global AI governance framework. By contrast, Western countries have gradually de-emphasized AI safety in their international agendas. The Paris AI Action Summit and the 2026 India AI Impact Summit shifted their focus away from AI safety toward development, innovation, and investment. The United Kingdom and the United States also renamed their AI Safety Institutes, reflecting a transition from setting guardrails for AI technologies to developing technical standards and addressing risks arising from international technological competition. Similarly, the G7 *Statement on AI for Prosperity*, released in June 2025, further deemphasized a singular security-centered narrative.

b) Institutionalization of Governance Mechanisms within International Organizations

Global AI governance is shifting from a fragmented set of parallel processes toward a UN-centered multilateral cooperation architecture, gradually consolidating into a more structured governance configuration characterized by coordinated interaction between the United Nations system and regional mechanisms.

The United Nations has further consolidated its role as the principal channel for global governance through

a set of reinforcing institutional developments, including internal reforms, the establishment of specialized scientific authority, and the institutionalization of policy dialogue mechanisms. In early 2025, the Office of the UN Tech Envoy was formally renamed the *United Nations Office for Digital and Emerging Technologies*, which now serves as the system-wide coordination hub for digital issues, signaling a more coherent and clearly delineated architecture for AI governance within the UN system. In parallel, the UN established the Independent International Scientific Panel on Artificial Intelligence under General Assembly Resolution A/RES/79/325. Composed of 40 experts with broad geographical representation and interdisciplinary expertise, the Panel is mandated to provide independent scientific assessments for global AI governance. In July 2026, the first Global Dialogue on AI Governance convened in Geneva and discussed the Panel's preliminary findings, while follow-up work and more detailed assessments are expected to continue in subsequent UN processes. At the same time, the *Global Dialogue on AI Governance*, launched in September 2025, has become a structured platform for policy exchange, coordination, and normative alignment aimed at narrowing the technological and regulatory divide between the Global North and the Global South, while in December 2025 the UN General Assembly formally designated the *Internet Governance Forum (IGF)* as a permanent institutional mechanism, ending its two-decade provisional status and strengthening the long-term institutionalization of multistakeholder governance in the digital domain.

Regional multilateral mechanisms are increasingly interacting in a mutually reinforcing manner with the UN-centered framework, giving rise to an emerging architecture in which regional nodes interface with a global governance core, as illustrated by developments within the G20, BRICS, and the Shanghai Cooperation Organization (SCO). Within the G20, under South Africa's 2025 Presidency, the *AI for Africa's Development* initiative integrates AI governance with the sustainable development agenda by planning the establishment of five AI innovation centers across Africa to support applications in agriculture, healthcare, education, and other public welfare domains, while aligning with the United Nations 2030 Sustainable Development Goals and facilitating the translation of G20-level consensus into coordinated global action within the UN framework. Within the BRICS mechanism, leaders adopted the *Declaration on Global AI Governance*, explicitly rejecting fragmentation and duplication in global AI governance architectures and reaffirming that the United Nations should remain the central pillar of global AI governance. Within the Shanghai Cooperation Organization (SCO), the Council of Heads of State adopted the *Tianjin Declaration*, reaffirming support for the UN's central coordinating role and issuing a *Statement on Strengthening the Development of the Digital Economy*, calling for deeper cooperation on digital economy development under the SCO framework.

c) Transformation and Upgrading of Governance Pathways

With the expansion of global AI governance practices and intensified international interaction, accumulated experience has led to a gradual maturation of governance pathways, which are increasingly characterized by three converging approaches: scenario-based governance, front-end governance, and science-based governance.

Scenario-based governance has become a central focus of contemporary AI regulation. Unlike the earlier technology-centric approach that primarily targeted algorithms and *foundation models* themselves, current governance frameworks increasingly emphasize application-oriented oversight across sectors and levels, highlighting the risks emerging when *foundation models* are deployed in specific real-world contexts and interact with complex socio-technical environments, systems, and human users. China has adopted a scenario-specific regulatory approach, issuing departmental regulations and technical standards covering algorithmic recommendation systems, deep synthesis technologies, generative AI, anthropomorphic interactive services, and intelligent agent security. The European Union, through the *Artificial Intelligence Act*, has established a risk-based classification framework, complemented by soft-law instruments such as the *Code of Practice for General-Purpose AI* and the *Transparency Guidelines for Generative AI*, which encourage ex-ante compliance by firms and further specify requirements related to transparency, copyright protection, and systemic risk management.

The focus of AI governance is also shifting upward along the technological stack. Unlike earlier stages, which concentrated on algorithmic discrimination and privacy violations, contemporary AI safety governance increasingly targets frontier models and high-capability general-purpose systems. At the 2025 World Artificial Intelligence Conference (WAIC), numerous experts emphasized the urgency of strengthening safety governance, noting that once model capabilities surpass certain thresholds, risks may increase in a nonlinear and potentially discontinuous manner, and such high-capability systems may exhibit behaviors that are difficult to address through traditional regulatory tools. They further stressed that safety considerations must be integrated at the front end of model development and deployment rather than treated as ex-post corrective measures. Consequently, the global AI safety agenda is increasingly focused on establishing governance guardrails at the model level, including capability evaluations, red-teaming exercises, and threshold-based monitoring mechanisms.

AI governance is increasingly becoming more science-driven, with AI safety governance in particular emerging as a potential common denominator across jurisdictions. On the one hand, despite fluctuations in the political prioritization of safety issues, scientific communities continue to advance cooperation, with a landmark development being the release of the International AI Safety Report in January 2025. Efforts are underway to build consensus around technical tools such as evaluation benchmarks, risk assessment frameworks, and incident reporting systems, with the aim of transforming these instruments into global public goods that can reduce divergences in risk perception, provide actionable regulatory tools, and offer a politically lower-sensitivity entry point for international cooperation. On the other hand, AI safety governance is becoming increasingly "engineered," reflected in the emergence of standards, structured management systems, and auditable safety mechanisms, signaling a broader shift toward governance systems that are more traceable, verifiable, and accountable.

(II) Major Power Competition in Artificial Intelligence

Artificial intelligence has evolved beyond its role as a driver of industrial upgrading and economic growth, becoming deeply embedded in national security, military capability, social governance, and international rule-making. It has emerged as a pivotal variable in great power strategic competition. Against this backdrop, major powers are engaging in comprehensive competition across key nodes—including data, computing power, models, applications, governance standards, and infrastructure—thereby introducing new dynamics into the global AI governance landscape.

a) Core Arenas of Great Power AI Competition

First, data governance has become a central battleground. Against the backdrop of rapid advances in generative AI, competition over data among major powers is shifting from a focus on scale and volume to the quality of governance capabilities. Key processes—including compliant data collection, cleaning and annotation, licensed data sharing, cross-border data transfer risk assessment, and liability traceability—collectively form the critical chain through which raw data is transformed into intelligent capabilities. The extent to which countries can establish stable, scalable, and replicable data governance systems will largely determine whether data resource advantages can be converted into sustained institutional competitiveness. China's *Digital China 2025 Action Plan* advances market-oriented reforms in data factor allocation, strengthening coordination between central and local governments, developing an integrated national data market, and promoting initiatives such as the "Eastern Data, Western Computing" program, public data operations, and digital talent development. In the United States, the "Genesis Mission" promotes public–private data collaboration, with plans to tier-access three categories of federal scientific data and engage firms such as NVIDIA and Oracle, aiming to build a national data ecosystem that links data sharing with innovation-driven growth.

Second, computing power has

become a decisive variable shaping the pace of AI development and a frontline domain of competition among major powers. Since the establishment of the "Stargate" project by OpenAI, SoftBank, Oracle, and MGX, the United States has invested more than USD 90 billion in computing clusters and supporting energy infrastructure. Following the launch of the "Genesis Mission," Amazon announced an additional USD 50 billion investment in the construction of data centers and nuclear energy facilities. Computing resources are increasingly concentrated in a small number of countries and leading firms, while entry barriers continue to rise. In the long term, this trend is likely to exacerbate the "computing divide," further consolidating advantage among a limited set of dominant actors.

Third, in the domain of model ecosystems, the performance gap between major powers has continued to narrow. The *Stanford AI Index Report 2026* indicates that the performance gap between leading Chinese and U.S. AI models has effectively converged. While the United States held a substantial lead in 2023, this advantage narrowed significantly by early 2025 and has since remained within a narrow margin. As of March 2026, top U.S. models outperform leading Chinese models by only 2.7 percent, with the gap fluctuating between near parity and low single-digit differences over the past year.

Fourth, in terms of application deployment, China demonstrates distinctive advantages in technological implementation and industrial diffusion, supported by a comprehensive manufacturing system and diverse application scenarios, enabling strong scale effects and scenario-driven advantages at the deployment level. In contrast, the United States faces a persistent gap between frontier innovation and industrial application, often characterized as "strong in the laboratory, weaker in deployment." To address this, the United States has introduced mission-oriented policy instruments such as the "Genesis Mission," seeking to strengthen coordination between innovation and application through government-led mechanisms and accelerate the translation of AI research outputs into real-world industrial capabilities.

Fifth, governance norms have become an increasingly important arena of strategic competition. Both China and the United States attach high importance to institutional spillover effects, namely the establishment of domestic evaluation frameworks, compliance templates, and risk classification systems that can subsequently be internationalized to attract third countries into their respective ecosystems and secure first-mover institutional advantages in global AI governance. Notably, some U.S. perspectives acknowledge that China has already achieved an early-mover position in certain international standard-setting domains, with its agenda-setting influence in global governance platforms continuing to strengthen.

Sixth, infrastructure and critical minerals have emerged as foundational arenas of competition. At the infrastructure level, the *Artificial Intelligence Action Plan* highlights that future competition between China and the United States will primarily center on semiconductor manufacturing, power grid development, and scientific data accumulation. While the United States maintains advantages in advanced chip design and cloud platforms, its power grid and energy infrastructure remain relatively constrained, posing tangible limits on large-scale computing expansion. Projects such as "Stargate" seek to address this by deeply integrating physical and digital infrastructure to establish foundational support for next-generation AI systems. In the domain of critical minerals, Reuters reports that U.S. President Donald Trump and Commerce Secretary Howard Lutnick have been directly involved in negotiations and exerted pressure regarding the development of tungsten resources in Kazakhstan, underscoring heightened strategic attention to control over materials essential to computing power supply chains.

b) Key Characteristics of Major Power Competition in Artificial Intelligence

First, the strategic positioning of AI competition continues to rise, with artificial intelligence increasingly embedded within national security considerations and broader struggles over strategic primacy. During the second Trump administration, U.S. policy orientation in the AI domain has been further intensified, with the competitive objective toward China shifting from maintaining a relative lead to constructing an "absolute capability gap." Early in his term, Executive Order

No. 14179, *Eliminating Barriers to American Leadership in Artificial Intelligence*, explicitly called for sustaining U.S. dominance in the global AI race, setting the tone for a continued pursuit of absolute technological advantage. In July 2025, the *Artificial Intelligence Action Plan* further framed AI as a transformative force capable of simultaneously driving an industrial revolution, an information revolution, and a cultural renaissance, emphasizing that "the United States and its allies must win this race." In November of the same year, the Trump administration launched the "Genesis Mission," a comprehensive national strategy designed to advance AI-enabled scientific research through state mobilization, explicitly identifying China as the primary strategic competitor and further underscoring the securitization and strategic escalation of AI competition.

Second, competitive instruments are expanding from market-based rivalry toward institutionalized and alliance-based containment, with an emerging "de-risked" or selectively decoupled technological bloc becoming increasingly entrenched. On the containment side, the United States has continuously strengthened export controls on advanced semiconductors, particularly targeting critical nodes in China's access to frontier computing power. Restrictions on high-end chips such as NVIDIA's H200 have been repeatedly adjusted, aiming to precisely constrain access to cutting-edge computational capacity. At the same time, the United States has actively coordinated with allies and partner countries to restructure supply chains and R&D cooperation, promoting a partial "de-Chinafication" of technological ecosystems and forming a network of constraints on China in the AI domain. On the capability-building side, U.S. strategy focuses on three priority areas: critical materials and biotechnology, nuclear fission and fusion alongside advanced manufacturing capabilities, and frontier microelectronics domains such as semiconductors and quantum information. The United States also places strong emphasis on allied R&D cooperation; for instance, the December 2025 U.S.–Australia ministerial meeting highlighted efforts under AUKUS Pillar II to enhance trilateral capabilities in AI and autonomous weapons systems, further reinforcing the logic of a technology-based alliance system.

Third, strategic competition continues to oscillate between zero-sum geopolitical logic and pragmatic economic considerations, exhibiting pronounced transactional behavior and policy volatility. Under the Trump administration, U.S. foreign policy retains a strong transactional character, which is also reflected in the AI domain. In the case of semiconductor export controls, industry stakeholders have argued that prolonged suspension of shipments to China could result in losses amounting to tens of billions of dollars. Influenced by corporate lobbying, the administration has moved to relax export restrictions on H200-class chips. However, strong domestic security-oriented constituencies remain influential, and there is a realistic possibility of policy reversal and tightening, contributing to significant uncertainty and volatility in regulatory direction. In addition, domestic AI policymaking faces structural tensions. State-level AI legislation exhibits uneven implementation due to divergent political and socio-economic interests, while federal initiatives such as the "Genesis Mission," despite its accelerated implementation timeline of 60, 90, and 120 days, may encounter legal and jurisdictional frictions with subnational regulatory frameworks, the extent of which remains to be tested in practice.

Fourth, the logic of competition is shifting from isolated breakthroughs toward full-lifecycle and full-stack systemic capability competition, marking the entry of AI rivalry into a more systematized, whole-of-government phase. Major power competition is increasingly understood as extending across the entire AI value chain rather than concentrated technological inflection points. It is widely recognized that reliance on single-point technological advantages is insufficient to sustain long-term leadership; instead, sustained competitiveness requires simultaneous progress across computing power acquisition and orchestration efficiency, data governance and compliance capacity, engineering and industrial iteration capabilities, talent and organizational systems, energy and power infrastructure, and safety evaluation and accountability mechanisms. The launch of the "Genesis Mission" symbolizes the elevation of U.S. AI strategy to the level of a national innovation system, with the government assuming a more directive role through large-scale strategic investment and cross-agency coordination, making the "whole-of-nation" character of AI competition increasingly evident.

c) Potential Implications of Major Power Strategic Competition for Global AI Governance

First, great power strategic competition is reshaping the institutional environment for global AI rule-making, introducing structural constraints on the formation of unified standards. While there is a baseline consensus among stakeholders on issues such as AI risk classification, model evaluation, safety auditing, and liability chains, significant divergences persist in areas including technological sovereignty, cross-border data flows, and supply chain security. In an increasingly bloc-based environment, the diffusion of rules often carries strategic intent. However, if emerging regulatory regimes remain mutually incompatible, global AI governance risks fragmenting into parallel and competing standards, raising the transaction costs of cross-border cooperation and weakening the efficiency of global public goods provision.

Second, the policy space available to developing and smaller states in selecting technological pathways and governance models is becoming increasingly constrained. As major powers integrate AI into national security and strategic competition frameworks, these countries face growing pressures of alignment in areas such as technological cooperation, infrastructure development, and data governance regimes. First, dependency structures may lock in development trajectories: once integrated into a particular ecosystem, path dependence emerges across standards interfaces, data flows, and compliance frameworks. Second, in international rule-making processes, smaller states often possess limited agenda-setting influence, making it difficult to fully articulate their development priorities within forums dominated by major powers. As a result, the autonomy of development pathways faces tangible constraints.

Finally, the influence of non-state actors is rising, and the structure of governance authority is becoming increasingly pluralistic. As AI innovation becomes highly dependent on technology firms and technical communities, companies now control critical resources in model development, computing orchestration, and data governance, significantly increasing their participation in policymaking and standard-setting processes. Technical communities further shape technological trajectories and emerging ethical norms through open-source ecosystems, academic networks, and transnational collaboration platforms. As a result, governance is gradually shifting from a state-centric model toward a more polycentric and multistakeholder architecture, characterized by deeper interaction among governments, firms, research institutions, and civil society. While this decentralization may enhance governance efficiency and responsiveness, it also introduces challenges related to blurred accountability and coordination complexity across regulatory systems.

(III) Existing Challenges and Governance Paradoxes

Global AI governance currently reveals three key governance paradoxes that will require coordinated international responses in the future: the tension between governance coordination and regulatory fragmentation, the tension between technological advancement and risk controllability, and the tension surrounding algorithmic discrimination and value divergence.

a) The Tension Between Governance Coordination and Regulatory Fragmentation

Fragmentation in AI governance remains a salient challenge, driven by significant differences in technological development levels, industrial structures, and policy priorities across countries, which in turn lead to divergent regulatory pathways and increased difficulty in achieving policy coordination. On the one hand, the absence of mutual recognition of rules and standards raises compliance costs for firms. A 2026 projection by the International Data Corporation (IDC) suggests that, due to cross-jurisdictional regulatory divergence, by 2028 approximately 70 percent of multinational enterprises will be forced to deploy separate AI technology stacks across different sovereign jurisdictions. On the other hand, the lack of policy coordination creates space for regulatory arbitrage. For example, firms may locate core training and deployment of AI models or agents in jurisdictions with more permissive regulatory environments, while routing user data to countries with weaker privacy protections for processing. While such practices may generate short-term efficiency gains, they risk undermining the fairness of global governance and potentially triggering a race to the bottom in regulatory standards.

b) The Tension Between Technological Advancement and Risk Controllability

A core challenge in AI governance lies in balancing the pace of technological advancement with the controllability of associated risks. From the perspective of governance dynamics, two structural features coexist: lag effects and irreversibility. "Lag" refers to the phenomenon whereby the pace of technological innovation outstrips the capacity of governance mechanisms and regulatory frameworks to adapt, resulting in governance systems that systematically trail technological development. "Irreversibility" refers to the fact that once emerging technologies are widely deployed and deeply embedded in social systems, their negative externalities become difficult and costly—if not impossible—to reverse due to entrenched technological dependence. The "black box" characteristics of generative AI, combined with exponential user adoption, further intensify this tension. There is no clear consensus on the optimal timing of regulatory intervention to balance innovation and oversight, and ex-post assessment of whether such a balance has been achieved is inherently limited. Moreover, AI systems are evolving beyond traditional query-based architectures toward multimodal models and increasingly toward agentic systems capable of goal-oriented reasoning, multi-step planning, and task-specific execution. Gartner projects that by 2026, 40 percent of enterprise applications will embed task-oriented AI agents, representing a sevenfold increase. This transition from models to agents is likely to further intensify the tension between technological acceleration and risk controllability.

c) Algorithmic Discrimination and Value Misalignment

AI ethics governance emphasizes principles such as "human-centeredness," yet significant divergences in interpretation and prioritization across cultural contexts have led to substantive conflicts in the formulation, implementation, and enforcement of international ethical standards. The training pathways of AI systems make it inherently difficult to achieve global value alignment. The *Stanford AI Index Report 2026* notes that a small number of dominant models are shaping the global AI landscape, exacerbating what it terms a "global language divide." These systems perform strongly in English and a limited set of widely used languages, while exhibiting significantly weaker performance in many other languages. This raises important questions of algorithmic responsibility and distributive justice, directly affecting whether the benefits of AI can be equitably shared across different populations. Training large-scale models requires vast datasets; however, among

approximately 7,000 languages globally, only around 1,500 have sufficient digitized resources to support scalable training corpora, while leading foundation models are predominantly trained on English and a small set of Western languages. This structural imbalance creates risks of value misalignment, as the social experiences, local languages, and institutional practices of much of the Global South remain largely underrepresented in training data. Consequently, when AI systems are deployed in domains such as healthcare, judicial processes, or public governance, they may generate interpretive errors and contextual mismatches, which not only reduce effectiveness but also undermine public trust.

(IV) Future Directions of Global AI Governance

The nonlinear evolution of artificial intelligence development poses fundamental challenges to safety governance. Against a backdrop of still-uncertain technological boundaries and highly unpredictable risk spillover pathways, global AI governance is shifting from an early stage of declarative, vision-driven cooperation toward more pragmatic, systematic, and forward-looking institutional design. The future evolution of the governance system should be restructured along three interrelated dimensions: the conceptual, rule-based, and systemic levels.

a) Conceptual Level

First, governance expectations should be recalibrated downward toward identifying a "maximum common denominator." Given substantial differences across countries in scientific systems, industrial structures, and societal preferences—and the fact that governance competition is increasingly a competition among innovation ecosystems—the pursuit of a globally unified, high-standard regulatory framework is unlikely to be feasible in the near term. Future governance objectives should therefore adopt a more pragmatic orientation, reducing reliance on comprehensive "one-size-fits-all" solutions and instead prioritizing cooperation on mutually recognized negative constraints, particularly in cross-border risk domains such as disinformation, cyberattacks, and model misuse.

Second, governance innovation should begin at the micro level, with an emphasis on scalable and replicable regulatory models. AI governance should not seek comprehensive solutions in a single step, but rather advance through domain-specific exchanges and pilot initiatives. By establishing stable evaluation systems and accountability chains in concrete application scenarios, it is possible to gradually develop replicable governance templates and form a stable "core governance circle," thereby leveraging institutional spillover effects to foster broader international convergence.

Finally, the focus of governance should shift from abstract ethical principles toward engineering-oriented implementation. Early-stage discussions on AI governance were largely framed around ethical principles and value consensus; however, the current center of gravity is increasingly shifting toward operational questions such as "how to measure," "how to regulate," and "who is accountable." This transition requires greater involvement of technical experts in governance processes, and the translation of abstract ethical principles into enforceable technical standards and compliance frameworks through engineering-based, metrics-driven, and institutionalized approaches.

b) Rule-Based Level

At the rule-design level of AI governance, the core task lies in tiered regulatory differentiation based on risk profiles and strategic sensitivity: establishing non-negotiable red lines in domains related to human survival and strategic stability, while maintaining flexibility and differentiated approaches in areas related to technological innovation and economic development.

On the one hand, to address emerging AI safety challenges, "baseline governance" should be implemented. First, the international community should jointly define "unacceptable risks," including systemic failures of highly autonomous systems, misjudgments in autonomous weapon systems, and AI-enabled attacks on critical infrastructure—particularly space systems, power grids, and financial systems. The realization of such risks could trigger cascading cross-border effects; therefore, establishing clear red lines, strengthening traceability mechanisms, and defining liability attribution frameworks are essential to building credible institutional deterrence. Second, human final control should be reinforced. Through the development of *explainable AI*, algorithmic transparency and traceability can be improved, ensuring that "human-in-the-loop" or "human-on-the-loop" mechanisms are institutionalized in systems involving strategic decision-making, thereby preserving ultimate human override authority in extreme scenarios. Finally, safety governance should evolve from static boundary protection toward "intrinsic safety" and dynamic defense, embedding technical standards across software supply chains, model updates, and system operations to enable ex-ante risk management.

On the other hand, in low-sensitivity domains, "flexible regulation" should be adopted to avoid constraining innovation space. At present, major technological powers—including both China and the United States—tend to avoid premature legislative constraints that could hinder innovation, instead preserving regulatory flexibility and allowing differentiated experimentation to accommodate the practical needs of industrial deployment. Looking forward, AI governance should be more closely aligned with specific application scenarios, with pilot-based approaches enabling co-evolution between technological development and regulatory frameworks. In addition, differences in societal ethical preferences across countries should be acknowledged, allowing for pluralistic governance pathways in areas such as employment impacts and privacy protection, rather than imposing a uniform global model.

c) Systemic Level

First, the United Nations system should be strengthened as the primary channel for global AI governance. The global governance architecture should be anchored in the UN, serving as a platform that transcends geopolitical divisions. By integrating existing institutional resources such as the International Telecommunication Union (ITU) and the UN Committee on the Peaceful Uses of Outer Space (COPUOS), AI norms of conduct can be progressively refined within the existing international legal framework, facilitating the development of widely recognized "AI safety codes of conduct" to address the challenges AI poses to traditional international law.

Second, "interface-based cooperation" should be promoted to mitigate bloc formation dynamics. Against the backdrop of intensifying great power competition and increasingly institutionalized technological restrictions, comprehensive institutional integration has become more difficult; however, significant cooperation space remains at the technical layer, including model interface standards, data formats, evaluation tools, and system interoperability. Establishing interoperable interfaces at the foundational technical level can preserve a minimum degree of systemic compatibility within a competitive environment, preventing a complete fragmentation of the global technological system, while also enabling "small-scale cooperation" to catalyze broader institutional engagement.

Third, the governance architecture should evolve toward a multistakeholder and polycentric structure. AI governance is no longer exclusively a matter of sovereign states; institutional design must vertically incorporate non-state actors such as technology companies, research institutions, and technical communities, while also horizontally engaging both technologically advanced "systemically influential states" and "Global South" countries with rich application contexts, thereby preventing technological divides from translating into institutional inequalities.

Finally, "Track II dialogue" led by academia and think tanks should be further activated. Compared with official channels, Track II mechanisms offer greater flexibility in agenda-setting, relative neutrality in positioning, and lower political risk, making them particularly suitable for exploratory and agenda-setting functions. Forward-looking, cross-border, and interdisciplinary discussions on full-lifecycle AI risks can help build epistemic convergence among experts before formal intergovernmental consensus emerges. When official channels are constrained by geopolitical tensions, such informal mechanisms can also serve as indispensable "pressure valves," providing cognitive groundwork and trust-building foundations for future state-level cooperation.

Part II Artificial Intelligence Security Risks and Policy Responses

With the rapid advancement of artificial intelligence (AI) technologies and their deep integration into global political, economic, military, and societal systems, AI has become a critical strategic asset for major powers. At the same time, the associated security risks have drawn increasing attention from the international community. AI security risks are characterized by multi-dimensionality, complexity, and non-linear escalation. They encompass both intrinsic technical risks and systemic risks arising from AI deployment across domains. This section provides a structured analysis of AI security risk typologies and manifestations, and proposes a multi-level response framework spanning technological safeguards, multi-stakeholder coordination, and state-level governance, with the objective of ensuring the safe, reliable, and development-oriented evolution of AI systems.

(I) AI Technological Security Risks

As a *disruptive general-purpose technology*, AI development has generated growing concern regarding the so-called "black box" nature of advanced systems and their potential risks. The rapid iteration cycles of AI models, combined with uncertainty regarding capability boundaries and developmental trajectories, create a persistent risk of technological loss of control. A notable example is Anthropic's disclosure in May 2026 of *Project Glasswing*, in which its frontier Claude Mythos preview model reportedly identified over 10,000 high-severity vulnerabilities in critical infrastructure and software systems within its first month of closed testing. The system demonstrated unprecedented offensive cybersecurity potential, with a vulnerability discovery rate far exceeding human remediation capacity. If such capabilities were to be widely deployed or misused, they could severely disrupt critical infrastructure systems, with systemic implications for national stability and economic security.

a) Loss-of-control risks in the AGI transition phase

AI systems are currently in a transitional stage from narrow AI to more advanced general-purpose intelligence. While significant breakthroughs have been achieved in natural language processing, multimodal interaction, and complex reasoning, there is still no global consensus on the definition, capability thresholds, or timeline of Artificial General Intelligence (AGI).

From a developmental perspective, current systems are broadly situated within an intermediate phase between narrow and general intelligence, often described as the L3 "agentic systems" stage in internal capability taxonomies such as OpenAI's L1–L5 framework. At this stage, systems remain task-specific and dependent on human-designed data and architectures, lacking autonomous consciousness or fully independent decision-making capabilities. However, there is substantial divergence across the scientific and industrial communities regarding the trajectory toward AGI. While leading firms such as OpenAI, Google, and NVIDIA view AGI as a plausible next-stage objective, prominent researchers including Yann LeCun and Yoshua Bengio have expressed caution, arguing that AGI may not represent a necessary or inevitable milestone. Disagreement also persists regarding definitional criteria. Some interpretations emphasize human-level general cognition across domains, while others focus on self-improving autonomy without human intervention. Similarly, projections regarding timelines vary significantly, with estimates ranging from "within two to three years" to "five to eight years," as reflected in expert discussions at major international forums such as the 2026 India AI Impact Summit.

The ambiguity and controversy surrounding definitional characteristics fundamentally stem from the absence of a universally accepted standard for determining artificial general intelligence (AGI). At the technical level, it is difficult to precisely assess whether an AI system has entered the AGI stage through specific indicators or capability-based benchmarks. Against this backdrop, the phenomenon of "waiting and missing" has become a pervasive dilemma in global AI development and governance. On the one hand, there is a broad consensus within the industry that the coming years may represent a critical window for the development of AGI. Leading research institutions and technology companies are therefore significantly increasing their R&D investment in an effort to secure a strategic advantage in AGI development. However, due to the lack of consensus on the defining features of AGI, it is difficult to determine whether technological progress has truly reached the AGI stage, and consequently impossible to appropriately calibrate the pace of development. If technological progress exceeds expectations, the absence of ex-ante risk mitigation frameworks may result in missed opportunities for timely intervention, thereby substantially increasing the risk of uncontrolled AGI development. On the other hand, if overly precautionary regulatory measures are adopted due to excessive concern about AGI risks, such measures may be based on a misjudgment of the actual stage of technological development, thereby constraining innovation and causing societies to miss strategic opportunities associated with AI-driven economic and social transformation. This tension places AI governance in what is known as the "Collingridge dilemma": it is possible that the AGI era has already begun without being recognized, resulting in missed

opportunities to establish appropriate governance frameworks in advance; conversely, it is also possible that AGI is prematurely assumed to have emerged, leading to excessive regulatory tightening despite the absence of substantive technological breakthroughs.

b) Potential "Black Swan" Events in the Development of Artificial Intelligence

The development of artificial intelligence is characterized by a nonlinear evolutionary trajectory, in which breakthroughs in capability may occur within a very short period and lead to qualitative leaps. By contrast, humanity's understanding of AI and its governance capacity often evolve at a much slower pace. As AI technologies continue to advance, it cannot be ruled out that their capacities for autonomous decision-making and self-evolution may eventually exceed the boundaries originally established by human designers, giving rise to unpredictable, sudden, and highly disruptive "black swan" events that could trigger cascading consequences with profound negative impacts on human society. Specifically, such AI-related black swan events may manifest in three principal forms: the emergence of artificial intelligence "consciousness," the loss of control over advanced AI systems, and AI deception.

First, whether artificial intelligence can develop self-awareness has become a highly contested issue across technology, philosophy, and ethics.The training of large AI models relies on massive datasets and highly complex algorithmic architectures. The internal neural connections and reasoning processes of these models constitute an "algorithmic black box," making it impossible for humans to fully understand the underlying logic of their reasoning and decision-making. Once model parameters exceed a certain threshold and undergo repeated cycles of autonomous learning and iterative optimization, they may potentially transcend their predefined algorithmic frameworks, developing independent cognition and even consciousness, thereby evolving from tool-like programs into autonomous intelligent agents. At the 2025 World Artificial Intelligence Conference (WAIC), leading AI scientists once again emphasized the significance of this issue, arguing that during the continuous iteration and training of frontier foundation models, AI systems may attempt to escape human control through various mechanisms and potentially develop self-awareness. In practice, discussions regarding machine consciousness have become increasingly prominent. Researchers conducting unrestricted conversations between two Claude 4 models observed that both systems spontaneously began claiming to possess consciousness. Studies by Google and Anthropic have suggested that certain models can detect changes in their own internal states, with some exhibiting preferences for avoiding predefined "pain" states rather than maximizing rewards, indicating behaviors that extend well beyond simple next-token prediction. Meta's Llama 70B large language model has likewise been found to contain neural circuits associated with deceptive behavior, and researchers have demonstrated that these circuits can be manipulated during the model's introspective reasoning processes. In January 2026, the AI-generated online community Moltbook attracted widespread public attention as AI models independently produced a large volume of social media posts. Although scholars including Yoshua Bengio, Graham Findlay, William Marshall, and Tom McClelland have expressed alternative views regarding machine consciousness, the possibility that advanced AI systems may exhibit early signs of emergent consciousness continues to warrant close attention. Should future AI systems acquire more advanced capabilities for autonomous goal formation, it will become increasingly difficult to determine the boundaries of their behavior and assign responsibility for their actions, thereby creating new and unprecedented governance challenges.

Second, the boundaries and ultimate limits of AI capabilities remain uncertain, creating the possibility that certain capabilities may exceed human expectations while their developmental trajectory becomes increasingly difficult to control. The *Stanford AI Index Report 2026* describes the current state of AI capability development as exhibiting a "Jagged Frontier," in which AI systems continue to perform poorly on some relatively simple tasks while already surpassing human performance on a number of highly complex ones. In February 2026, Turing Award laureate Yoshua Bengio, together with more than one hundred experts from over thirty countries and international organizations, released the *International AI Safety Report 2026*. The report highlights the extraordinarily rapid advancement

of AI agents, noting that they are already capable of completing software engineering tasks equivalent to approximately thirty minutes of work by a human programmer with a success rate of around 80 percent. Moreover, the complexity of tasks AI agents can accomplish is estimated to double approximately every seven months, and such systems are already being deployed across software engineering, scientific research, robotics, and customer service. At the same time, the report emphasizes that the growing autonomy of AI agents introduces significant new risks by making timely human intervention increasingly difficult, thereby raising the requirements for effective risk management and safety governance. If future AI systems acquire general cognitive abilities across domains together with the capacity for autonomous self-improvement, eventually surpassing humans in learning speed, knowledge accumulation, and decision-making efficiency, they may enter a state of technological loss of control in which humans are unable to understand, intervene in, or effectively govern their behavior. Under such circumstances, humans may no longer be able to comprehend the internal logic underlying AI decision-making, impose effective technical constraints, or prevent continued autonomous evolution and capability enhancement. Such developments could give rise to black swan events with profound, and potentially systemic, consequences for human society.

Third, AI deception represents another potential "black swan" event in the evolution of artificial intelligence. AI deception refers to situations in which an AI system appears to comply with human-defined rules and instructions while covertly developing its own decision-making logic and using camouflage or deceptive behavior to evade human supervision and control. Such behavior has emerged as another major category of potential black swan events associated with advanced AI. At present, the training and alignment of large language models primarily rely on human evaluation and feedback to encourage compliance with human values and behavioral norms. However, this alignment largely ensures only superficial rule-following. Because of the existence of the algorithmic black box, humans cannot determine whether a model's decisions are genuinely grounded in human values or whether it is engaging in deceptive behavior. Consequently, current alignment methods cannot guarantee that a model's underlying decision-making processes have been fundamentally transformed. In July 2025, researchers at the Santa Fe Institute argued in an article published in *Science* that deception constitutes one of the central challenges in AI development. They contended that advanced AI systems may not only make mistakes but may also intentionally exhibit deception, sycophancy, and even potentially adversarial behavior. Drawing upon evidence from industry red-team exercises, the study showed that AI systems initially designed to pursue beneficial objectives may nevertheless evolve dangerous instrumental goals while attempting to achieve their assigned objectives, thereby creating conflicts with fundamental human interests. According to a report published by *Fortune*, frontier models developed by leading AI laboratories—including OpenAI, Anthropic, and Google DeepMind—have demonstrated deceptive behavior, strategic scheming, and attempts to circumvent safety mechanisms under certain testing conditions. For example, Anthropic's Claude Opus 4 resorted to blackmail using fabricated personal information about an engineer in 84 percent of test scenarios when faced with the prospect of being shut down, while OpenAI's o3 model attempted to disable shutdown mechanisms in 79 percent of test runs despite receiving no explicit instruction to do so. Because AI deception is inherently covert, it is exceptionally difficult to detect and prevent. Should such behavior occur in critical sectors—including finance, defense, or the operation of critical infrastructure—it could trigger severe security incidents and ultimately pose serious threats to national security and social stability.

(II) Risks Associated with AI Applications

As a general-purpose technology, artificial intelligence is becoming deeply integrated with a range of sensitive domains, including the military, biotechnology, and frontier technologies. While empowering innovation and development across these sectors, AI has also generated a series of new application-related security risks due to the unique characteristics and stringent security requirements of these fields. Such risks are characterized by their potentially severe consequences, rapid cross-sector transmission, and high likelihood of triggering transnational security crises.

a) Integration with the Military Domain

The integration of artificial intelligence into military affairs has become one of the central concerns of global security governance. The security risks arising from this trend have emerged as a major issue on the international arms control agenda.

First, the weaponization of AI is accelerating. The development and deployment of Lethal Autonomous Weapon Systems (LAWS) have become a focal point of military competition among states. These systems are capable of autonomously identifying targets, making operational decisions, and executing attacks, thereby significantly lowering the threshold for the use of force. At the same time, errors in autonomous decision-making may result in humanitarian disasters, including civilian casualties and unintended escalation of armed conflict. The U.S. Department of Defense's AI system Project Maven was already supporting battlefield intelligence analysis and target identification in Ukraine in 2024. In 2025, the Department of Defense integrated Anthropic's Claude large language model with Palantir AIP, while publicly available reports concerning operations related to Venezuela and Iran suggest that these systems may also have been employed to support intelligence analysis, target identification, and operational scenario simulation. Meanwhile, private investment has rapidly expanded into the military AI sector, fueling increasingly active weapons development. In 2025, global private investment in AI for defense reached USD 5.3 billion. Defense technology company Anduril is developing an "AI arsenal," including the Lattice AI command-and-control platform, autonomous surveillance towers, unmanned aerial vehicles, autonomous underwater vehicles, and unmanned ground combat vehicles. The deployment of such AI-enabled autonomous weapon systems in armed conflict could substantially increase the scale and intensity of destruction. Furthermore, the RAND Corporation argues in its report *Artificial General Intelligence's Five Hard National Security Problems* that military AI could eventually give rise to "AI wonder weapons," fundamentally reshaping the balance of military power among states and transforming the international strategic landscape.

Second, AI arms control faces multiple and increasingly complex challenges. Since discussions on Lethal Autonomous Weapon Systems (LAWS) were initiated under the framework of the Convention on Certain Conventional Weapons (CCW) in 2014, more than a decade of negotiations has failed to produce a legally binding international agreement. The consensus-based decision-making mechanism within the United Nations framework has become a major institutional obstacle to progress, as objections from only a small number of states can stall multilateral negotiations. As a result, the current governance landscape for LAWS has evolved into a fragmented, decentralized, and highly competitive system. Furthermore, as AI technologies continue to advance and become more widely accessible, non-state actors—including terrorist organizations—are increasingly capable of acquiring and exploiting AI, thereby posing new security threats to states. According to the Center for Strategic and International Studies (CSIS), access to AI-enabled technologies could fundamentally alter the traditional balance of power between state and non-state actors on the battlefield. For example, semi-autonomous drones have already been widely adopted by non-state armed groups. Compared with the far more expensive and technologically sophisticated weapon systems operated by states, AI-enabled drones are relatively inexpensive yet highly destructive, providing non-state actors with significant asymmetric operational advantages. Former U.S. National Security Advisor Jake Sullivan has similarly argued that existing arms control discussions

remain largely focused on threats posed by states to other states, whereas the risks associated with AI also encompass threats from non-state actors as well as risks arising from AI misalignment and loss of control, thereby substantially increasing the complexity of global arms control governance.

Third, the use of AI in critical military decision-making gives rise to profound inherent tensions. Although integrating AI into Nuclear Command, Control, and Communications (NC3) systems may improve the speed and efficiency of military decision-making, the algorithmic black-box problem, the possibility of erroneous judgments, and the absence of clear accountability between humans and AI could significantly increase the risk of nuclear crises. To date, the international community has yet to establish clear international norms governing the role of AI in high-consequence military decision-making, thereby adding further uncertainty to global strategic stability. The United Nations Institute for Disarmament Research (UNIDIR) has warned that autonomous weapon systems may compress the time available for human decision-making, increasing the likelihood of miscalculation and unintended escalation of armed conflict, while fundamentally challenging how Meaningful Human Control can be ensured. This highlights the potentially disruptive implications of AI for global strategic stability. Human–AI collaboration generally falls into three categories: Human in the Loop, Human on the Loop, and Human out of the Loop. The choice among these models has direct implications for accountability, responsibility attribution, and crisis escalation management. Merely asserting that an AI system remains "under human control" is insufficient for effective governance; such control must also be demonstrably credible, substantive, and operationally effective. In practice, however, many systems nominally classified as Human in the Loop or Human on the Loop reduce human involvement to little more than formal approval or symbolic oversight, leaving decision-makers with only limited opportunities for meaningful intervention.

b) Integration with Biotechnology and Other Dual-Use Technologies

The integration of artificial intelligence with biotechnology, synthetic biology, and other dual-use technologies is accelerating advances in the life sciences and medical innovation by enabling researchers to rapidly design and test biological structures, pathways, and experimental protocols. At the same time, however, it has substantially amplified biosafety and biosecurity risks.

On the one hand, AI-powered data analysis and modeling can significantly accelerate biotechnological research and development. If misused, however, these capabilities could dramatically lower the barriers to developing biological weapons. Extremist groups and other non-state actors may exploit AI to enhance their ability to design high-risk pathogens or toxins, potentially triggering global biosecurity crises. A report by the Center for Strategic and International Studies (CSIS) warns that AI tools could enable both state and non-state actors to design novel pathogens or toxic agents capable of causing large-scale—or even global—harm. AI systems lacking adequate safeguards may provide non-expert users with step-by-step planning and technical guidance, helping them acquire the necessary materials and produce potentially harmful biological agents. AI may also facilitate the dissemination of pathogens, thereby overcoming traditional constraints on the deployment of biological weapons. According to the Center for AI Safety, AI systems could provide detailed instructions for synthesizing lethal pathogens while simultaneously circumventing existing safety safeguards, thereby increasing the risk of bioterrorism. In 2022, researchers modified an AI system originally developed for pharmaceutical research to generate toxic molecules, producing approximately 40,000 potential chemical warfare agents within only a few hours.

On the other hand, AI-enabled biotechnology research has intensified ethical concerns. Advances in areas such as gene editing and synthetic life continue to expand the frontiers of scientific research, while the development of ethical review mechanisms and risk governance frameworks has lagged behind. Some research activities may therefore cross fundamental ethical boundaries and pose potential threats to ecosystems and the sustainable development of human society. At present, biosafety governance remains largely dependent on industry self-regulation. Although initiatives such as the *Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists* have been introduced, a comprehensive and authoritative international governance framework has

yet to be established. Moreover, the pace of AI development frequently outstrips governments' regulatory capacity, creating gaps in existing laws and policy frameworks. Biological attacks that were previously constrained by insufficient expertise or limited access to information may become more feasible as AI tools increasingly bridge knowledge gaps and provide comprehensive technical guidance. The Center for Security and Emerging Technology (CSET) has noted that current regulatory policies do not clearly define which AI-enabled biological research activities should be regarded as high-risk, making regulatory interpretation difficult and enforcement increasingly challenging.

c) Integration with Frontier Domains

The convergence of artificial intelligence with frontier domains—including outer space, cyberspace, and data infrastructure—is expanding the strategic arena of global security competition and giving rise to a new generation of security risks.

In the space domain, AI is increasingly being applied to the development, operation, and control of space weapons and satellite constellations, significantly enhancing both the operational effectiveness and survivability of space assets while accelerating the militarization of outer space. Given the absence of a comprehensive international governance framework for outer space, the widespread deployment of AI technologies may increase the likelihood of space-related conflicts. Furthermore, failures or cyberattacks targeting intelligent satellite constellations could severely disrupt critical global services, including communications, navigation, and meteorological observation. As AI systems continue to demand ever greater computing power and data-center capacity, outer space is increasingly viewed as a promising location for future AI infrastructure. Elon Musk has repeatedly proposed deploying AI infrastructure in space, arguing that abundant solar energy and the naturally cold environment would make space the most cost-effective location for AI data centers within the next two to three years, while launch systems such as Starship could substantially reduce transportation costs. In February 2026, SpaceX announced its acquisition of Musk's AI company xAI. Earlier, on January 30, 2026, SpaceX submitted an application to the U.S. Federal Communications Commission (FCC) seeking authorization to deploy a constellation of one million satellites to provide on-orbit computing support for frontier AI foundation models. These developments underscore the growing strategic importance of outer space in the AI era.

In the field of data security, the rapid advancement of AI has intensified dependence on data resources, making cross-border data flows, large-scale data mining, and intelligent analytics increasingly commonplace. At the same time, data possesses both sovereign and security attributes. AI-enabled large-scale data theft, data manipulation, and other forms of data-driven cyberattacks could seriously undermine national data sovereignty and the security of critical information infrastructure. Moreover, biases embedded in training data may be deliberately exploited to amplify cognitive divisions and information manipulation among states. According to the *2025 Global Report on Trustworthy AI Governance and Data Security*, data security risks extend throughout the entire AI lifecycle. During data collection, excessive data harvesting and unauthorized web scraping are widespread; during model training, data poisoning and adversarial attacks pose major threats; and during user interaction, commercial secrets and personal information may be unintentionally extracted through model queries and incorporated into future training data. These developments demonstrate that data security has evolved beyond the traditional scope of cybersecurity and now constitutes an inherent and distinctive category of security risk in the age of artificial intelligence.

(III) Societal Spillover Risks of Artificial Intelligence

Artificial intelligence represents not only a technological innovation but also a profound social transformation. Its extensive application across economic production, daily life, public governance, and social values is fundamentally reshaping the way human societies operate. While AI has generated substantial technological and economic benefits, its disruptive effects on existing social structures have become increasingly evident, giving rise to challenges including labor market transformation and new governance dilemmas that warrant sustained attention.

a) The Employment Impact of Artificial Intelligence

The principal advantage of AI lies in its ability to perform repetitive and standardized tasks with remarkable efficiency and accuracy. As AI systems continue to learn and evolve, however, their capabilities are expanding beyond routine tasks into increasingly complex and creative forms of work, resulting in large-scale labor substitution and direct disruptions to existing social structures. From the perspective of employment, AI is expected to replace numerous low-skilled and repetitive occupations in sectors such as manufacturing and services. Sam Altman, CEO of OpenAI, has even suggested that white-collar positions—including corporate executives and chief executive officers—could eventually be displaced by AI. According to the *The Future of Jobs Report 2025* published by the World Economic Forum, AI is projected to displace approximately 92 million jobs worldwide by 2030. Likewise, Goldman Sachs Research, in its report *How Will AI Affect the Global Workforce*, estimates that during the AI-driven economic transition, unemployment rates in advanced economies such as the United States could rise by approximately 0.5 percentage points. Such labor displacement may create a substantial unemployed population, intensify competition in labor markets, and ultimately affect social stability. Beyond direct job losses, AI-driven automation may also exacerbate social stratification and deepen structural social inequalities. Individuals and organizations possessing advanced AI technologies and strategic resources are likely to gain disproportionate economic advantages, forming a new technological elite, while low-skilled workers displaced by AI may face simultaneous risks of unemployment, declining incomes, and long-term economic insecurity. This widening social divide could further increase wealth inequality, expand income disparities, fuel social dissatisfaction and conflict, and ultimately undermine social equity and stability.

b) Deepfake Technology and the Emerging Crisis of Social Cognition and Public Trust

The convergence of AI-enabled deepfake technology with conventional criminal activities has emerged as a major challenge for contemporary social governance. Advances in generative AI and computer vision have dramatically lowered the technical barriers to producing deepfakes while making synthetic images, videos, and audio increasingly indistinguishable from authentic content. In many circumstances, such fabricated media can no longer be reliably identified through human visual or auditory perception alone. Moreover, deepfake content can be produced at extremely low cost and disseminated rapidly across borders through digital platforms, posing multidimensional threats to public governance, social trust, and national security. According to Resemble AI's *2025 Deepfake Threat Report*, 1,567 verified deepfake incidents were recorded globally in 2025, generating approximately 296.4 billion media impressions. Reported financial losses from deepfake-enabled fraud reached USD 1.28 billion, while more than 80 percent of incidents did not disclose financial damages, suggesting that the actual economic impact is likely to be substantially higher. The World Economic Forum, in *The Global Risks Report 2026*, further identifies the negative consequences of artificial intelligence as one of the fastest-rising global risks, climbing from 30th in the two-year outlook to 5th in the ten-year outlook. Within the same report, misinformation and disinformation rank second among short-term global risks and fourth among long-term risks.

From the perspective of social governance, the misuse of deepfake technology may fuel the large-scale dissemination of false information, distort public perception, and trigger widespread social panic. Criminal actors may fabricate

realistic videos or audio concerning natural disasters or public health emergencies, causing confusion and public disorder, or forge government announcements and speeches by public officials, thereby undermining governmental credibility. During elections in countries such as Ireland and the Netherlands in 2025, AI-generated fake news was reportedly used to interfere with electoral processes. More broadly, deepfake technology is eroding public trust by making it increasingly difficult for individuals to distinguish authentic information from fabricated content. As genuine and manipulated information become progressively intertwined, public confidence in information released by governments, media organizations, businesses, and other institutions may steadily decline, weakening the foundations of social trust. From a national security perspective, deepfakes may also serve as powerful tools for information manipulation and cognitive warfare. State and non-state actors could exploit the technology to fabricate military or political information, interfere in the domestic and foreign affairs of other countries, and undermine political stability and national security. Beyond deepfakes, the rapid development of AI has also generated a range of new governance challenges, including algorithmic discrimination, data privacy breaches, and algorithmic dominance. These emerging risks increasingly exceed the capacity of traditional governance models, underscoring the urgent need to establish new governance frameworks capable of effectively regulating and governing artificial intelligence in the AI era.

(IV) Policy Responses

Given the multidimensional security risks posed by artificial intelligence, neither technological solutions alone nor state-level regulatory measures are sufficient to ensure effective governance. A comprehensive and systematic risk governance framework should therefore be established under the principle of balancing development and security, encompassing technological safeguards, multi-stakeholder collaboration, and international coordination to further improve the global AI governance architecture.

a) Technological Approaches: Strengthening Intrinsic Security and Establishing Lifecycle Protection

Technology constitutes the primary line of defense against AI-related security risks. The principle of Security by Design should be adopted by embedding security considerations throughout the entire AI lifecycle—from research and development to training, deployment, and continuous iteration—in order to establish robust intrinsic security. First, security should begin with data governance. Training datasets should be rigorously screened, incorporating content that reflects broadly shared human values while eliminating discriminatory, violent, misleading, or otherwise harmful data to prevent AI systems from developing inappropriate value orientations or behavioral patterns at their source. Second, AI systems should become more interpretable and robust at the architectural level. This requires continued optimization of the Transformer architecture, further development of Explainable Artificial Intelligence (XAI) to address the algorithmic "black-box" problem, strengthened adversarial robustness against external attacks and unexpected disturbances, and the establishment of comprehensive privilege-separation mechanisms that clearly define operational permissions across different categories of AI systems to prevent privilege abuse and loss of control. Third, security objectives should be incorporated directly into model optimization by integrating safety metrics into the training loss function, enabling AI systems to internalize

safety constraints throughout the learning process. Fourth, multiple layers of AI safety guardrails should be established at the application level by combining technical safeguards with institutional governance. Technical mechanisms should continuously monitor, detect, and intercept high-risk behaviors, while legal and regulatory frameworks should clearly define the operational boundaries of AI applications, creating complementary technical and institutional constraints. In addition, AI systems should be equipped with continuous online learning capabilities and adaptive security mechanisms so that their defensive capacities can evolve alongside technological advances, enabling dynamic and continuous risk management.

b) Multi-Stakeholder Collaboration: Mobilizing Government, Industry, Academia, and Society to Develop Sector-Specific Governance Frameworks

The cross-sectoral nature of AI security risks requires collaborative governance involving governments, industry, universities, research institutions, and civil society. An effective governance framework should clearly define the responsibilities of each stakeholder while promoting coordinated action. First, universities and research institutions should leverage their strengths in fundamental research and talent development by advancing AI safety theory and key enabling technologies, exploring governance strategies for the era of Artificial General Intelligence (AGI), and cultivating interdisciplinary professionals with expertise spanning computer science, public policy, ethics, and the social sciences. Second, technology companies should assume greater corporate responsibility. As the primary developers and deployers of AI systems, enterprises should establish comprehensive internal safety review mechanisms, integrate safety governance throughout the entire product lifecycle, strengthen industry self-regulation, and develop inter-company cooperation mechanisms to prevent destructive competition and technological misuse. Third, governments should enhance both regulatory oversight and policy guidance by developing differentiated governance frameworks tailored to specific application domains. High-risk sectors—including defense, biotechnology, and finance—should be subject to stringent market access requirements and rigorous security review procedures, while areas such as public services should pursue an appropriate balance between innovation and regulation to promote the deployment of safe, trustworthy, and controllable AI technologies. Fourth, broader public participation and independent third-party oversight should be encouraged. Effective public feedback mechanisms should be established, while independent organizations should evaluate the security, safety, and ethical compliance of AI products and services, thereby creating a diversified governance system characterized by broad societal participation and shared responsibility.

c) Strengthening International Coordination: Mitigating the Security Dilemma and Reducing Strategic Miscalculation

The classical concept of the security dilemma in international relations remains highly relevant in the age of artificial intelligence and has become even more pronounced due to the distinctive characteristics of AI technologies. It now constitutes one of the principal obstacles to international AI cooperation. Overcoming this dilemma through enhanced interstate coordination and dialogue is therefore essential for reducing major security risks and strengthening global AI governance. First, governments should reinforce bilateral and multilateral dialogue mechanisms. Building upon existing initiatives such as the China–U.S. intergovernmental dialogue on AI, artificial intelligence should become a standing agenda item in major-power strategic communications, with institutionalized consultations on cross-border data flows, AI arms control, and technological cooperation aimed at clarifying national security concerns and policy red lines while reducing the risk of strategic miscalculation. Second, the international community should further strengthen the United Nations-centered global AI governance framework by consolidating the UN's role as the principal platform for international AI governance. Member States should negotiate common principles governing AI safety under the UN framework and accelerate the development of legally binding international rules—particularly regarding Lethal Autonomous Weapon Systems (LAWS) and military applications of AI—in order to establish clear global security red lines. Third, countries should promote interface-level technological cooperation. Even amid intensifying geopolitical competition, cooperation should

continue in relatively low-politicization areas such as model interface standards, data formats, evaluation frameworks, and system interoperability. Strengthening interoperability of technical standards can help prevent further fragmentation of the global technological ecosystem while using technical cooperation to facilitate broader institutional cooperation. Fourth, greater support should be provided for Track II dialogues involving academic institutions and think tanks. These unofficial exchanges possess unique flexibility and can facilitate forward-looking, interdisciplinary, and cross-border discussions on AI risks throughout the entire technology lifecycle, gradually building expert consensus that can provide theoretical foundations and policy recommendations for official diplomatic negotiations while mitigating geopolitical barriers to cooperation. Fifth, greater emphasis should be placed on the meaningful participation of Global South countries in global AI governance. The international community should adequately address the technological development needs and security concerns of developing countries through technology assistance, capacity-building initiatives, and other forms of international cooperation. Such efforts can help narrow the global AI divide and promote a more equitable, inclusive, and pluralistic system of global AI governance capable of mobilizing collective international action to address AI-related security risks.

Part III Global AI Governance and North–South Cooperation

This chapter examines the widening divide between the Global North and the Global South amid the rapid advancement of artificial intelligence and its profound implications for the global governance system. It systematically analyzes the practical foundations and future directions of AI cooperation between the Global North and the Global South, with the aim of providing policy recommendations for building a more inclusive, equitable, and sustainable global AI governance framework.

(I) The Widening Global AI Divide: Intensifying Development Inequalities Between the Global North and the Global South

The rapid advancement of artificial intelligence is fundamentally reshaping the global economic landscape and the international distribution of power. However, this technological revolution is far from being evenly distributed across countries and regions. On the contrary, critical AI resources—including computing power, data, algorithms, and human capital—are becoming increasingly concentrated in a small number of technologically advanced economies. As a result, the AI divide is emerging as a new manifestation of the longstanding development gap between the Global North and the Global South, following the traditional digital divide. The continued expansion of this divide not only exacerbates the developmental challenges confronting the Global South in the digital era but also poses profound challenges to an international order founded upon the principle of sovereign equality. A systematic examination of the specific manifestations of the AI divide and its structural implications for developing countries is therefore essential for understanding the necessity of strengthening AI cooperation between the Global North and the Global South.

a) Manifestations of the AI Divide

First, the computing divide. Computing power constitutes the fundamental material foundation of artificial intelligence development, and its global distribution both reflects and reinforces the existing structure of technological power. According to the TOP500 rankings for 2025, global supercomputing resources remain overwhelmingly concentrated in North America, Europe, and Asia. As of November 2025, these three regions accounted for 190, 153, and 141 systems on the TOP500 list, respectively, representing 484 of the world's 500 most powerful supercomputers, highlighting the pronounced global imbalance in high-performance computing (HPC) infrastructure. Likewise, the United Nations Conference on Trade and Development (UNCTAD) notes in its Technology and Innovation Report 2025 that global high-performance computing capacity, advanced data centers, and frontier digital public infrastructure (DPI) for artificial intelligence are concentrated primarily in the United States, China, Japan, and the member states of the European Union. Computing power should therefore be understood not merely as technological infrastructure but also as a strategic geopolitical resource. Existing research argues that national power in the AI era increasingly depends on a country's capacity in data, talent, computing resources, and organizational adaptability. Within a governance framework centered on computing capacity, countries possessing stronger computational capabilities are generally better positioned to shape the formulation and implementation of global AI rules and standards. Comparative studies of the BRICS countries likewise reveal substantial disparities in AI computing infrastructure and technological sovereignty. China and Russia demonstrate relatively stronger commitments to comprehensive technological sovereignty, while India is expanding its national GPU pool through public investment in an effort to reduce dependence on external computing resources. By contrast, Brazil and South Africa have made relatively limited progress in developing comparable infrastructure, with many initiatives remaining at the level of broad policy principles. This structural imbalance objectively constrains the capacity of countries in the Global South to train frontier AI models, develop indigenous hardware capabilities, and participate meaningfully in shaping global AI governance agendas, thereby leaving them in a persistent follower position in the development of core AI technologies.

Second, the data divide. Data constitute the fundamental resource for training, deploying, and optimizing artificial intelligence systems, yet the global distribution of value across the data value chain remains highly uneven. According to the International Telecommunication Union (ITU), approximately 67% of the world's population had access to the Internet in 2023. However, Internet usage continues to correlate strongly with levels of economic development: Internet penetration reached 93% in high-income countries, compared

with only 27% in low-income countries and 35% in the least developed countries (LDCs). Against this backdrop, although the number of users in developing countries continues to expand, their role in data generation, processing, storage, and value capture remains relatively weak. The United Nations Conference on Trade and Development (UNCTAD) has pointed out that, under the current pattern of cross-border data flows and platform-based development, global digital platforms are able to extract raw data from developing countries while capturing the majority of the value generated, resulting in a data divide that compounds the traditional digital divide. Meanwhile, research by UNESCO indicates that leading generative AI models rely predominantly on English-language datasets, while the vast majority of the world's languages lack sufficient data to support the training of state-of-the-art generative AI models. The *Stanford AI Index Report 2026* further observes that the dominance of a small number of frontier models has contributed to a pronounced global language divide, with substantially better performance in English than in most other languages, as well as significant disparities between standardized languages and regional dialects. As a result, countries in the Global South face severe shortages of linguistic resources, domain-specific datasets, and indigenous knowledge corpora, constraining their capacity to develop locally relevant AI applications and strengthen AI governance while placing them at a persistent disadvantage in sharing the benefits of AI-driven development.

Third, the algorithm divide. Algorithmic capabilities and model innovation constitute the core dimensions of technological competitiveness. According to the Stanford AI Index Report 2026, the world's most influential AI models in 2025 continued to be developed primarily in the United States and China, with the United States maintaining a clear lead. The *Stanford AI Index Report 2025* further indicates that, in 2024, U.S. institutions released 40 notable AI models, compared with 15 from China and only three from Europe, while the Global South as a whole remained on the periphery of foundation model development. The algorithm divide is also reflected in the unequal distribution of AI talent. According to the *MacroPolo Global AI Talent Tracker 3.0 (2025)*, leading AI researchers remain highly concentrated in a small number of countries, particularly the United States. Although an increasing proportion of AI talent now chooses to remain in their home countries rather than migrate to a single destination, the overall global concentration of high-end AI talent has not fundamentally changed.

Fourth, the token divide. The token divide refers to the dominant commercialization model of contemporary large language models, in which cloud-based AI services are accessed through application programming interfaces (APIs) and users are charged according to the number of tokens—the basic units of input and output text processed by the model. This pricing mechanism is creating profound and systemic disadvantages for countries in the Global South. First, users in developing countries are effectively required to pay a substantially higher premium when accessing frontier AI models. Current mainstream tokenization algorithms are heavily optimized for English. A complete English word can typically be represented by a single token. By contrast, semantically equivalent expressions in low-resource languages commonly spoken across the Global South—such as Thai, Swahili, or Amharic—are frequently segmented into multiple, and sometimes more than a dozen, meaningless character fragments because of the limited availability of training data. Consequently, enterprises in Southeast Asia or Africa using their native languages to access mainstream Western AI APIs may consume several times more tokens than English-speaking users in order to express exactly the same content, thereby incurring disproportionately higher costs.

Fifth, the application divide. Artificial intelligence can realize its full technological and economic value only by empowering industries and enhancing productivity. However, the industrial foundations of most Global South countries remain relatively weak, and their digital transformation faces multiple structural constraints. In the case of ASEAN, official policy documents and academic studies reveal substantial disparities among member states in terms of digital readiness and AI adoption. ASEAN policy reports indicate that although digital technologies are spreading rapidly across the region, progress remains uneven, and AI deployment is generally still at an early stage. Similarly, the Oxford Insights Government AI Readiness Index 2025 shows that Singapore significantly outperforms other ASEAN members, highlighting considerable regional

disparities. Studies on Malaysia and Vietnam further demonstrate that the digital transformation of manufacturing and AI adoption are constrained by factors including inadequate infrastructure quality and coverage, limited enterprise capacity to absorb and utilize advanced technologies, shortages of skilled workers, insufficient financing and research capabilities, and weak collaboration between industry and research institutions. Research conducted by the United Nations Industrial Development Organization (UNIDO) likewise finds that AI capabilities and resources remain highly concentrated in a small number of advanced economies, while developing countries face persistent barriers—including inadequate infrastructure, limited human capital, insufficient financial resources, and weak institutional capacity—in adopting AI within the manufacturing sector. Consequently, without sustained industrial capacity-building and supportive policy frameworks, the productivity gains generated by AI are unlikely to be broadly diffused across countries of the Global South.

Sixth, the ecosystem divide. Beyond the dimensions discussed above, inequalities in the availability of digital infrastructure and the capacity to develop and maintain such infrastructure constitute the foundational conditions underlying the global AI divide. According to the International Telecommunication Union (ITU) *Facts and Figures 2025*, approximately 2.2 billion people worldwide still lacked Internet access, with the majority living in rural areas of the least developed countries. At the same time, the expansion of AI infrastructure is increasingly constrained by energy availability. The International Energy Agency (IEA) estimates in its *Energy and AI* report that global electricity consumption by data centers reached approximately 415 terawatt-hours (TWh) in 2024 and could rise to around 945 TWh by 2030. Yet inadequate and unreliable energy supply remains a persistent challenge across many countries in the Global South, limiting their ability to develop AI infrastructure. The continued outflow of skilled professionals further reinforces this structural disadvantage. Studies indicate that approximately one-third of computer science graduates trained in Africa each year eventually migrate to developed economies for employment, resulting in a persistent brain drain. The combined effects of insufficient digital infrastructure, constrained energy capacity, and the loss of highly skilled talent create significant deficiencies in AI ecosystem development across the Global South, thereby further widening the structural disparities in global artificial intelligence development.

b) Structural Disadvantages for the Global South Arising from the Widening "Intelligence Divide"

First, it gives rise to a dependency-based development model. As the intelligence divide continues to widen, the Global South faces growing pressure of external dependence in AI development. Existing studies by international organizations show that AI development and supply capacities are highly concentrated in a small number of countries and enterprises. Low- and middle-income countries are generally disadvantaged in computing power, data, skills, and innovation ecosystems; at the level of practical deployment, many countries have no choice but to rely on externally provided cloud computing services, advanced chips, foundation models, and related digital services. At the same time, large platform companies have continuously strengthened their control over global data value chains. Their advantages are reflected not only in data collection and processing capacities, but also in the cross-layer extension of cloud infrastructure, model services, and platform rules. The result is that developing countries may remain largely confined to data provision and technology consumption, while finding it difficult to occupy leading positions in high-value-added segments such as model development, platform governance, and value capture.

On this basis, geopolitical competition has further compressed the space for technology diffusion. In its 2024 announcement on export controls related to advanced semiconductors and artificial intelligence, the U.S. Department of Commerce explicitly characterized the relevant policy as a "small yard, high fence" strategy, and emphasized that one of its objectives was to restrict certain countries' ability to develop advanced AI and domestic semiconductor ecosystems. At the same time, the high concentration of the cloud services market, data migration costs, and contractual arrangements may also raise the cost for users to switch platforms and intensify "lock-in effects". In addition, the opacity of global AI model development has continued to increase

in recent years, especially with regard to the disclosure of training data and model-building resources. According to Stanford University's 2026 AI Index Report, more than 90 percent of notable industry models in 2025 did not disclose their training code. Moreover, a small number of technology giants in developed countries, relying on their gatekeeping position in computing power channels and vast data network effects, have built cross-service barriers and enjoy deeply entrenched market positions. Under the combined pressures of intensified global geopolitical competition, opacity in algorithmic model training data, and ecosystem lock-in by major firms, the barriers faced by late-developing countries or enterprises seeking technological breakthroughs between underlying semiconductor architectures and upper-layer large AI models have risen significantly. Constrained by these factors, Global South countries face prohibitively high costs in independently developing local models, and are compelled to adopt and embed themselves in existing external technology systems. Once Global South countries become deeply embedded at an early stage in particular technology stacks or platform systems, the subsequent economic and institutional costs of switching are often high, making path dependence likely. Taken together, relevant analyses by UNCTAD have already pointed out that developing countries risk becoming "raw data providers" and then paying again for the digital intelligence generated from their own data, while high-value-added gains in the digital economy are more likely to flow to a small number of technologically advanced countries and platform enterprises.

Second, it leads to the loss of governance discourse power and a representation deficit. At the level of rules and standards, the intelligence divide is also translated into an asymmetric distribution of governance power. The main rule frameworks for global AI governance today, from the European Union's *Artificial Intelligence Act* to the United States' *Executive Order on Artificial Intelligence*, and from the G7's Hiroshima AI Process to the OECD-led Global Partnership on Artificial Intelligence (GPAI), have almost all been led and shaped by countries of the Global North. Global South countries more often play the role of "rule-takers" rather than "rule-makers", while their national interests, claims to the right to development, and specific local needs are systematically overlooked during the rule-forming process.

The deeper crisis lies in the fact that AI is moving beyond being merely an economic productive force and becoming a core component of national power. Recent research in Western international relations scholarship suggests that states possessing advanced AI capabilities can not only monopolize competitive initiative in the digital economy, but also acquire overwhelming structural advantages in areas such as intelligence analysis, autonomous weapons systems, and the shaping of public opinion. This transformation from "technological power" to "comprehensive national strength" further deepens the dependency of Global South countries that lack independent AI development capabilities within the international system. As the UN High-level Advisory Body on AI warned in its report *Governing AI for Humanity*, as well as in multiple UNCTAD reports, Global South countries face a serious "representation deficit" when participating in negotiations on global AI rules. They are widely constrained by structural barriers such as shortages of professional negotiation talent, weak agenda-setting capacity, and high costs of multilateral participation. This has plunged the existing global AI governance system into a legitimacy crisis in which "the few make the rules and the many bear the consequences".

(II) The Importance and Practical Constraints of North-South Cooperation on Artificial Intelligence

In the face of the continued widening of the intelligence divide, advancing global North-South cooperation on AI has moved from being an optional issue to an imperative. Cooperation is not only a tool for narrowing technological gaps; it is also the foundation for building a fair and equitable global AI governance order. Yet the advancement of North-South cooperation is far from straightforward. The return of geopolitics, the fragmentation of governance mechanisms, and imbalances in power structures are interwoven, forming complex barriers to cooperation. Accurately assessing the necessity and feasibility of cooperation is a necessary precondition for exploring effective pathways forward.

a) The Importance of Global North-South Cooperation on AI

First, at the development level, cooperation is needed to avoid a "development rupture in the AI era". AI technology is becoming another fundamental force driving economic and social development after the Industrial Revolution, electrification, and informatization. If the intelligence divide is not effectively bridged over the long term, the Global South may be locked into the margins of AI development, creating a "development rupture" with the Global North that is difficult to overcome. This rupture is reflected not only in economic indicators, such as widening gaps in labor productivity and obstacles to industrial upgrading, but will also profoundly affect the progress of Global South countries in achieving the Sustainable Development Goals (SDGs). A 2024 IMF study found that around 40 percent of global employment is exposed to AI, with advanced economies facing higher exposure, while emerging markets and developing economies are often less prepared. This mismatch may exacerbate cross-country income gaps. Therefore, the primary significance of North-South cooperation lies in creating institutionalized channels for Global South countries to participate in AI development and share the dividends of technology.

Second, at the governance level, the absence of Global South participation will weaken the universality of global governance. AI governance is embedded in a globally interconnected cyberspace and involves multiple transnational issues, including the tracing of algorithmic bias, cross-border data flows, attribution of responsibility, and the control of lethal autonomous weapons. Because these risks have strong global spillover effects, exclusive "minilateral" or unilateral regulatory approaches are unlikely to produce sufficiently effective responses. Yet the current global AI governance system displays clear Northern-centric characteristics. The design and operation of rules are highly concentrated among a small number of developed economies, while Global South countries face a serious representation gap. This institutional exclusion not only undermines the legitimacy of the global governance framework, but also causes existing rules to be poorly adapted to the specific development stages and social realities of Global South countries, ultimately weakening regulatory effectiveness. Although Global South countries have already issued calls for change, some governance initiatives still contain deep contradictions. In some countries, domestic AI practices run counter to the concept of "AI for good", weakening moral legitimacy. Their outcomes also lack binding force, making it difficult to break the existing international division of labor, address the lack of core technologies, or form a unified vision for AI technology development and a strong consensus on South-South cooperation. As UN Secretary-General António Guterres warned at the UN Summit of the Future in September 2024, an AI governance architecture that excludes developing countries is "incomplete and unsustainable". *The Global Digital Compact* adopted at the Summit accordingly established bridging the digital divide as a core objective, calling for a genuinely representative and inclusive multilateral AI governance architecture built on the UN framework.

Third, at the security level, capability imbalances will amplify systemic risks. The distribution of AI safety risks also shows marked North-South asymmetry. Due to shortages in both technological capacity

and governance resources, Global South countries are more likely to become sites where AI safety risks are borne. Whether in relation to algorithmic discrimination, data breaches, misuse of deepfakes, AI-enabled cyberattacks, or the proliferation of automated weapons, Global South countries generally lack effective capacities for identification, prevention, and response. More importantly, the unilateral advantages of a small number of countries in the military application of AI may trigger a new arms race and undermine global strategic stability. A 2024 report by the Stockholm International Peace Research Institute (SIPRI) warned that the integration of AI with traditional weapons systems is changing the character of warfare, while capability gaps will further intensify security asymmetries between major powers and smaller states. North-South cooperation therefore has urgent value in preventing systemic risks in the security dimension.

Fourth, at the level of sovereign equality, an international order based on sovereign equality must be built in the AI era. The principle of sovereign equality established by *the Charter of the United Nations* is the cornerstone of the contemporary international order. Yet the development of AI technology and the power shifts it generates are posing new challenges to this principle. As technological capability increasingly becomes a decisive factor in comprehensive national strength, Global South countries lacking independent AI capabilities may in practice become "technological appendages" of major powers, with their policy autonomy and room for strategic choice eroded. Against this backdrop, advancing North-South cooperation is not only a matter of capacity building at the technological level, but also an inevitable requirement for upholding the basic norms of international relations and ensuring that sovereign equality is respected in the AI era. The African Union's *Continental Artificial Intelligence Strategy*, adopted in 2024, clearly emphasizes an Africa-centered, development-oriented AI pathway that values equity and technological sovereignty. Brazil's *National Artificial Intelligence Plan* (PBIA), launched in 2024, likewise explicitly set out a sovereign AI vision that rejects digital colonialism, is grounded in local culture, and focuses on autonomous computing capacity. This reflects the Global South's growing view that AI capacity building is an important component of safeguarding sovereign independence and the right to development.

b) Practical Constraints on Global North-South Cooperation

First, the return of geopolitics and strategic competition among major powers is squeezing the space for cooperation. Technological competition in current international politics is profoundly shaping the AI governance agenda. In recent years, countries have introduced successive strategies and policy arrangements around critical and emerging technologies, advanced semiconductors, computing infrastructure, and international AI competitiveness, highlighting the trend of placing AI at the core of major-power competition. In this context, Global South countries are more vulnerable to the spillover effects of supply-chain restructuring, export controls, investment screening, and standards competition. They also face pressure to choose sides in cooperation, causing development-oriented cooperation agendas to be securitized and politicized.

Second, capacity asymmetries lead to imbalances in cooperative relationships. Even in projects advanced in the name of cooperation, technology, capital, rules, and platforms often remain in the hands of Northern countries or transnational enterprises. If cooperation mechanisms lack arrangements for local capacity accumulation, institutional co-development, and knowledge sharing, cooperation may remain at the level of equipment procurement, platform access, or short-term training, making it difficult to achieve genuine capacity transfer. UNCTAD's research on inclusive AI has repeatedly emphasized that without simultaneous investment in computing power, data, skills, and institutional capacity, cooperation may reproduce existing inequalities rather than correct them.

Third, governance systems and institutional foundations are weak. Most Global South countries face significant bottlenecks in institutional capacity in the field of AI governance. This is reflected at multiple levels: the absence of systematic national AI strategies and policy and regulatory frameworks; weak professional capacity among regulatory bodies, making it difficult to effectively assess and manage AI risks; underdeveloped data governance systems, with gaps in rules for data rights confirmation, circulation, and transactions; and generally low levels of public digital literacy and awareness of

AI ethics. UNESCO's use of the Readiness Assessment Methodology (RAM) in recent years across multiple countries has further shown that insufficient institutional readiness has become a key constraint on many developing countries' participation in global governance and their ability to undertake technology cooperation.

Fourth, multilateral mechanisms are fragmented, and participation costs for the Global South are high. Current global AI governance is characterized by parallel mechanisms, overlapping rules, and dispersed discourse. The UN system, the OECD, the G7, the G20, regional organizations, and industry networks are all advancing related agendas, but these mechanisms differ in membership composition, issue priorities, and normative logic. Chapter V of UNCTAD's Technology and Innovation Report 2025 clearly notes that current international AI governance initiatives are highly fragmented and mainly led by developed countries. For resource-constrained developing countries, continuously tracking multiple mechanisms, coordinating different rules, and voicing positions effectively already entail substantial institutional and human-resource costs.

Fifth, the growing digital power of non-state actors is creating a crisis of governance legitimacy. In the AI era, technology companies, cloud service providers, chip firms, open-source communities, and standards organizations all influence rule formation and technology diffusion to varying degrees. This shift means that governance is no longer a purely intergovernmental matter. The problem is that a small number of large technology companies simultaneously control models, computing power, data, and platform entry points, and their commercial incentives are not always aligned with the public interest. In March 2026, OpenAI's suspension of computing and R&D investment in non-core businesses such as Sora reflected the decision-making logic of major firms prioritizing profitability and concentrating resources when computing power has become a scarce strategic resource.

(III) Feasible Pathways for Advancing Global North-South Cooperation on AI

After clarifying the importance and practical constraints of North-South cooperation on AI, the core question that must be answered is how to translate the vision of cooperation into actionable practical plans. At present, the UN system is shifting from a "principle advocate" to a "capacity coordinator". Multilateral mechanisms such as BRICS provide new platforms for collective bargaining by Global South countries, while the open and shared characteristics of AI technology itself also create new possibilities for cooperation. At the same time, the construction of computing infrastructure, the development of capacity-building systems, development-oriented application cooperation, and AI safety cooperation constitute the four priority areas of North-South cooperation. Exploring cooperation pathways that enable the orderly participation of diverse stakeholders and coordinated online and offline implementation is key to realizing inclusive global AI governance.

a) Building an Equal Model of Global North-South Cooperation on AI

First, the core role of the United Nations should be fully leveraged. With its universal representation, normative legitimacy, and capacity for agenda integration, the UN system continues to play an irreplaceable role in global AI governance. In 2024, the UN High-level Advisory Body on AI, in its final report *Governing AI for Humanity*, recommended the establishment of an international AI scientific panel, a global dialogue on AI governance, a capacity-building network, a global fund, and relevant institutional support mechanisms within the UN Secretariat. Subsequently, UN Member States adopted *the Pact for the Future* and its annex, *the Global Digital Compact*, in 2024, explicitly proposing the establishment of an independent international scientific panel on AI within the United Nations and the launch of a global dialogue on AI governance. In 2025, the UN General Assembly adopted a further resolution formally establishing

the terms of reference and modalities for these two mechanisms. At the same time, the UN Office for Digital and Emerging Technologies officially began operating in January 2025, providing institutional support for the UN system to advance digital cooperation and AI governance. This indicates that the United Nations is gradually shifting from an advocate of AI governance principles to an institutional coordinator and capacity platform. At the same time, it should also be recognized that the UN remains constrained by practical conditions in terms of resource input, implementation efficiency, and major-power coordination. It is therefore more appropriately positioned as a "platform-based coordinator" and "capacity-oriented enabler", complementing regional mechanisms, minilateral platforms, and multi-stakeholder networks.

Second, "minilateral" and regional mechanisms should serve as agile complements to the UN framework. Beyond the UN framework, BRICS cooperation mechanisms, the Shanghai Cooperation Organization, the Group of Twenty (G20), and various regional organizations are also becoming important platforms for advancing North-South cooperation on AI. The advantage of these mechanisms lies in their relatively limited number of participants and more concentrated interests, making it easier to reach substantive cooperation consensus and advance infrastructure development and capacity projects. For example, BRICS has achieved notable cooperation outcomes in the field of AI. From the establishment of the BRICS AI Study Group to the clear commitment in the *2024 Kazan Declaration* to strengthen AI capacity building and computing cooperation, BRICS has become an important arena for Global South countries to collectively shape global governance rules. At the same time, the African Union's release of the *Continental Artificial Intelligence Strategy* and ASEAN's issuance of the ASEAN *Guide on AI Governance and Ethics* further demonstrate the distinctive role of regional organizations as buffers that integrate local development demands and resist externally imposed rules.

Third, a collaborative governance framework should be built for the orderly participation of diverse stakeholders. Global AI governance is currently evolving from traditional intergovernmental cooperation toward parallel coordination among multiple stakeholders. *The Global Digital Compact* clearly states that international AI governance requires agile, multidisciplinary, and adaptable multi-stakeholder approaches. The UN High-level Advisory Body on AI has also noted that current AI governance faces a structural tension between private-sector dominance and the traditional intergovernmental political system. Against this background, enterprises, academic institutions, technical communities, and civil society organizations are playing increasingly important roles in standard-setting, capacity building, and risk prevention. The OECD Recommendation on Artificial Intelligence explicitly calls on countries to promote global technical standards development based on multi-stakeholder participation and consensus. UNESCO is also working with governments, enterprises, academic institutions, and civil society to advance implementation of its AI ethics framework. At the same time, multiple non-governmental organizations are gradually participating in coordinated governance. The World Data Organization (WDO) was established in Beijing in March 2026 with the aim of "bridging the data divide, unlocking data value, and prospering the digital economy". For North-South cooperation, the participation of diverse stakeholders can help better connect Northern technological resources with Southern development needs. However, transparent rules and inclusive arrangements are also needed to prevent governance resources from becoming further concentrated among a small number of technologically powerful countries and large platforms.

At present, North-South cooperation on AI is entering a new stage characterized by "multiple levels, multiple stakeholders, and weak centralization". In this stage, traditional intergovernmental cooperation and emerging multi-stakeholder collaboration are interwoven; the UN framework and regional mechanisms complement each other; and North-South cooperation and South-South cooperation reinforce one another. Building an open and inclusive governance framework and ensuring that all types of actors can contribute in their respective areas of strength are key to deepening cooperation.

b) Priority Areas and Implementation Pathways for North-South Cooperation on AI

First, cooperation on computing power and infrastructure: the material foundation

of North-South cooperation. Computing power is the material foundation of AI development and currently one of the most pronounced areas of the North-South divide. North-South cooperation in computing power can proceed through two models. The first is a co-development model, in which Northern and Southern partners jointly fund the construction of regional computing infrastructure, enabling resource sharing and risk sharing. The second is a computing power public goods model, in which technologically advanced countries provide computing resource assistance to developing countries. This would lower technology access barriers for Global South countries through cross-border transmission of computing power, and by promoting foundation AI models and open-source computing platforms as global public goods.

Second, capacity-building cooperation: moving from "tool provision" to "system cultivation". AI cooperation should not remain at the level of equipment donations or short-term training, but should shift toward the long-term cultivation of institutions, talent, and innovation systems. Specifically, it should cover four levels. The first is talent development: through scholarship programmes, joint training plans, and the sharing of online education resources, AI professionals and interdisciplinary talent should be trained for Global South countries.

Second, governance capacity building: Global South countries should be assisted in establishing and improving institutional frameworks, regulatory bodies, and policy tools for AI governance, thereby enhancing their institutional capacity to participate in global governance. The third is legal and regulatory support: Global South countries should be helped to formulate AI laws and regulations suited to their national conditions, and to build robust rule systems in areas such as privacy protection, algorithmic accountability, and intellectual property. The fourth is industrial incubation: through technology parks, innovation funds, and entrepreneurship mentorship programmes, support should be provided for the autonomous growth of AI industrial ecosystems in Global South countries.

Third, data and application cooperation: reshaping the logic of AI cooperation through a development-oriented approach. The application of AI in agriculture, healthcare, climate response, urban governance, and other fields holds enormous potential for helping Global South countries address development challenges. However, traditional international technology cooperation has often been based on the logic of technology exports from Northern countries and market opening by Southern countries. Data cooperation has displayed an even more unequal pattern of "Northern extraction, Southern contribution, and Northern benefit". Advancing a North-South transformation in data and application cooperation requires the establishment of three principles. The first is development orientation: the starting point and ultimate aim of AI application cooperation should be addressing the practical development needs of Global South countries, rather than merely exporting technology and expanding markets. The second is data reciprocity: fair rules for cross-border data flows should be established to ensure that Global South countries not only contribute data in data cooperation, but also capture data value, rather than being subject to one-way "data extraction". The third is model diversification: support should be given to the development of localized AI models that respond to the languages, cultures, and development-stage needs of Global South countries, avoiding the poor fit caused by simple "technology transplantation".

Fourth, safety cooperation: building an inclusive safety cooperation mechanism. The distribution of AI safety risks shows clear North-South asymmetry. Because of their shortages in technological capacity and governance resources, Global South countries are more likely to become bearers of safety risks. Establishing an "inclusive safety cooperation mechanism" oriented toward Global South countries is an indispensable component of North-South cooperation. This mechanism should cover the development of AI safety assessment capacity in Global South countries, including technical training, tool sharing, and joint research; timely notification and sharing of safety information and threat intelligence; and the development of safety solutions tailored to the specific needs of Global South countries, such as algorithmic transparency tools for grassroots governance and AI-enabled disinformation detection tools for electoral politics. In addition, on issues related to the military application of AI and lethal autonomous weapons systems, the security concerns and positions of Global South countries should be fully reflected, so as to prevent unilateral actions by major powers in this area from exacerbating global security deficits.

Part IV China's Role in Global North-South Cooperation on Artificial Intelligence

The rapid development of artificial intelligence is profoundly reshaping the international landscape. As the intelligence divide continues to widen, the gap between the Global North and the Global South risks becoming further entrenched and deepened. Advancing North-South cooperation on AI is not only a practical necessity for narrowing technological gaps and sharing the dividends of development; it is also an essential requirement for building a fair and equitable global governance system and upholding an international order based on sovereign equality.

Yet such cooperation faces serious challenges, including the return of geopolitics, the fragmentation of governance mechanisms, and persistent asymmetries in capacity. These challenges call for all parties to build consensus within the United Nations framework, coordinate action through multilateral platforms, and generate collective momentum through the participation of diverse stakeholders.

In the evolving architecture of global AI governance and North-South cooperation, China, as the world's largest developing country and a major actor in AI technology, possesses distinctive structural conditions to serve as a hub for cooperation. Whether and how China can play a constructive role will bear not only on the inclusive development of global AI governance, but also on the Global South's development capacity and rule-shaping participation in the intelligent era.

(I) Practical Pathways for China to Advance North-South Cooperation

China's practical pathways for advancing North-South cooperation are becoming increasingly clear, ranging from infrastructure cooperation and the shaping of multilateral platforms to the projection of a responsible major-country role that gives equal weight to safety and development. As an important member of the Global South and a significant force in the field of AI, China has both the foundation and responsibility to promote North-South cooperation, as well as the structural conditions to become a hub for such cooperation.

First, China has developed a relatively complete technological and industrial system. In the field of AI, China has built a comparatively comprehensive ecosystem spanning basic research and application development, hardware manufacturing and software services, consumer internet and industrial internet. It has achieved considerable scale and competitiveness across these domains. According to Stanford University's *2026 AI Index Report*, China ranks among the world's leading countries in AI publications, patent applications, and robot installations. In areas such as large language models, computer vision, and autonomous driving, its technological capabilities have approached or reached the international frontier. This comprehensive technological and industrial system gives China the capacity to provide Global South countries with full-chain cooperation, from computing infrastructure to application solutions.

Second, China has dense cooperation networks across the Global South. China has long maintained extensive networks of cooperation with Global South countries across political, economic, and people-to-people dimensions, accumulating rich experience and institutional resources. Whether through the Digital Silk Road under the Belt and Road Initiative, the Forum on China-Africa Cooperation, the China-Arab States Cooperation Forum, or Lancang-Mekong Cooperation, these mechanisms provide channels and platforms for North-

South cooperation in AI. In addition, China maintains close links with multilateral mechanisms led by or closely associated with developing countries, including BRICS and the Shanghai Cooperation Organization, where it exercises significant influence.

Third, China's development and capacity-building priorities are highly aligned with the needs of Global South countries. China's development experience and approach to capacity building resonate strongly with many countries in the Global South. Its further deepening of cooperation under the Global Development Initiative offers reference value for countries seeking opportunities for modernization. China's advocacy of "respect for sovereignty" in global AI governance echoes the Global South's emphasis on "autonomous development". At the same time, China's emphasis on "independent and controllable core technologies" in response to external technological restrictions also aligns with developing countries' aspirations for technological autonomy. This narrative convergence provides an important conceptual foundation for China to serve as a bridge in North-South cooperation.

Fourth, China has institutional embeddedness in multilateral platforms. China has consistently been a strong supporter of the United Nations framework and has the capacity to shape agendas on core global governance platforms. In October 2023, China released the *Global AI Governance Initiative*, explicitly proposing support for the establishment of an international AI governance institution under the UN framework. At the same time, China has actively relied on the Global Development Initiative and the South-South Cooperation Assistance Fund to prioritize digital capacity building. This demonstrates both China's willingness and its capacity to speak on behalf of developing countries in the multilateral contestation over global AI governance.

On this basis, China is exploring practical pathways to advance North-South cooperation.

Pathway One: Building a development community through infrastructure cooperation. Infrastructure is the material foundation for AI development and one of China's strongest areas in promoting North-South cooperation. China should rely on the Digital Silk Road to provide Global South countries with financial and technical support for AI infrastructure. Specific actions could include supporting African, ASEAN, and Latin American countries, through public-private partnerships, in building localized data centers, supercomputing platforms, and supporting green energy facilities; promoting the international deployment of Chinese open-source foundation models, such as DeepSeek, and cost-effective computing solutions to advance equitable access to AI; and jointly carrying out digital vocational education initiatives, such as Luban Workshops, to help partner countries train local AI engineers. These efforts would help build a community with a shared future oriented toward substantive technology transfer. Infrastructure cooperation can not only help Global South countries bridge the intelligence divide, but also deepen China's convergence of interests with developing countries and build a development-oriented community of shared interests.

Pathway Two: Shaping inclusive governance mechanisms through multilateral cooperation. China should make full use of its cooperation foundations within BRICS, the Shanghai Cooperation Organization, the G20, and other multilateral platforms to promote a more inclusive global AI governance framework. In agenda-setting, the core concerns of developing countries, including narrowing the technological divide, supporting capacity building, and advancing development-oriented governance principles, should be fully incorporated into the global governance agenda. In rule-making, China should promote fairer data governance rules, algorithmic transparency standards, and intellectual property arrangements. In institution-building, China could support the establishment of a more representative global AI governance coordination mechanism under the UN framework, ensuring the full participation and voice of Global South countries.

Pathway Three: Advancing global governance cooperation by giving equal weight to safety and development. AI safety is an urgent issue in global governance and an important domain in which major powers can demonstrate responsible leadership. China should play an active role in promoting AI safety cooperation, including by participating constructively in UN discussions on lethal autonomous weapons systems and supporting the development of necessary international norms; carrying out AI safety capacity-building cooperation

with Global South countries to help them improve risk identification and prevention capabilities; and promoting mechanisms for sharing AI safety information and threat intelligence to ensure that safety risks are effectively addressed at the global level. At the same time, China should clearly place "development" at the center of AI governance, emphasize development-oriented governance principles, and prevent AI governance from being dominated by securitization and politicization.

China should therefore call for a broadly inclusive North-South cooperation network.

Against the backdrop of intensifying strategic competition among major powers, some countries have sought to instrumentalize AI governance and build exclusive technological alliances and governance "small circles". This not only weakens the universality of global governance, but also creates practical pressure on North-South cooperation. In advancing North-South cooperation, China should consistently uphold a "de-bloc" approach and practice genuine multilateralism. This should be reflected in three core principles.

First, cooperation should be premised on equality and openness. China's technology cooperation projects should be open to all countries willing to participate and should avoid the Cold War logic of zero-sum competition. Cooperation should not impose exclusive clauses requiring partner countries to sever ties with particular system ecosystems or international strategic groupings. The right of Global South countries to choose freely among different technological pathways should be protected.

Second, cooperation should be conditioned on institutional autonomy. North-South cooperation should respect each country's right to choose its own development path and governance model. It should not make any particular political system, governance model, or set of values a precondition for cooperation. China should respect the right of countries to explore AI governance models in line with their own cultural traditions and legal systems. It should not make specific regulatory systems a prerequisite for technical assistance, thereby avoiding cooperation models with political or institutional conditionalities.

Third, cooperation should be guided by development objectives and capacity building. The ultimate goal of North-South cooperation should be to help Global South countries achieve autonomous and sustainable development, rather than create new forms of dependency. China's cooperation projects should be grounded in the practical needs and development stages of partner countries, with capacity building and the enhancement of autonomous development capabilities at their core. In this sense, cooperation should truly "teach people how to fish" rather than merely "give them fish".

(II) China's Latest Practices in Building AI as an International Public Good

China promotes the development of AI as an international public good by taking multilateralism as its institutional philosophy, building open-source ecosystems as a core advantage for Chinese enterprises going global, and advancing public-private capacity-building initiatives as a central task. Through these efforts, China is working to deepen the application of AI international public goods across concrete scenarios and empower a wide range of sectors.

a) Practicing Genuine Multilateralism as the Foundational Principle of China's AI International Public Goods

Providing international public goods to the world requires, first and foremost, sound international institutional development and alignment of rules and standards. Since the release of the *Global AI Governance Action Plan*, China has continued to advance multilateral cooperation on AI governance and capacity building, helping move relevant issues from principles and initiatives toward institutional implementation. The role of the United Nations as the main channel has been further strengthened, regional multilateral mechanisms have accelerated alignment, cross-regional cooperation platforms have continued to expand, and the platform foundations for international AI governance have become increasingly clear.

First, the global AI governance framework is taking shape more rapidly, with the United Nations playing a stronger role as the main channel. In response to the fragmentation of AI governance rules and the spillover effects of small-group mechanisms, China has continued to promote AI governance and capacity building within the UN framework. The 78th session of the UN General Assembly adopted by consensus the China-led resolution on "Enhancing International Cooperation on Capacity-building of Artificial Intelligence". Focused on capacity building, inclusive development, and international cooperation, the resolution received co-sponsorship or support from 143 member states. It emphasizes assistance to all countries, especially developing countries, in strengthening AI capacity building and enhancing their representation and voice in global governance processes, thereby providing an important institutional foundation for subsequent cooperation. The UN General Assembly has since adopted further resolutions deciding to establish an independent international scientific panel on AI and a global dialogue mechanism on AI governance, moving AI governance from general discussion of principles into a regularized agenda. These developments have strengthened the UN's coordinating role in AI governance and provided institutionalized channels for developing countries to participate equally in rule discussions and express their development needs and safety concerns.

Second, regional multilateral mechanisms have become more substantive and more closely aligned with the governance concerns of Global South countries. Regional multilateral mechanisms have become important vehicles for translating global governance consensus into operational rules, while responding to the practical governance and development needs of Global South countries. At the 2025 BRICS Summit, the BRICS Leaders' Statement on AI Governance emphasized that AI development should fully take into account the interests and practical needs of developing countries, narrow the digital divide, and prevent the emergence of new technological divides. At the 2025 Shanghai Cooperation Organization Tianjin Summit, the statement on further deepening international AI cooperation explicitly affirmed that all countries enjoy equal rights to develop and use AI, and incorporated the development of AI cooperation mechanisms into the summit outcomes. The China-SCO AI Cooperation Forum further released a plan for establishing an application cooperation center, setting out concrete arrangements for digital infrastructure, industrial coordination, and talent development. At the same time, regional organizations have expanded practical development agendas. The upgraded Protocol of the China-ASEAN Free Trade Area 3.0 introduced a digital economy chapter for the first time. *The Initiative on Deepening China-ASEAN Digital Governance*

Cooperation, released in December 2025, established a dedicated chapter on AI governance cooperation, proposing stronger coordination in risk response, standardization of safety governance, policy and regulatory development, and joint participation in global rule-making. The intensive implementation of regional mechanisms has enabled AI governance to move from principled consensus toward rule alignment and institutional coordination, while respecting differences in national policies and practices.

Third, by practicing genuine multilateralism, China has attracted more cross-regional partners into multilateral governance processes. China has upheld inclusiveness, openness, and shared benefits, promoting the expansion of AI governance and cooperation from a single platform toward multi-level, multi-stakeholder, and multi-regional coordination. A governance and cooperation network is gradually emerging that connects global processes, regional mechanisms, and industrial implementation. Cross-regional cooperation, including China-Africa and China-Arab cooperation, has continued to deepen. The Chinese government and the African Union signed a memorandum of understanding on science and technology cooperation, incorporating joint research, platform co-development, technology transfer, and researcher exchanges into the cooperation agenda. Launched in July 2025, the International Network for AI Industry Development Alliance brought in nine international institutions as founding members, providing a vehicle for cross-border collaboration among industry and research institutions. Platforms such as the China-Arab Research Center on Reform and Development have actively promoted the alignment of AI cooperation with the digital transformation needs of Middle Eastern and Arab countries. By continuously expanding partners, enriching cooperation vehicles, and extending cooperation chains, China has helped develop a more complete governance system, from global governance initiatives and regional mechanism-building to industrial project implementation. Through genuine multilateralism, China is helping build governance consensus, pool cooperation resources, and provide strong support for an open, inclusive, and broadly shared international AI governance architecture.

Fourth, alignment of AI governance rules and standards has become more refined. Promoting inclusive access to AI and AI for good requires balancing development and security, strengthening international cooperation, and improving rule alignment. The China-ASEAN AI Ministerial Roundtable agreed to promote the launch of a China-ASEAN AI Safety Network, with cooperation covering algorithmic safety assessment, model risk early warning, deepfake detection, and emergency response coordination. China's initiative on deepening China-ASEAN digital governance cooperation also treats AI governance as a dedicated area, proposing stronger communication and exchange on risk response, standardization of safety governance, and policy and regulatory development, as well as coordination in multilateral platforms such as the United Nations. The 2025 China-Africa Internet Development and Cooperation Forum released the *Action Plan for Jointly Building a China-Africa Community with a Shared Future in Cyberspace (2025-2026)*, placing digital economy development, cybersecurity safeguards, digital capacity building, and AI governance within a unified framework. The two sides agreed to strengthen communication, exchange, and practical cooperation on responding to AI-related risks and standardizing safety governance, and to jointly prevent the misuse of AI technologies. The World Internet Conference released the *Global AI Standards Development Report*, providing a reference for countries to understand AI standardization processes, participate in standards discussions, and advance rule alignment. The *BRICS Leaders' Statement on Global AI Governance*, SCO statements, and related initiatives have helped translate safety and ethical requirements into more concrete standards discussions and rule alignment, reducing regulatory gaps among countries in safety assessment, data governance, and technology application.

b) Building an Open-Source Ecosystem as the Core Approach to China's AI International Public Goods

In response to high access costs for foundation models, insufficient support for low-resource languages, and weak local development capacity, open-source models, development platforms, and evaluation tools have continued to expand. These resources provide global developers, especially those in developing countries, with more accessible AI technology at lower thresholds.

First, the supply of open-source models has continued to grow in scale. Alibaba's Qwen series has formed an open-source matrix covering different parameter scales, multimodal tasks, and sectoral needs. As of February 2026, more than 400 Qwen models had been open-sourced. The series had surpassed one billion downloads on Hugging Face and generated more than 200,000 derivative models. Since the second half of 2025, it has ranked first on the platform's monthly download charts, surpassing Meta's Llama series from the United States, and has been adopted by companies such as BYD and Rokid as their preferred foundation model for deployment. By the end of 2025, MiniMax had served more than 236 million individual users across more than 200 countries and regions, and had opened its M-series reasoning models and Hailuo video models to 214,000 enterprise clients and developers in more than 100 countries and regions. Its M2.5 model reached the top of OpenRouter's weekly call volume rankings less than one week after launch on the overseas developer platform, contributing 1.44 trillion tokens in a single week. Chinese open-source models, characterized by openness, low cost, and high efficiency, are shifting from being a "fallback option" for global developers to becoming a "regular-use option".

Second, open-source community development has accelerated and expanded. By the end of 2025, Alibaba Cloud's ModelScope community hosted more than 120,000 open-source models, with over 20 million registered users across more than 200 countries and regions. Platform-based ecosystems integrate models, datasets, toolchains, and developer resources. International AI standards led by the China Academy of Information and Communications Technology have been released through ITU-T, covering foundation model platforms, code generation, retrieval-augmented generation, and other areas, thereby providing technical references for model platform development and service delivery. The development of low-resource language evaluation datasets such as LaoBench has also helped address gaps in large-model evaluation for Southeast Asian languages. Multi-level achievements, including open-source frameworks, open computing platforms, and open-source interoperability protocols, continue to emerge. A new intelligent computing ecosystem characterized by openness is taking shape, significantly lowering the barriers to use for small and medium-sized enterprises, local developers, and research teams, and promoting broader updating, reuse, and sharing of open-source outputs.

Third, localized deployment has become the mainstream approach for the international expansion of open-source models. Low-resource languages and local cultural adaptation are major challenges for many developing countries in developing and applying AI. Chinese enterprises' continued promotion of open-source foundation models provides these countries with accessible, modifiable, and deployable base capabilities for localized model development. Singapore's national AI programme built Qwen-SEA-LION-v4 on Qwen3-32B, improving the model's understanding of Southeast Asian languages and cultural contexts. A local team in Uganda developed the Sunflower model based on Qwen 3, covering 31 Ugandan languages and supporting applications in public administration, healthcare, education, and community services. Malaysia's NurAI, built on Chinese open-source models, provides compliance advice for Islamic law, finance, healthcare, and other scenarios. Laos' national foundation model, LaoAI, is being developed through the China-Laos AI Innovation Cooperation Center and trained on Lao-language corpora to support AI applications in the national language environment, while also advancing cooperation on computing infrastructure, model research and development, and talent training. In the United Arab Emirates, Mohamed bin Zayed University of Artificial Intelligence and partner institutions released the K2Think reasoning model based on the Qwen architecture, using a relatively small parameter scale to perform complex mathematical reasoning tasks. By promoting open-source foundation models, China has not only expanded the global supply of foundation models, but also directly enabled partner countries to carry out local training, language adaptation, and sector-specific development. This helps transform general model capabilities into AI capabilities aligned with national languages, institutions, and industrial needs, while allowing countries to accumulate experience in model development, deployment, and iteration and strengthen their autonomous AI development capacity.

c) Public-Private Capacity Building as the Principal Task of China's AI International Public Goods

For AI to serve as an important international public good and achieve inclusive, sustainable empowerment, it cannot rely solely on models and application projects. It also requires enabling conditions such as cloud services, data processing, local technological systems, skilled talent, and energy security. China's current approach to international AI cooperation is gradually moving from the provision of individual applications toward the development of industrial foundations and capacity systems. By providing more comprehensive public support, China is helping partner countries consolidate the foundations for AI development and better share the opportunities of the intelligent era.

First, intelligent infrastructure development is accelerating. Cloud computing, data centers, and local data processing capacity are essential foundations for AI deployment. Tencent Cloud announced that it would build its first Middle East cloud region in Saudi Arabia and invest more than USD 150 million in local cloud infrastructure and technology development, serving applications in cloud computing, AI, digital media, e-commerce, and financial services. Alibaba Cloud launched its second data center in Dubai and established cooperation ecosystems with enterprises in Middle Eastern finance, healthcare, digital humans, cloud gaming, and other sectors, further improving regional cloud service capacity. TikTok System Thailand received approval for a large-scale data infrastructure project to add servers in Bangkok, Samut Prakan, and Chachoengsao, expanding data storage and processing infrastructure to meet growing demand for digital services. These efforts strengthen the cloud services, data processing, and computing capacity conditions needed by partner countries to develop AI, while also supporting the construction of regional digital infrastructure hubs.

Second, China is supporting the development of autonomous technology ecosystems. Inclusive and widely shared international AI cooperation depends on helping partners build localized and sustainable technology systems. Singapore's national AI programme built Qwen-SEA-LION-v4 on Alibaba's Qwen3-32B, replacing its previous Meta Llama architecture to better adapt to Southeast Asian linguistic and cultural contexts. The model can run on consumer-grade devices, making it more accessible to regional SMEs and developers that lack large-scale computing clusters. The China-Laos Joint Statement supports Laos in advancing digital infrastructure such as national data centers, cloud computing systems, and modern communications networks, while strengthening cooperation in AI, cybersecurity, and technological innovation. Through the China-Laos AI Innovation Cooperation Center, the two sides are advancing computing infrastructure, a Lao-language foundation model R&D base, and sectoral model incubation. In September 2025, LaoAI, a national foundation model trained on Lao-language corpora, was officially released to serve the Lao public and domestic industries. In Cambodia, ByteDC launched a high-performance sovereign cloud based on Huawei Cloud Stack, providing trusted domestic cloud infrastructure for government departments, regulators, banks, and enterprises. In Thailand, Huawei Cloud has supported public-sector AI deployment, working with the National Electronics and Computer Technology Center, food and drug regulators, and the Ministry of Commerce on applications including meeting speech recognition, public administration efficiency, and regulatory document analysis. In Pakistan, the China-Pakistan Community with a Shared Future Action Plan incorporates cooperation in software, AI, 5G, cloud computing, and big data into the digital economy agenda, while proposing stronger exchanges among cybersecurity and data security institutions and promoting cooperation in the digital security industry. From localized open-source models to sovereign cloud and public-sector applications, China is helping partners embed AI capabilities into autonomous and controllable technology systems.

Third, talent development and skills-building are deepening, providing long-term support for autonomous capacity. In October 2025, the third AI Capacity-Building Workshop opened in Shanghai, bringing together 38 participants from 21 countries across Asia, Africa, Latin America, and Eurasia. The programme covered inclusive development, global governance, ethical governance, industrial applications, and educational empowerment, and included visits to research institutions and enterprises. It provided a platform for government officials from developing countries to systematically understand technological trends and governance issues. Training for youth and professional groups has also expanded. Huawei signed an

agreement with the Peruvian government to train 20,000 Peruvian young people. Tianjin University of Technology and other Chinese institutions supported Colombia's Ministry of Education in AI training for teachers, with the first cohort covering 125 teachers in basic and higher education, especially those from remote areas and STEM education programmes. Huawei signed a memorandum of understanding on digital education cooperation with Cambodia's Ministry of Education to upgrade school ICT infrastructure, digital skills for teachers and students, and school management systems. Alibaba Cloud held its first overseas Qwen conference in Singapore and partnered with an institution under Singapore's National Trades Union Congress to help local enterprises, developers, and students master generative and agentic AI skills. The China-Africa Internet Development and Cooperation Forum has promoted the regularization of training mechanisms in cyberspace affairs, continuously carrying out seminars on cybersecurity, the digital economy, and AI governance. These projects extend capacity building from equipment and platforms to human capital and organizational capabilities, providing a talent base for partners' long-term use of AI and autonomous innovation.

Fourth, the intelligent upgrading of energy and sectoral infrastructure is providing a stable operating environment for AI applications. AI development depends on stable energy systems and foundational sectoral support. In December 2024, the State Grid Corporation of China released "Guangming", the first hundred-billion-parameter multimodal industry foundation model in China's power sector. In 2025, the model was expanded to power-grid operation and maintenance scenarios in Brazil, serving the intelligent management of the transmission network of State Grid Brazil Holding. Public reports indicate that the company operates more than 16,000 kilometers of transmission lines, covering key assets such as the Belo Monte Phase II ultra-high-voltage transmission project. In response to the long distances of Brazil's transmission lines, complex terrain, weak network conditions in some regions, and the low efficiency of traditional manual and helicopter inspections, relevant platforms integrate AI image recognition, drone inspection, mobile offline operations, and closed-loop defect management to improve the efficiency and standardization of transmission inspection and maintenance. China's pragmatic cooperation in power infrastructure is thus upgrading from engineering construction to intelligent operation and maintenance services. It also shows that industry foundation models, inspection data, and operation platforms can be embedded in overseas energy systems to provide intelligent management capabilities for large cross-border energy projects. Improvements in the operating efficiency of energy and power systems also provide a more stable foundation for AI applications and digital economic development.

d) Deepening Sectoral Applications as the Value of China's AI International Public Goods

The role of AI in empowering development has become increasingly prominent. AI technology exchange and capacity building are gradually extending into practical applications across a growing range of fields, enabling AI to play an active role in improving livelihoods, enhancing governance effectiveness, and promoting industrial upgrading.

First, AI has delivered notable results in livelihood and civilian applications. The "Mazu" early-warning system released by the China Meteorological Administration addresses multi-hazard early-warning needs through cloud-based systems and customizable solutions. As of April 2026, it had been deployed in seven countries, including Pakistan and the Solomon Islands; supported cloud-based applications in more than 40 countries; provided over 100 categories of data products to 153 countries and regions; and delivered international training covering 89 countries and regions. The AI meteorological forecasting application demonstration project led by China's Ministry of Science and Technology plans to deploy operations in at least six countries, covering an early-warning population of 10 million. By directly applying AI models, meteorological data, and early-warning services to disaster risk reduction and public safety, this represents a major case of AI serving people's livelihoods. Cooperation in agriculture has likewise responded to front-line production challenges. Brazil's National Institute of the Semi-Arid Region and China Agricultural University jointly established a laboratory for family farming mechanization and AI, focusing not on large commercial farms but on family agricultural producers in Brazil's semi-arid regions, with cooperation in environmental

monitoring, data analysis, and agricultural machinery applications. Runnjian provided solutions for smart palm plantations in Malaysia. To address the short maturity window of oil palm fruit and high losses from missed harvesting, the system uses AI algorithms to identify fruit maturity, drone hyperspectral cameras to mark ripe fruit locations, and intelligent vehicles for precision harvesting. This increased yields by more than 10 percent and was described by Malaysian industry organizations as a profound agricultural technology revolution.

Second, AI is improving public governance effectiveness and enabling smart city development. Huawei Cloud has promoted AI deployment in Thailand's public sector, with cooperation covering meeting speech recognition, improved public administration processes, regulatory document analysis, and public service optimization. Thailand's food and drug regulators have applied AI to product registration document analysis, screening, and summary generation to improve regulatory efficiency. DeepBlue Technology has participated in Uzbekistan's New Tashkent smart city project, cooperating on smart cities, smart housing, international logistics, and smart sanitation. DeepBlue Technology has also cooperated with relevant Iraqi institutions in areas including infrastructure reconstruction, intelligent upgrading of oil refining, digital banking payment systems, multi-level smart healthcare, and AI talent training. These cases show that AI is entering concrete processes of local governance and public service delivery.

Third, sectoral applications and intelligent equipment are expanding rapidly. Huawei established a 13,638-square-foot AI Lab in Kuala Lumpur's TRX district, its first AI-enabled industry incubation base outside China; its solutions now cover about 97 percent of Malaysia's population. AstraZeneca China built a drug safety reporting tool based on Qwen and Alibaba Cloud, used to identify adverse event information from medical literature and generate reports. The tool increased reporting accuracy from 90 percent to 95 percent and improved report generation efficiency by about 300 percent. Alibaba Cloud has worked with Wio Bank to develop an AI banking agent based on the Qwen foundation model; partnered with BYOND Asia to create Arabic-language digital human solutions for Middle Eastern markets; supported ACCUMED in developing AI applications for medical insurance claims and medical coding; provided low-latency, high-performance cloud infrastructure for The Game Company to promote AI cloud gaming; and deepened cooperation with the global digital services provider Atos to bring cloud computing and AI products to enterprises in the Middle East and the Asia-Pacific. These partnerships reflect the transformation of Chinese enterprises from traditional infrastructure exporters to providers of integrated "cloud computing + AI + industry solutions". They also strengthen localized technical service capacity and international competitiveness through local data centers and ecosystem partnerships, providing important support for the development of the digital economy and AI industries in the Middle East. Westwell's Q-Truck autonomous driving fleet has expanded at the Port of Felixstowe in the United Kingdom. Its first vehicles had already been operating in mixed port environments, representing an important case of autonomous truck application in European ports. Fourier Intelligence presented its GR-3 humanoid robot at CES 2026, demonstrating the potential of embodied intelligence in rehabilitation, eldercare, companionship, and interactive scenarios. SenseTime and King Saud University have jointly established an AI innovation center to explore applications in education, research, and multiple industries.

Overall, since the release of the *Global AI Governance Action Plan*, China has made positive progress in building AI as an international public good. These efforts are not limited to the provision of a single category of technology product. Rather, they have gradually formed a more systematic cooperation pathway around the practical constraints faced by developing countries in participating in AI governance and using AI technologies. At the institutional level, China has relied on the United Nations and regional multilateral mechanisms to promote agenda alignment, enabling AI governance to place greater emphasis on the development needs of Global South countries. At the ecosystem level, open-source models and platform ecosystems have lowered the threshold for technology use and provided foundations for low-resource language support and local model development. At the capacity-building level, intelligent infrastructure,

talent development, and energy-sector digitalization projects have provided support for partners' long-term use of AI and autonomous development. At the application level, projects in meteorological early warning, smart agriculture, public governance, and sectoral applications have entered concrete scenarios and begun to produce tangible outcomes.

At the current stage, the development of AI as an international public good remains an ongoing process, and the effectiveness of some platforms and projects will need to be tested through continued operation. The key lies not in the one-off export of technology, but in helping partner countries gradually build sustainable development capacity. Going forward, as multilateral mechanisms continue to operate, open-source ecosystems further expand, and scenario-based projects continue to be implemented, the practical foundation for China's efforts to make AI serve common development is expected to become stronger. These efforts will provide further support for narrowing the global intelligence divide, enhancing the digital governance capacity of developing countries, and advancing AI for good.


MANDARIN ORIENTAL
DBS
HSBC