CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

JUNE 2025

How Some of China's Top AI Thinkers Built Their Own AI Safety Institute

Scott Singer, Karson Elmgren, and Oliver Guest

How Some of China's Top AI Thinkers Built Their Own AI Safety Institute

Scott Singer, Karson Elmgren, and Oliver Guest

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Abbreviations

AIIA	Artificial Intelligence Industry Alliance (China)
AISI	AI safety institute (term used globally)
BAAI	Beijing Academy of Artificial Intelligence
Beijing-AISI	Beijing Institute of AI Safety and Governance
CAICT	China Academy of Information and Communications Technology
CAIS	Center for AI Safety (United States)
CAS	Chinese Academy of Sciences
CCID	China Center for Information Industry Development
CCP	Chinese Communist Party
CnAISDA	China AI Safety and Development Association
I-AIIG	Institute for AI International Governance (China)
IDAIS	International Dialogues on AI Safety
MIIT	Ministry of Industry and Information Technology (China)
STEM	science, technology, engineering, and mathematics
UK AISI	United Kingdom AI Security Institute (previously UK AI Safety Institute)
U.S. AISI	United States AI Safety Institute

Executive Summary

Since the January 2025 release of the DeepSeek-R1 open-source reasoning model, China has increasingly prioritized leveraging artificial intelligence (AI) as a key engine for economic growth, encouraged AI diffusion domestically, and continued to pursue self-sufficiency across the AI stack. Yet while China has been investing heavily in AI development and deployment, it has also begun to talk more concretely about catastrophic risks from frontier AI and the need for international coordination. The February 2025 launch of the China AI Safety and Development Association (CnAISDA, 中国人工智能发展与安全研究网络)—China's self-described counterpart to the AI safety institutes (AISIs) that the United Kingdom, United States, and other countries have launched over the last two years—offers a critical data point on the state of China's rapidly evolving AI safety conversation.

Despite its potential importance, little has been publicly reported on CnAISDA. What is it? How did it come about? And what does it signal about the direction of Chinese AI policy more broadly? This paper provides the first comprehensive analysis of these questions.

What Is CnAISDA?

Function. As of this writing, CnAISDA's primary function is to represent China in international AI conversations, including those with other AISIs, underscoring China's willingness to engage on frontier AI issues outside its traditionally preferred venue, the United Nations. Unlike the United Kingdom's and United States' AISIs, CnAISDA does not currently appear to be structured to carry out substantial domestic functions, such as independently testing and evaluating frontier AI models.

Structure. CnAISDA integrates multiple existing Chinese AI-focused institutions into a network structure. Rather than being a new stand-alone agency to govern AI, CnAISDA is more of a coalition to represent China abroad, as well as to advise the government. This avoids the need for the government to "pick winners" in China's policy ecosystem. CnAISDA's leaders have credibly claimed that the organization has the Chinese government's support, though the exact relationship is unclear.

Personnel. CnAISDA provides a formal platform for influential experts with strong pre-existing government and international connections. This arrangement elevates key policy entrepreneurs, including Fu Ying, a former vice minister of foreign affairs; Andrew Yao, China's sole Turing Award winner; and Xue Lan, an important external adviser to the powerful State Council.

How Did CnAISDA Come About?

The formation of CnAISDA represents the culmination of years of strategic positioning by major policy entrepreneurs within China's AI governance ecosystem. Their collective efforts evolved as they engaged in both a growing international conversation around frontier AI risks and a burgeoning AI development and governance ecosystem shaped by a range of distinctive domestic priorities.

CnAISDA emerged from a years-long evolution of Chinese interest in AI safety, beginning with concerns from a small group of Chinese scientists in the late 2010s. The inclusion of high-level figures from China's AI community in international forums and publications—such as the 2023 AI Safety Summit at Bletchley Park in the United Kingdom and a global statement on AI extinction risk released by the San Francisco–based nonprofit Center for AI Safety (CAIS)—legitimized AI safety as a potential policy priority, though secondary to accelerating economic growth. Following the Bletchley summit and the formation of the first AISIs in the United States and United Kingdom, an internationally connected group of policy entrepreneurs within China developed a body that could engage in global AI governance conversations on frontier AI risks while fitting within China's domestic political context.

What Does CnAISDA Signal About the Direction of Chinese AI Policy More Broadly?

The establishment of CnAISDA presents promise for global AI governance, elevating experts who appear genuinely concerned about catastrophic AI risks and are motivated to build common international standards to reduce them. A recent speech from Chinese President Xi Jinping suggests that the CnAISDA group may be influencing the thinking of China's leadership on these issues and laying a foundation for regulatory action.[1]

However, CnAISDA faces significant challenges. Engagement with the United States may be challenging for CnAISDA due to (1) uncertainty about the future of the U.S. Center for AI Standards and Innovation, the rebranded U.S. AISI; (2) the current U.S. administration's emphasis on AI opportunity rather than safety concerns; and (3) the broader context of hawkish attitudes in both countries. Additionally, motivation within China to address catastrophic risks, as elsewhere, might be limited. The government's backing of CnAISDA likely stems primarily from aspirations for global participation. While Chinese leaders have signaled concern about AI safety, their immediate priority has been promoting AI innovation to stimulate economic growth, creating a potential tension for CnAISDA between engagement in international safety efforts and the pursuit of China's development-focused domestic agenda. The Shanghai World AI Conference in July will provide an early litmus test: how seriously top Chinese leaders engage with frontier AI safety, and whether concrete commitments follow, will shine light on the level of CnAISDA's domestic influence.

Despite these challenges, CnAISDA's emergence represents a significant victory for a group of policy entrepreneurs in China that has long warned about catastrophic AI risks. While many AI policy thinkers have proposed treaties as central mechanisms to reduce shared risks from AI, CnAISDA's establishment offers a different path: it demonstrates how international-borne ideas about AI safety can naturally diffuse across different political systems, albeit taking different forms depending on national contexts. For international stakeholders, the organization provides both opportunities for engaging on AI safety and for gaining insights into China's evolving approach to frontier AI governance. It opens up several possible pathways for addressing shared catastrophic risks even as strategic competition intensifies in other dimensions of AI development.

Introduction

The release of DeepSeek-R1 in January 2025 catalyzed a global reckoning around the international competitiveness of frontier AI models produced in China, shaking financial markets worldwide and accelerating Chinese efforts to develop and diffuse advanced AI systems. In the weeks that followed, DeepSeek Chief Executive Officer Liang Wenfeng was quickly invited to meet with China's highest leadership, including Premier Li Qiang[2] and President Xi Jinping himself,[3] who has emphasized the need to push for rapid AI diffusion and self-sufficiency across the AI stack. While Chinese leadership has pushed to accelerate AI development with unmistakable urgency, questions remain about whether their focus on the safety and security of these increasingly powerful models is advancing with equal rigor and determination.

DeepSeek's breakthrough confirms that China's frontier AI capabilities have achieved global competitiveness. As a result, it has become increasingly essential for Western technology firms, national security analysts, and global civil society actors to understand China's AI safety ecosystem. China's approach to AI safety and security will shape its economic competitiveness in AI markets, the risk profile of its AI-enhanced military systems, and the emergence of potential global externalities, including catastrophic risks such as the facilitation of dangerous pathogen development or even AI systems that may escape from human control.

This paper examines the China AI Safety and Development Association, China's self-proclaimed counterpart to the AI safety institutes that have emerged globally over the past two years. Three fundamental traits characterize CnAISDA. First, it prioritizes representation in

international fora over domestic functions, positioning China as an engaged participant in global governance discussions without imposing binding frontier AI safety requirements that could hinder the competitiveness of its domestic developers. Second, CnAISDA leverages a networked architecture that brings together existing expertise across multiple institutions. In doing so, China avoids designating a single "winner" in its AI safety ecosystem, providing the Chinese Communist Party (CCP) with greater flexibility. Third, while it has an ambiguous relationship with the Chinese government, CnAISDA provides a formal platform for a set of influential experts with preexisting strong governmental and international connections to engage in global AI governance discussions on behalf of China. It creates valuable opportunities for resourceful policy entrepreneurs to leverage their position in China's international AI engagement efforts to potentially shape the country's rapidly evolving domestic frontier AI policy space.

Drawing on dozens of interviews and a survey of hundreds of relevant Chinese-language media and CCP documents, this paper describes how the formation of CnAISDA represents a milestone for China's AI safety ecosystem. Yet it also lays out the organization's limitations. While CnAISDA's establishment demonstrates the influence of scientists and policymakers concerned about frontier AI risks, the CCP's support for the organization may stem in large part from the party's aspirations for global participation rather than deeply held concerns about AI safety.

This tension creates a critical inflection point for China's AI safety community. Though CnAISDA has enabled China to join international AI governance conversations, it has yet to translate this engagement into substantive AI safety–oriented domestic policies, particularly regarding rigorous safety evaluations of Chinese frontier models. By analyzing the motivations behind CnAISDA's formation, this paper illuminates the complex path forward for China's AI safety leaders who are navigating global engagement under China's distinctive political constraints. This analysis not only sheds light on the likely direction of Chinese AI policy but also offers a case study of how ideas can diffuse into different political contexts without requiring formal international institutions, revealing critical informal pathways for global AI coordination.

In establishing CnAISDA, China has created an important platform for international engagement in global AI discussions outside its traditionally preferred United Nations (UN) system. While Chinese domestic AI policy remains focused on economic growth, the CCP leadership's support of a frontier AI safety institution signals that it recognizes the value in having government-connected expertise in this domain. Though primarily internationally facing today, CnAISDA could eventually pave the way for China to build more intragovernmental infrastructure to monitor and mitigate catastrophic risks generated from its companies' increasingly capable AI models. July's World AI Conference in Shanghai will offer early insight into how much CnAISDA is shaping China's domestic AI policy conversation.

This paper follows a three-part analytical framework to understand CnAISDA's significance for both China's AI ecosystem and the global policy landscape. First, it explores CnAISDA's institutional design and key actors, explaining how its networked structure differs fundamentally from stand-alone AISIs in the United States and United Kingdom. Second, the paper traces the historical roots that shaped CnAISDA's formation, revealing how both domestic AI safety advocates and international developments influenced its creation. Finally, it examines the strategic implications and open questions raised by CnAISDA's emergence, particularly regarding global coordination opportunities, domestic regulation in China, and the pathways for policy entrepreneurship in China's frontier AI landscape.

The Global Discussion on Frontier AI Risks

CnAISDA was not established in a vacuum; it joins a growing number of AISIs or equivalent institutions globally.[4] Broadly speaking, AISIs are government-backed organizations with a responsibility for reducing AI risks and sometimes catastrophic risks in particular. While AISIs vary in their name, scope, and institutional structure, in this paper, they are viewed as national and regional institutions empowered to tackle challenges connected to AI risk reduction.

The landscape of AISIs reflects significant variation in institutional mandates and risk priorities. Some institutions focus primarily on addressing immediate AI safety concerns, such as privacy violations, bias in AI systems, or harmful AI generated content such as child sexual abuse material. Others prioritize catastrophic risks that could have far-reaching societal consequences, including AI-enabled attacks on critical infrastructure or potential loss of control scenarios with advanced AI systems. Catastrophic risks are often linked to the most advanced (or "frontier") AI systems in particular and are therefore sometimes also discussed under the heading of "frontier AI risks."

In terms of institutional structure, CnAISDA deviates from early models like those in the United Kingdom and United States, which were established as entirely new government organizations. Instead, CnAISDA follows the approach of several newer AISIs—in Canada, France, and India—by integrating a coalition of existing institutions under a coordinated mandate (see Table 1). However, it distinguishes itself from many recent AISIs by returning to a core focus that motivated the earliest institutes: addressing catastrophic risks from advanced AI systems.

Table 1. AISIs Around the World

Jurisdiction and institution	Launch date	Explicit substantive focus on catastrophic risks?	Institutional structure
United Kingdom (AI Security Institute, formerly the AI Safety Institute)[i]	November 2023	Yes	New government organization
United States (Center for AI Standards and Innovation, formerly U.S. AI Safety Institute)[ii]	November 2023	Yes	New government organization
Japan (AI Safety Institute)[iii]	February 2024	No	New government organization
European Union (via the AI Office)[iv]	May 2024	Yes	The AI Office is a new institution but was not established specifically as an AISI
Singapore (AI Safety Institute)[v]	May 2024	No	Mandate given to existing organization
South Korea (AI Safety Institute)[vi]	November 2024	Yes	New government organization
Canada (AI Safety Institute)[vii]	November 2024	Somewhat	Mandate given to a grouping of existing organizations
France (National Institute for the Evaluation and Security of AI, abbreviated INESIA in French)[viii]	January 2025	No	Mandate given to a grouping of existing organizations
India (IndiaAI Safety Institute)[ix]	January 2025	No	Mandate given to a grouping of existing organizations
China (Chinese AI Safety and Development Association)	February 2025	Yes	Mandate given to a grouping of existing organizations

[i] Department for Science, Innovation and Technology and AI Safety Institute, "Introducing the AI Safety Institute," UK Government, November 2, 2023, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[ii] U.S. Department of Commerce, "At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety," November 1, 2023, https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial.

[iii] Ministry of Economy, Trade and Industry (METI), "Launch of AI Safety Institute," February 14, 2024, https://www.meti.go.jp/english/press/2024/0214_001.html.

[iv] European Commission, "Commission Establishes AI Office to Strengthen EU Leadership in Safe and Trustworthy Artificial Intelligence," May 29, 2024, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2982.

[v] Infocomm Media Development Authority (IMDA), "Digital Trust Centre Designated as Singapore's AI Safety Institute," May 22, 2024, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/digital-trust-centre.

[vi] Ministry of Science and ICT (MSIT), "'AI Safety Institute' Launched Following the AI Seoul Summit in May," press release, November 27, 2024, https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&nttSeqNo=1058.

[vii] Innovation, Science and Economic Development Canada, "Canada Launches Canadian Artificial Intelligence Safety Institute," news release, November 12, 2024, https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html.

[viii] Direction générale des Entreprises, "Le Gouvernement annonce la création de l'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle (INESIA)," Ministère de l'Économie, des Finances et de la Souveraineté industrielle et numérique, January 31, 2025, https://www.entreprises.gouv.fr/espace-presse/le-gouvernement-annonce-la-creation-de-linstitut-national-pour-levaluation-et-la.

[ix] Ministry of Electronics & Information Technology, "With Robust and High-End Common Computing Facility in Place, India All Set to Launch Its Own Safe & Secure Indigenous AI Model at Affordable Cost Soon: Shri Ashwini Vaishnaw," January 30, 2025, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2097709.

Sources: The linked primary sources in the table, as well as earlier overviews from Gregory C. Allen and Georgia Adamson, "The AI Safety Institute International Network: Next Steps and Recommendations," CSIS, October 30, 2024, https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations, and Alex Petropoulos, "The AI Safety Institute Network: Who, What and How?," Centre for Future Generations, September, 10, 2024, https://cfg.eu/the-ai-safety-institute-network-who-what-and-how/.

This focus on catastrophic risks is particularly noteworthy because evidence for these risks, while still inconclusive, continues to accumulate. Reports from frontier AI developers themselves, including Anthropic[5] and OpenAI,[6] have highlighted concerning capabilities in advanced systems. These reports document instances where AI systems demonstrate deceptive behaviors to achieve objectives,[7] feign alignment with human values,[8] and in some cases circumvent oversight mechanisms specifically designed to ensure their safety. These emerging capabilities pose substantial challenges for risk mitigation efforts,[9] such as in developing effective measurement and testing protocols for advanced AI systems.

Making Sense of China's AISI Equivalent: Institutional Design and Key Actors

CnAISDA's networked structure brings together three types of institutions: prestigious academic institutions such as Tsinghua University, government-backed research centers such as the Beijing Academy of Artificial Intelligence (BAAI), and research groups housed within the Ministry of Industry and Information Technology (MIIT). Rather than building new bureaucracy, China has assembled its leading organizations already engaged in technical and policy research on advanced AI.[10]

Three fundamental characteristics of CnAISDA reveal China's strategic priorities in AI governance. First, CnAISDA emphasizes international representation over domestic functions such as testing and evaluations, positioning China as an engaged participant in global governance discussions without imposing binding frontier AI safety requirements on domestic developers. Second, by leveraging existing expertise across multiple institutions, China avoids designating a single "winner" in its AI safety ecosystem, providing the CCP with greater policy flexibility. Third, while maintaining an ambiguous relationship to government, CnAISDA elevates the platform for experts with strong government connections, particularly those at Tsinghua University.

International Engagement as a Primary Function

This institutional design supports CnAISDA's apparent primary purpose: serving as a centralized hub for engagement with international counterparts. At both an in-person event[11] and in an English language op-ed,[12] the organization was explicitly described as "China's version of an AI Safety Institute (AISI)." Older AISIs often combine both domestic-level work and international engagement. For example, AISIs in the United States and United Kingdom carry out safety evaluations on AI systems through companies located in those

countries; they, for example, test for offensive or dual-use biological or cybersecurity capabilities.[13] They also conduct research and development, such as exploring "safety cases"[14]—rigorously structured rationales that developers create to articulate clearly why their AI systems are unlikely to cause catastrophic harm.[15] At the same time, these AISIs have various formal or informal engagements with counterparts elsewhere. In contrast, CnAISDA shows few signs of assuming domestic-level AI governance and safety functions. Rather, it consolidates existing expertise to create a unified front for China's international engagement on frontier AI risks.

By consolidating expertise from multiple institutions under one banner, CnAISDA creates a unified voice for China in international AI governance discussions without necessarily constraining domestic innovation through new regulatory mechanisms. Critically, it reveals China's willingness to engage on frontier AI issues outside of the UN system, traditionally China's preferred mode of engagement.[16] Its international focus was further reflected both in the substance and location of its launch event on the sidelines of the Paris AI Action Summit in February 2025, a critical juncture we discuss in greater depth below.

Core Institutions That Shape China's Networked Approach

CnAISDA brings together China's most significant institutions focused on frontier AI risks (see Table 2). In so doing, it has created a designated entity for international representation in important global AI fora without picking a domestic champion to own foundational frontier AI policy functions, providing the CCP with regulatory flexibility. While each participating organization maintains its independent identity, their collective involvement in CnAISDA elevates their political capital and potential influence on China's AI governance trajectory.

While CnAISDA does not itself test and evaluate frontier models like the U.S. and UK AISIs do, several member institutions have already conducted substantial work on AI safety evaluations. These include the China Academy of Information and Communications Technology (CAICT), Shanghai AI Laboratory, and Beijing Academy of Artificial Intelligence. These groups conduct evaluations ranging from testing for conventional risks that concern the Chinese government (such as outputs that China considers harmful to its national image) to assessing catastrophic risks in ways similar to international evaluations (such as examining whether models could advise users on creating dangerous chemicals).[17] However, the quality and nature of these evaluations is sometimes unclear.[18]

The most surprising participant in CnAISDA is the China Center for Information Industry Development (CCID). CCID has engaged in some AI policy work internationally, such as participating in track 2 dialogues on the digital economy.[19] Nonetheless, compared to the other institutions involved in CnAISDA, CCID has been a small player in international AI policy discussions, raising questions about its specific contribution to the association.

CCID sits under MIIT.[20] It conducts a variety of activities, including providing research and technical services to the government and offering operational support to industry bodies. CCID is described as the lead coordinating entity for over twenty organizations, including the China Semiconductor Industry Association. Another arm of its work is in the military industry, including what it describes as testing services and incubation and conversion of the products of military-civilian innovation.[21]

With little connection to AI specifically, CCID's role in CnAISDA could be to provide coordination and support services, as it seems to do for industry associations. On the other hand, as a body broadly focused on industrial development, CCID may be intended as a representative of the development side of the AI ecosystem within CnAISDA to balance out the safety constituents.

The real impact of CnAISDA will likely emerge not through the association itself but through the enhanced influence of the existing institutions and experts who have gained increased political capital through its successful launch. These include Chinese policy and legal experts involved in ongoing debates over frontier AI regulation, as well as technical specialists focused on AI safety research.[22]

Elevation of Outside Expertise Connected to the Government

CnAISDA elevates existing experts and heightens their international platform, revealing a distinctive approach to policy entrepreneurship in China's technology governance. This arrangement creates a strategic balance: while CnAISDA maintains a deliberately ambiguous governmental status, it provides an enhanced platform for experts who possess strong connections to government decisionmaking channels. Rather than creating entirely new bureaucratic structures, CnAISDA formalizes and amplifies the influence of established voices in China's AI governance discussions. In bringing together leaders representing a broad network of research institutions, universities, and ministerial units, CnAISDA allows its AI luminaries to maintain their independent domestic functions while gaining greater visibility and legitimacy in international forums through their collective association.

Table 2. Institutions Comprising China's AI Safety and Development Association

Institution	Description	Individuals publicly named as experts involved in CnAISDA
Beijing Academy of Artificial Intelligence (BAAI, 北京智源研究院)	State-backed research institution, primarily focused on AI development but has done some research relevant to safety. BAAI was added to the U.S. government's "Entity List" in March 2025.	Wang Zhongyuan (王仲元, President of BAAI) Zhang Hongjiang (张宏江, Founding Chairman of BAAI)
China Academy of Information and Communications Technology (CAICT, 中国信息通信研究院)	Influential think tank housed within MIIT, focusing on a range of emerging technology issues. On AI specifically, it has written papers about governing large AI models and has carried out AI evaluations.	Wei Kai (魏凯, Director of AI Research Institute at CAICT) Wei Liang (魏亮, Vice President of CAICT)
China Center for Information Industry Development (CCID, 中国电子信息产业发展研究院)	Research group within MIIT. Its role within CnAISDA might be to coordinate the various participating institutions.	Hu Guodong (胡国栋, Director of Key Laboratory of AI Scenario Application and AI System Evaluation at MIIT)
Institute of Automation, Chinese Academy of Sciences (CAS, 中国科学院自动化研究所)	Research group led by Zeng Yi, a scientist active in representing China in influential international fora such as the United Nations.	Zeng Yi (曾毅, Professor at Institute of Automation, CAS)
Peking University (北京大学)	One of the two most prestigious universities in China and the country's second-oldest. It is unclear which specific parts of the university are involved in CnAISDA.	No publicly listed experts on its website.
Shanghai Artificial Intelligence Laboratory (上海人工智能实验室)	Similarly to BAAI, a state-backed research institution. The laboratory has done work both on AI development and AI safety, and its leadership has expressed concerns about AI safety.	Zhou Bowen (周伯文, Director and Chief Scientist of Shanghai AI Laboratory)
Shanghai Qi Zhi Institute (上海期智研究院)	Research group led by Andrew Yao.	Andrew Yao (姚期智, Dean of Shanghai Qi Zhi Institute)
Tsinghua University (清华大学)	One of the two most prestigious universities in China with especially strong expertise in STEM fields, including AI. Tsinghua houses multiple institutes engaged in frontier AI safety and governance, including: · Institute for AI International Governance (I-AIIG). I-AIIG researchers have published papers about the governance of advanced AI and have taken part in track 2 dialogues about risks from advanced AI, such as the Brookings-(China) Center for International Security and Strategy dialogue. · Institute for Interdisciplinary Information Sciences.	Xue Lan (薛澜, Dean of Institute for AI International Governance; Schwarzman College) Xu Wei (徐葳, Professor at Institute for Interdisciplinary Information Sciences) Andrew Yao (姚期智, Dean of Institute for Interdisciplinary Information Sciences; College of AI)

Sources: For the participating institutions and experts, see CnAISDA's website, accessed May 28, 2025. For the descriptions, unless otherwise indicated, information is drawn from Oliver Guest, "Chinese AI Safety Institute Counterparts," Institute for AI Policy and Strategy, October 30, 2024, https://www.iaps.ai/research/china-aisi-counterparts as well as CnAISDA's website. For the Entity List, see Bureau of Industry and Security, U.S. Department of Commerce, "Additions to the Entity List," Federal Register 90, no. 60 (March 28, 2025): 14046–14052, https://www.federalregister.gov/documents/2025/03/28/2025-05427/additions-to-the-entity-list.

While CnAISDA unites various elements of China's AI ecosystem, Tsinghua University emerges as the clear intellectual and organizational nucleus of the association: CnAISDA's listed phone number and address are both in Tsinghua. Additionally, Tsinghua is well-represented among CnAISDA's experts—and in particular among those who have publicly expressed concern about catastrophic AI risks.[23]

Tsinghua's centrality within CnAISDA reflects its unique position in China's political and academic landscape. In China's governance system, academics often serve as main conduits for policy ideas, providing technical expertise and international perspectives and frequently acting as advisers to high-level leaders inside the CCP. University presidents are generally CCP officials; consequently, research directions and institutional priorities remain consonant with broader national objectives. Within this system, Tsinghua is a preeminent institution in China for science, technology, engineering, and mathematics (STEM).

Furthermore, Tsinghua's historical orientation as an internationally facing institution enhances its role. Since its founding with American educational influences, the university has maintained extensive international academic partnerships that bring global talent to Beijing. This existing international infrastructure and outlook naturally make it a critical part of China's AI policy conversation.

Tsinghua University has attracted a combination of AI luminaries privy to China's foreign policy agenda, domestic AI policy, and international AI developments. Significant Tsinghua figures include Turing Award winner Andrew Yao, top AI policy expert Xue Lan, and former vice minister of foreign affairs Fu Ying, and each brings complementary expertise to China's AI safety discussions (see Box 1). This concentration of STEM policy talent has made Tsinghua the natural institutional home for frontier technology governance discussions. The university has cultivated a reputation for producing China's technical elite while maintaining robust connections to government decisionmakers. Its position at the nexus of technical expertise and political influence enables Tsinghua to bridge theoretical concerns about AI safety with practical policy considerations. These figures could also form part of a potential secretariat for CnAISDA, should one be formally created.

While Tsinghua plays a unique role at the intersection of international AI policy, technical research, and start-up incubation, CnAISDA also incorporates expertise from other influential institutions, such as Zeng Yi, a professor at the Institute of Automation in the Chinese Academy of Sciences.

Box 1. Notable Figures Involved in Building CnAISDA

Andrew Yao

Yao has been described as a "giant" in the field of computer science.[i] He is a Turing Award winner and has received a public letter of praise from Xi for his scientific work.[ii] His organization within Tsinghua, the Institute for Interdisciplinary Information Sciences, houses the Yao Class, one of China's top undergraduate STEM programs and feeder of some of China's leading AI start-ups and research institutions.[iii] He also leads a relatively new organization within Tsinghua, the College of AI. The college describes its goals as performing pioneering AI research and cultivating top AI talent.[iv]

Yao has been one of the most outspoken Chinese voices warning about severe AI risks at the international level. He is one of the main conveners of the International Dialogues of AI Safety (IDAIS)[v] Under the aegis of IDAIS, Chinese and Western experts have published a series of statements warning of existential risks to humanity from rogue or maliciously used advanced AI systems.

He has also made strong claims about AI risks at conferences *within* China. For example, at the 2024 World AI Conference & High-Level Meeting on Global AI Governance in Shanghai, he said the following:[vi]

> We have suddenly found a way to create a new species that is many, many, times more powerful than we are. And are we sure we can live with it? Certainly, if we don't do anything, we are going to be eliminated. There's absolutely no question about it. Whether due to the nature of the computer or due to bad, malicious, actors, I think there would be a lot of destruction.

[i] Matt Sheehan, "China's Views on AI Safety Are Changing—Quickly," Carnegie Endowment for International Peace, August 27, 2024, https://carnegieendowment.org/research/2024/08/china-artificial-intelligence-ai-safety-regulation?lang=en.
[ii] Xinhua, "习近平给中国科学院院士、清华大学教授姚期智的回信," 中华人民共和国中央人民政府, July 12, 2024, https://www.gov.cn/yaowen/liebiao/202406/content_6956875.htm.
[iii] Matt Sheehan, "China's Views on AI Safety Are Changing—Quickly,"Carnegie Endowment for International Peace, August 27, 2024, https://carnegieendowment.org/research/2024/08/china-artificial-intelligence-ai-safety-regulation?lang=en.
[iv] "094 College of Artificial Intelligence, Tsinghua University, October 17, 2024, https://yz.tsinghua.edu.cn/en/info/1014/1471.htm.
[v] "International Dialogues on AI Safety," https://idais.ai/.
[vi] "2024世界人工智能大会暨人工智能全球治理高级别会议开幕式," World AI Conference 2024, July 4, 2024, https://online2024.worldaic.com.cn/forumdetail?uuid=4f83f98e8efe4224994a7da45e8a4986.

Xue Lan

Xue leads I-AIIG, a research institution within Tsinghua University focused on emerging technology policy.[i] I-AIIG has provided recommendations for Chinese domestic AI governance, as well as analysis of AI governance efforts elsewhere. I-AIIG has also organized international conferences that include discussions on frontier AI risks. Additionally, Xue is the dean of Tsinghua's prestigious Schwarzman Scholars master's program; about 80 percent of the program's student body originates from outside China.[ii]

Xue has been outspoken at the international level about frontier AI risks. He is a signatory of two of the three statements from IDAIS. Along with Yao, he is among the who's who of high-profile authors of the "Managing Extreme AI Risks Amid Rapid Progress" journal article in *Science*.[iii] The article warns of rapid advancements in AI systems that are so capable they can be used for large-scale misuse—or even cause "an irreversible loss of human control over autonomous AI systems."

He has also worked on frontier safety within China. He was the chair of an advisory body for the Ministry of Science and Technology, a body that published a document on "Ethical Norms for New Generation Artificial Intelligence" in 2021.[iv] These principles included a call to ensure that AI is always under human control. Additionally, he was the lead drafter of a paper published by several Chinese institutions describing AI safety as a global public good.[v]

i Karson Elmgren and Oliver Guest, "Chinese AISI Counterparts," Institute for AI Policy and Strategy, October 2024, https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/672261e1cb8fe024f2d10f9c/1730306539334/Chinese+AISI+Counterparts.pdf.

ii "Schwarzman Scholars Admissions Brochure," Schwarzman Scholars, https://www.schwarzmanscholars.org/wp-content/uploads/2023/03/Schwarzman-Scholars-Admissions-Brochure.pdf

iii Yoshua Bengio et al., "Managing Extreme AI Risks amid Rapid Progress," Science 384, no. 6698 (May 20, 2024): 842–845, https://www.science.org/doi/10.1126/science.adn0117.

iv National New Generation Artificial Intelligence Governance Specialist Committee, Ethical Norms for New Generation Artificial Intelligence Released, translated by Etcetera Language Group, Inc., edited by Ben Murphy, Center for Security and Emerging Technology, October 21, 2021, https://cset.georgetown.edu/publication/ethical-norms-for-new-generation-artificial-intelligence-released/.

v "AI Safety as Global Public Goods Working Report," Shanghai Jiaotong University, July 5, 2024, https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20240704/%E7%A0%94%E7%A9%B6%E6%8A%A5%E5%91%8A%E6%89%8B%E5%86%8C-04.pdf.

Fu Ying

Fu is not mentioned on the CnAISDA website. But she appears to be closely connected to it, publishing an op-ed discussing the launch of the association.[i]

She is China's former vice minister of foreign affairs and one of the highest-ranking women in the history of the ministry. She now leads Tsinghua's Center for International Security and Strategy, which has been engaged in long-standing international dialogues on AI and national security with U.S. think tanks.[ii] As early as 2019, Fu led the development of a set of principles for AI, which includes the idea that AI should be safe/secure and controllable.[iii]

[i] Fu Ying, "Cooperation for AI Safety Must Transcend Geopolitical Interference," South China Morning Post, February 12, 2025, https://www.scmp.com/opinion/china-opinion/article/3298281/cooperation-ai-safety-must-transcend-geopolitical-interference.
[ii] Ryan Hass and Colin Kahl, "Laying the groundwork for US-China AI dialogue," Brookings Institution, April 5, 2024, https://www.brookings.edu/articles/laying-the-groundwork-for-us-china-ai-dialogue/.
[iii] Jeffrey Ding, "ChinaAI #67: Fu Ying on AI + the International Order," ChinAI (newsletter), July 5, 2024, https://chinai.substack.com/p/chinai-67-fu-ying-on-ai-the-international?open=false#%C2%A7feature-translation-fu-yings-preliminary-analysis-of-ai-international-relations.

Zeng Yi

Zeng is a professor at the Institute of Automation, Chinese Academy of Sciences[iv] His research focuses on brain-inspired artificial intelligence, AI ethics and governance, and AI safety.

At the international level, Zeng has been active in AI safety and governance forums. He served as a member of the UN High-level Advisory Body on Artificial Intelligence[v] and has briefed the UN Security Council on AI risks,[vi] warning about potential extinction risks from advanced AI systems. He is also a contributor to the International AI Safety Report 2025,[vii] which attempts to build international consensus around the capabilities and risks of advanced AI. He signed the Center for AI Safety's statement on extinction risk from AI, which equates mitigating AI extinction risk with societal-scale threats like pandemics and nuclear war.[viii]

Within China, Zeng sits on the National Governance Committee of New Generation Artificial Intelligence, a high-level body that helps shape China's AI governance policies.[ix] He directs the Beijing Institute of AI Safety and Governance (Beijing-AISI),[x] which appears to carry out safety-relevant evaluations of Chinese models.

[iv] Will Henshall, "Yi Zeng," TIME, September 7, 2023, https://time.com/collection/time100-ai/6308795/yi-zeng/.
[vi] United Nations, "Members of the High-level Advisory Body on Artificial Intelligence," accessed June 6, 2025, https://www.un.org/en/ai-advisory-body/members.
[vi] United Nations Security Council, "Artificial Intelligence: Opportunities and Risks for International Peace and Security," 9381st meeting, October 18, 2023, UN Web TV, https://webtv.un.org/en/asset/k1j/k1ji81po8p.
[vii] Yoshua Bengio et al., International AI Safety Report 2025, Department for Science, Innovation and Technology and AI Safety Institute, January 29, 2025, https://www.gov.uk/government/publications/international-ai-safety-report-2025.
[viii] "Statement on AI Risk," Center for AI Safety, https://www.safe.ai/work/statement-on-ai-risk.
[ix] Zhang Na, "加强量子人工智能伦理治理," 中国社会科学网, March 21, 2025, https://www.cssn.cn/skgz/bwyc/202503/t20250321_5859153.shtml#:~:text=Zeng%20Yi%2C%20a%20member%20of%20the%20National%20New%20Generation%20Artificial%20Intelligence%20Governance%20Professional%20Committee.
[x] Beijing Institute of AI Safety and Governance, "Leaders and Scientists," accessed June 6, 2025, https://beijing.ai-safety-and-governance.institute/people.

Tracing the Origins of Frontier AI Governance in China

CnAISDA's establishment represents the culmination of years of strategic positioning by policy entrepreneurs within China's AI ecosystem. Understanding how these domestic AI safety advocates successfully navigated China's political landscape can offer critical insight into the future trajectory of China's AI policy landscape. The establishment of CnAISDA emerged from a dual process: engaging in international AI safety conversations while navigating a complex constellation of domestic AI developments and priorities. It also offers a remarkable case study of how ideas can diffuse into different political contexts without requiring formal international institutions. The story of how China's policymakers simultaneously navigated these domestic and international environments may illuminate genuinely tractable pathways for informal global AI coordination more broadly.

Concerns about catastrophic risks from AI have percolated among technical experts in the Chinese ecosystem since at least the late 2010s.[24] They first surfaced significantly in the policy ecosystem in 2021 with the publication of "Ethical Norms for New Generation Artificial Intelligence" by the Ministry of Science and Technology, which included a call to "ensure that AI is always under human control."[25] That year, China introduced regulation on recommendation algorithms;[26] the year after, it began to regulate deepfakes through its "deep synthesis" regulation,[27] which requires synthetically generated content to be conspicuously labelled. While Chinese policymakers clearly saw risks tied to powerful future AI systems, their immediate priority seemed to be "content security"—ensuring that generated outputs did not undermine the strictly managed information environment the CCP has cultivated.[28] This focus also motivated China's 2023 generative AI regulation, which requires providers to conduct safety/security evaluations focused primarily on preventing politically sensitive content.[29]

By 2023, China's domestic AI conversation had evolved in two significant ways that would shape CnAISDA. First, the focus on content security had moderated to place increased emphasis on promoting AI innovation, a manifestation of the CCP's desire to stimulate economic growth in the aftermath of its COVID-19 lockdowns.[30] This was reflected, for example, in the final version of China's generative AI regulation in July 2023, which, in contrast to the more restrictive draft version, called for equal emphasis on development and safety/security in AI.[31]

Second, the conversation inside China had evolved from a narrow focus on current risks and harms—as viewed by the CCP—to include a more forward-looking conversation around potential catastrophic risks from future systems. Technical research in China related to AI safety, including in some cases explicit attention to catastrophic risks, had shown a noticeable uptick.[32] For example, Chinese scientists began producing increasingly sophisticated work on AI alignment, such as comprehensive surveys of methods to ensure AI systems remain aligned with human intentions as they grow more capable.[33] Chinese scientists also

developed safety benchmarks like SALAD-Bench that specifically evaluate large language models for catastrophic risk dimensions, such as enabling chemical, biological, radiological, and nuclear threats; cyber attacks; and psychological manipulation capabilities.[34]

Parallel to these domestic developments, Chinese experts were becoming increasingly embedded in international forums dedicated to frontier AI safety.

The International Currents That Changed China's AI Safety Conversation

While China's domestic AI safety conversation was evolving internally, a parallel process of international engagement was simultaneously reshaping how Chinese actors approached frontier AI risks. This international engagement took shape through a series of coordinated statements, international summits, and scientific collaborations that positioned Chinese experts alongside Western and global counterparts in acknowledging frontier AI risks (see Figure 1). The engagement established a pattern of participation that would eventually lead to a more formal institutional representation of China's interests in frontier AI safety abroad.

In May 2023, leading Chinese academics and industry leaders joined over 100 leading scientists and thinkers globally in signing a one-sentence statement published by CAIS that declared, "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[35] Later in October, some of China's most prominent AI scientists joined Western counterparts in signing an open letter in Oxford.[36] Its headline claim was that "coordinated global action on AI safety research and governance is critical to prevent uncontrolled frontier AI development from posing unacceptable risks to humanity." These developments signaled Chinese experts' growing recognition of AI safety as a critical international concern requiring coordinated action across geopolitical boundaries.

The early expressions of concern from Chinese experts would soon transform into more formalized institutional engagement at a watershed international summit. The international origin story of China's AI Safety Institute counterpart, like that of other AISIs, can be traced to Bletchley Park, England. The United Kingdom organized an international summit on AI and decided to make major risks at the frontier of AI capabilities the primary focus. It also strategically calculated that it was critical to invite China,[37] despite generating significant controversy for doing so.[38] Held in November 2023, the summit ultimately brought together the United States, China, European Union, United Kingdom, and twenty-five other countries to sign a statement known as the Bletchley Declaration. The declaration noted the "potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of frontier AI models."[39]

At the summit, the United Kingdom and United States simultaneously announced the creation of the first-ever government institutions dedicated to contributing to the safety

of AI systems—the UK and U.S. AISIs.[40] In addition to participating in AI discussions globally, the UK AISI announced plans for testing and evaluating models for future risks, such as whether increasingly powerful large language models can make it easier for a novice to launch a cyber attack.[41]

Beyond establishing domestic institutions, the United Kingdom also worked to establish a global consensus on the state of frontier AI capabilities and risks that could ground AI governance conversations globally. It engaged Turing Award-winning AI scientist Yoshua Bengio, based in Canada, to head the effort. The resulting International AI Safety Report was overseen in part by three top Chinese scientists.[42] Zeng Yi served on the expert advisory panel while Zhang Ya-Qin served as senior advisers. Additionally, the report's writing team included Kwan Yee Ng of Concordia AI, a Beijing-based organization that began convening safety discussions in China.[43]

The Bletchley summit had come at a critical juncture in China's own AI governance story—just a few weeks earlier, China had launched its Global AI Governance Initiative,[44] outlining its vision for international AI governance engagement. But at Bletchley, Beijing joined a truly global discussion, one in which its leading scientists and policymakers were exposed to new institutions that were taking testing and evaluations and global, government-backed coordination to address frontier AI risks seriously. Ideas organically developing within China's frontier AI ecosystem were integrated into an international movement to ensure the safety and security of increasingly powerful AI systems globally.

China's AI Thinkers Maneuver to Establish Their Own Institute

For some of China's leading AI thinkers, the establishment of the UK and U.S. AISIs in late 2023 opened up the possibility for more formalized avenues of AI safety coordination. And they became increasingly motivated to set up a counterpart institution of their own. The opportunity to participate in the global AI governance conversation galvanized a select group of policy entrepreneurs—strategically positioned at the intersection of government connections and international expertise—to bridge international ideas and situate them within China's evolving domestic AI landscape.

Initially, despite the Chinese AI community's increasing engagement with safety concerns, there was no immediate indication that China might establish its own AISI—not until the prestigious BAAI conference in June 2024, which included a panel discussion with Chinese and Western participants about AI safety.[45] Responding to comments from U.S.-based AI luminary Max Tegmark, several of the Chinese panelists expressed enthusiasm about a potential Chinese AISI. Then, in July, the CCP gave its most prominent, authoritative endorsement of AI safety as a policy priority yet. The decision of the Third Plenum,[46] a meeting for CCP leadership to unveil its economic and social vision for the next five years, included a goal of "establishing an AI safety supervision and regulation system."[47] While the document left unstated what kinds of risks such a system was intended to address, it included this

goal in the context of major public safety issues such as pandemics and cybersecurity. The official explainer for the decision referred to risks to employment, privacy, and "the norms of international relations."[48] But a follow-up *People's Daily* op-ed by a senior government official in Jiangsu Province discussing this goal also called explicitly for research on "frontier safety technology" and "safety and controllability of general-purpose large models."[49] Though the Third Plenum decision did not necessarily imply that China planned to establish an AISI, some took it as a sign that top leadership might view frontier AI safety as a priority.

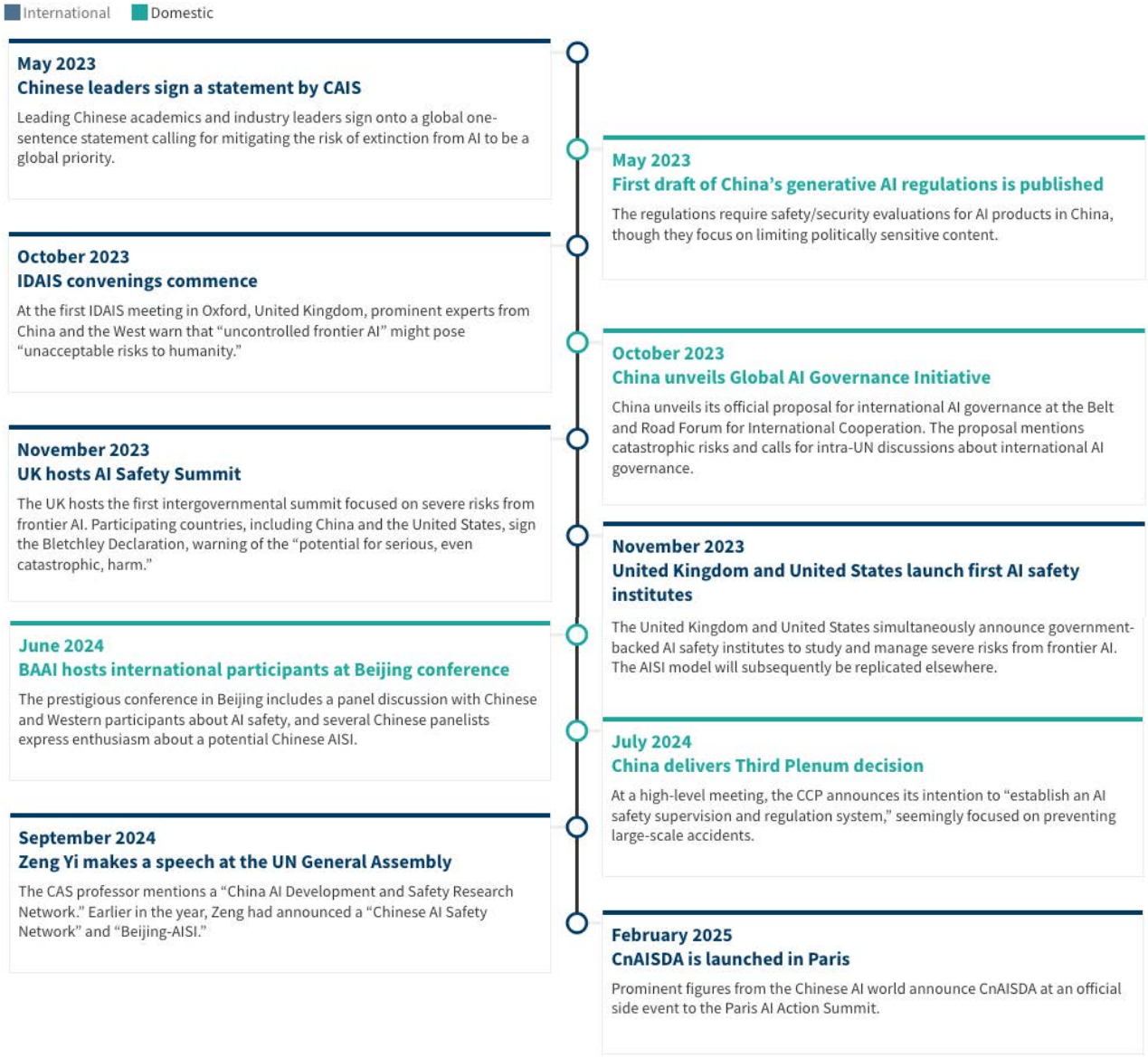
Over the next few months, rumors began to spread about a possible Chinese AISI being established. Chinese stakeholders seemed to be debating what form it would take, who would lead it, and who would be involved. A number of different groups also appeared to be making bids to become China's AISI, with institutions being established in both Beijing and Shanghai and websites popping up that purported to represent a Chinese AISI. In September 2024, Zeng Yi gave a speech at the United Nations in which one slide mentioned a "China AI Development and Safety Research Network," which was presented as roughly analogous to an AISI.[50] Zeng had previously announced a "Chinese AI Safety Network" in June 2024[51] and later led the "Beijing-AISI,"[52] which functions like an AISI for the Beijing municipality, a Chinese administrative unit with similar status to a province. Over the remainder of the year, there were several indications that an official, government-backed, national Chinese AISI was likely to emerge, perhaps imminently. But months passed with no official public announcement.

China's AI policy community was not alone in exploring how to engage more deeply in the global frontier AI conversation. Internationally, the number, institutional form, and remit of AISIs expanded globally.[53] By the Paris AI Action Summit in February 2025, ten jurisdictions had launched AISIs. Yet, while international participation had increased, the initial AISI focus on catastrophic risks had diminished. For example, when France[54] and India[55] announced their AISI counterparts in January, neither discussed the possibility that advanced AI systems might be difficult to keep under control. Moreover, both France and India joined Canada in establishing a new structure for their AISIs: as opposed to having just one entity serving as its AISI, these countries distributed the mandate among a network of several existing organizations. These design choices raised questions for China should it establish its own AISI—how to define its structure and scope—situating a domestic institution within an increasingly diverse set of AI-focused institutions internationally.

China's AI Safety Institute Launches in Paris

The Paris AI Action Summit finally brought clarity to both the nature and structure of China's AISI equivalent. CnAISDA made its public debut at an official side event titled "Promoting International Cooperation on AI Safety and Inclusive Development." Co-hosted by the Institute for AI International Governance and Shanghai Qizhi Institute, the launch event's focus reinforced CnAISDA's primary function as a vehicle for international engagement rather than domestic functions.

Figure 1. How China's AI Safety Institute Was Built



A document presented at CnAISDA's launch event (see Appendix A) notes that "various countries have established AI safety research organizations, including national AI safety institutions and associations, to accelerate evaluation, research, and standard setting," largely paralleling the three functions the UK AISI had outlined when it was first established in 2023.[56] CnAISDA envisions a global AI safety governance framework that combines UN engagement with global AI safety institute engagement to "[foster] global dialogue on common risks and [share] best practices." The initiative calls for collaboration to address "misuse, abuse, and malicious use" risks, including both near-term risks such as the manipulation of public opinion and disinformation, along with speculative risks like the deployment of AI by terrorist organizations. It also calls for establishing scientific consensus on risks and redlines and for "early warning thresholds for AI systems that may pose catastrophic or existential risks to humans." And it expresses a desire to enhance international cooperation on standards setting, technological safeguards, and global AI safety capacity building.

While much of CnAISDA was developed quietly, it now has a clear public champion. Fu Ying, China's former vice minister of foreign affairs and one the of highest ranking women in the ministry's history, announced both at a Paris AI Action Summit side event[57] and an English-language op-ed in the *South China Morning Post* that CnAISDA is equivalent to other AISIs globally.[58] Her op-ed raised concerns about the safety of current AI applications, along with possible risks from future systems. Acknowledging today's geopolitical environment, she wrote, "Realistically, few can see much promise as geopolitical tensions continue to cast a shadow over scientific collaboration." She pushed for the great AI powers to carve out a lane to address catastrophic risks.

China's AI ecosystem possesses many of the world's leading developers of open-source systems. In her op-ed, Fu concluded with an argument for their safety compared to the closed-source systems deployed by cutting-edge developers in the West. In so doing she directly challenged Yoshua Bengio, who has warned of the potential for catastrophic risks from open-sourcing increasingly powerful frontier AI models. In countering Bengio, she noted that while open-source models may be more vulnerable to misuse, their transparent nature can improve problem detection.

Outstanding Questions and Strategic Implications

Several open questions remain regarding CnAISDA at a functional level. First, while Fu notes that CnAISDA was "established with government support," it is not precisely clear what the relationship is between CnAISDA and the Chinese government. Several

governmental bodies, including the Cyberspace Administration of China, Ministry of Science and Technology, and MIIT, have collectively shaped China's AI regulatory ecosystem.[59] But it is possible that the National Development and Reform Commission, China's macroeconomic regulator that has recently played a larger role in AI governance, was the executive department that lent its support to CnAISDA. Where CnAISDA settles within or around China's bureaucracy will offer clues into the broader balance of power among the executive departments shaping China's domestic AI regulation.

Second, it is unclear what responsibilities CnAISDA will ultimately assume and to what extent they will go beyond those of the existing member institutions. The UK AISI was established with a £50 million annual budget (approximately $68 million) and over 100 full-time staff,[60] but at present, CnAISDA does not appear to have separately dedicated staff, research teams, or budget. The website does not seem to have a Chinese-language version and is only available in English. In any case, even if CnAISDA does not represent net-new capacity in Chinese AI governance on either domestic or international topics, providing a clear docking point for international coordination could be helpful to facilitate effective governance.

Finally, it seems that CnAISDA's official English name could still be subject to revision. Fu's op-ed refers to China's AI Safety and Development "Network," as opposed to "Association."[61] China is not unique in potentially changing the name of its AISI soon after launch. The United Kingdom, for example, renamed its institute about a year after the AISI's establishment, and the UK AISI itself evolved out of the earlier Frontier AI Taskforce. Most recently, the United States renamed its AISI to the Center for AI Standards and Innovation.

These naming considerations are clearly strategic and reflect sensitivity to the need to balance both international and domestic legitimacy. Foundational documents outlining the CCP's vision for frontier AI policy, such as its generative AI regulation, refer to the importance of placing equal emphasis on both development and safety/security.62 CnAISDA's name clearly builds on this notion but shapes its messaging based on linguistic context. In the English name, "safety" comes before "development," while in Mandarin the order is reversed. This subtle difference may reflect an effort to emphasize safety concerns when engaging with international audiences while maintaining development as the primary focus domestically. The fluctuation between "association" and "network" might also indicate changing priorities between faithfully translating the Mandarin name and other goals, such as avoiding confusion with the similarly named International Network of AI Safety Institutes. The relevant term in Mandarin, 网络 (wǎngluò) has remained consistent and is directly translated as "network."

U.S.-China Coordination Amid International Uncertainty

The establishment of CnAISDA—designed explicitly for coordination with global counterparts—represents China's formal entry into the international AI safety governance conversation. However, this milestone arrives at a moment of significant flux in the international landscape, particularly in U.S. policy toward frontier AI governance.

The U.S. approach to AI safety has undergone substantial recalibration following the 2024 presidential transition. Within hours of taking office in January 2025, President Donald Trump rescinded the 2023 AI executive order that had established much of the previous administration's framework for AI governance.[63] This shift was further emphasized at the Paris AI Action Summit, where Vice President JD Vance articulated a new priority framework, stating, "The AI future is not going to be won by hand-wringing about safety."[64] While not dismissing risk concerns entirely, the administration has clearly pivoted toward what Vance termed "AI opportunity."

These policy shifts have created significant uncertainty for the institutional infrastructure of U.S. government–based efforts to address frontier AI safety risks. The U.S. AISI—a cornerstone of the previous administration's approach and inaugural chair of the International Network of AI Safety Institutes[65]—has faced an unclear future, especially when its founding director resigned shortly after the presidential transition.[66] Meanwhile, the U.S.-led international coordination network is now confronting potential funding gaps, as $3.8 million of the $11 million the United States committed to the network was designated through the U.S. Agency for International Development, which has been largely gutted.[67] However, there may be reason for cautious optimism: in June 2025, the Department of Commerce rebranded the U.S. AISI as the Center for AI Standards and Innovation, keeping many of its core functions.[68]

Prospects for bilateral U.S.-China engagement on catastrophic AI risks remain limited. U.S. policymakers' skepticism about technical cooperation with China, especially amidst concerns about inadvertently improving dual-use capabilities.[69] Those risks, coupled with proposed legislation to restrict AI research collaboration with China,[70] suggests minimal U.S. appetite for engagement with CnAISDA. In addition, BAAI has been placed on the U.S. Bureau of Industry and Security's Entity List,[71] which could erect further barriers to interaction with CnAISDA's constituent institutions.

Regardless of how U.S. policy evolves, CnAISDA may find engagement opportunities with other international partners. The UK AISI has previously reached out to Chinese organizations,[72] and Singapore has collaborated with BAAI on AI red-teaming:[73] adversarial testing designed to identify vulnerabilities and safety risks in AI systems. Which partners engage with CnAISDA will likely shape its agenda and focus areas.

Given this complex and uncertain international environment, CnAISDA faces both limitations and opportunities in its early development. The shifting U.S. approach to AI safety governance creates a potential leadership vacuum in global coordination efforts that China could partially fill through strategic engagement with receptive international partners. Nonetheless, to the extent CnAISDA's leadership believes that the safety and security of U.S. models will impact Chinese national security, its leaders will need to explore creative ways to develop channels for coordination with U.S. counterparts, such as by developing domestically palatable counterparts to initiatives the United States is already engaged in.

Amid Boisterous AI Development, Questions on China's AI Safety and Governance Loom

Domestically, the establishment of CnAISDA arrives at a particularly salient, even triumphant, moment for Chinese AI development in the post-DeepSeek-R1 era. China's domestic policy AI zeitgeist has focused on leveraging AI as an engine for economic growth.[74]

While China's leadership may be excited by DeepSeek's potential, they may soon need to grapple with emerging risks that their own technical AI safety community is increasingly documenting.[75] This research, along with the tangible ways AI is shaping Chinese society, could shape their future governance decisions. Advancing frontier AI capabilities carries ill-understood risks, which could arise first in a Chinese lab rather than abroad. For Chinese leaders, homegrown risks sprouting up within China's Great Firewall may be seen as more likely to cause headaches than the products of foreign companies blocked off from the average citizen to begin with.

To the extent senior Chinese leaders indeed believe catastrophic risks may be hiding in the near future of AI, this means that the ball is more in their court to anticipate and prevent them. If a major incident occurs with AI, attention could quickly swing back from promoting innovation to managing risk.

Growing evidence suggests this reality is not lost on China's highest leaders. Chinese officials have begun signaling more explicit concern about catastrophic AI risks in diplomatic and domestic contexts. In March 2025, Chinese Ambassador to the U.S. Xie Feng framed AI development as potentially "opening Pandora's box" and advocated U.S.-China cooperation on global AI governance.[76] Even more telling was the CCP Central Committee Politburo's study session dedicated to AI—only the second such collective session on this topic and a

clear indicator of leadership priorities. The April 2025 study session's official readout warned of "unprecedented risks and challenges" from advanced AI systems and outlined specific policy responses, including building "systems for technical monitoring, risk warning, and emergency response."[77]

These high-level statements and concrete policy prescriptions suggest potential regulatory action may be forthcoming.[78] They also lend significant credibility to CnAISDA's mission, indicating that the organization may help shape China's approach to AI safety despite representing a minority position within China's broader AI policy landscape—a landscape that remains heavily focused on development and competitiveness.

The Pathway Forward for China's AI Safety Policy Entrepreneurs

CnAISDA's emergence in Paris demonstrates the tenacity of China's AI ecosystem and Beijing's strong desire to engage in the international safety conversation. Having established their AISI, China's frontier AI policy pioneers must navigate choppy waters both globally and at home. Internationally, they may struggle to find other AISIs willing to engage in politically sensitive but potentially important technical collaboration, such as on joint testing and evaluations.

Constituent groups within CnAISDA have demonstrated the ability to diffuse the international AI safety conversation within China in creative ways. At the May 2024 Seoul AI Summit, frontier AI developers pledged to produce safety frameworks that establish concrete thresholds for risk that would trigger safety-mitigating action known as the Frontier AI Safety Commitments.[79] At the time, one Chinese company, Zhipu.AI, signed onto the commitments, while two others, 01.AI and MiniMax, later signed on the sidelines of the 2025 Paris summit.

More notably, however, China has established its own domestic version of the commitments,[80] paving a pathway to internalize international ideas around safety frameworks within China's domestic context. China's Artificial Intelligence Industry Alliance (AIIA),[81] a prominent industry consortium guided by MIIT, released commitments signed by seventeen Chinese AI companies,[82] including DeepSeek, Alibaba, and Tencent. Whether the CAICT and AIIA can leverage these commitments as a pathway to harmonize international standards remains to be seen. Nonetheless, the very existence of a domestic set of commitments reveals the upside of an approach in which international ideas diffuse naturally within China's domestic context on the basis of its own national interest.[83]

The existence and possible alignment of industry commitments represent clear victories for China's AI safety policy entrepreneurs. But the real challenge will be whether CnAISDA can translate the substantial political capital it developed following a successful launch in Paris into meaningful policy change in China's domestic context. At this stage, the question remains definitively unanswered.

Conclusion

The emergence of CnAISDA marks a pivotal moment in both China's approach to frontier AI governance and the global coordination landscape. The organization's distinctive institutional architecture—focused internationally rather than domestically, networked rather than centralized, and strategically positioned between government and academia—represents China's efforts to balance international opportunities with domestic political realities. This design enables meaningful participation in global AI safety conversations while preserving regulatory flexibility at home.

For CCP leadership, CnAISDA offers a useful vehicle to stay involved in global conversations on frontier AI; it is a passport to important conversations about what could be the most important technology of the century. But for a small but powerful group of leading AI policymakers, CnAISDA offers a platform that legitimizes their authentic concern for frontier AI risks. For the international community, the key question is whether this group will be able to shape China's international priorities, and more importantly, China's approach to ensuring the safety and security of its own AI models. At the very least, CnAISDA's establishment offers compelling evidence that this group has made substantial progress in shaping the AI safety conversation in China over the last two years. And increasingly concrete language describing mechanisms to combat risks suggests that they may be playing an increasingly important role in China's domestic AI conversation.

The institution's emergence illuminates a crucial pathway for global coordination: when technical ideas about AI safety complement rather than challenge national interests, they can transcend geopolitical barriers. As frontier capabilities advance rapidly across multiple countries, this insight suggests that natural diffusion of safety principles through domestically anchored institutions may prove more resilient than rigid international frameworks struggling to accommodate divergent priorities.

For international stakeholders seeking to engage China on AI safety, it may be more productive to focus less on formal architectures and more on creating conditions where shared technical concerns can be independently incorporated into China's domestic governance approach. This model of parallel evolution guided by common technical understanding offers a pragmatic path forward for addressing shared catastrophic risks even as strategic competition intensifies in other dimensions of AI development.

Appendix 1. Initiative on Promoting International Cooperation on AI Safety and Inclusive Development (Document Excerpt)

The below text was included in a document CnAISDA shared at an event that I-AIIG and the Shanghai Qizhi Institute hosted on the sidelines of the 2025 Paris AI Action Summit.

English Version

Artificial intelligence (AI) is a new area of human development. While bringing tremendous opportunities, it also presents risks and challenges. From intergovernmental AI summits to various consensuses formed within the scientific community, existing steps by the international community show various countries' strong willingness toward cooperation on AI safety and governance despite geopolitical tensions. Currently, various countries have established AI safety research organizations, including national AI safety institutes and associations, to accelerate evaluation, research, and standard setting. We believe it is essential to seize this opportunity to encourage different entities, such as national AI safety institutes, international organizations, companies, civil organizations, and individuals, to uphold the principles of extensive consultation, joint contribution, and shared benefits. Strengthening collaboration among these stakeholders will help collectively advance AI safety and governance. To this end, we propose:

1. **Building an inclusive global AI safety governance framework.** We aim to build a framework in which the United Nations acts as the central platform for governance consultations, and AI safety institutes around the world play a proactive role in fostering global dialogue on common risks and sharing best practices. Through the concerted efforts of diverse stakeholders and platforms, we shall ensure that the international community keeps abreast of the potentially most advanced AI systems and is able to curb the proliferation and operation of dangerous models.

2. **Strengthening international cooperation to prevent AI misuse, abuse, and malicious use.** While respecting international law and local legal systems, we support joint efforts of the international community to combat the manipulation of public opinion and disinformation. We must collaborate to prevent and combat the illicit use of AI by terrorist organizations, extremist forces, and transnational organized crime groups.

3. **Promoting international cooperation to prevent substantial AI risks.** We facilitate cooperation between the international scientific community and governments to establish scientific consensus in terms of risks and red lines, and set early warning thresholds for AI systems that may pose catastrophic or existential risks to humans. We encourage countries to increase investment in the research and development of AI safety technologies, and ensure that AI technologies are safe, controllable, and reliable.

4. **Facilitating a transparent and accountable cooperation mechanism for AI safety governance.** We aim to strengthen information sharing and cooperation among scientific research institutions, enterprises, and governments in safety standards, regulations, and technology research and development. We encourage collaboration among standardization organizations and promote the joint development and sharing of international multilateral platforms.

5. **Enhancing international cooperation on technological safeguard.** We call on governments to improve AI safety, controllability, and reliability across aspects such as science, technology, and engineering. This includes establishing international AI security research networks, developing scientific plans, setting up expert panels for global AI risk identification and assessment, and promoting the international exchange of security research and technical tools.

6. **Strengthening global AI safety capacity-building.** We urge AI-leading countries to support AI-developing nations in building their AI governance capabilities through infrastructure development, international cooperation platforms for AI capacity -building, safety testing and validation platforms, and strategic alignment and policy exchanges associated with AI.

Simplified Chinese Version

人工智能是人类发展的新领域，在带来新机遇的同时也产生新的风险挑战。从政府间的人工智能峰会到科学界形成的各类共识，国际社会目前的积极举措表明，即使在地缘政治局势趋于紧张的情况下，各国在人工智能发展和安全治理方面仍然怀有强烈的合作意愿。目前，各国陆续设立了安全研究所、安全研究网络等各种不同形式的人工智能安全研究机构和组织，加快推进测评、研究和标准制定工作。我们认为，应当抓住这个时机，推动各国人工智能安全研究机构、国际组织、技术企业、科研院校、民间机构和公民个人等各类主体秉持共商共建共享的理念，加强合作，协力共同促进人工智能安全与治理。为此，我们倡议：

——建立包容性的全球人工智能安全治理体系。构建以联合国为核心的风险治理协商平台，同时，发挥各国人工智能安全研究机构的重要作用，就共同面临的人工智能风险组织对话，分享人工智能安全治理最佳实践。通过多方和多平台的共同努力，确保国际社会充分了解可能存在的最先进人工智能系统，并具备遏制危险模型分发和运营的手段。

——加强国际合作防范人工智能误用、滥用和恶用。在尊重国际法和各国法律框架前提下，支持国际社会共同打击制作与传播虚假信息的行为，合作防范和打击恐怖主义、极端势力和跨国有组织犯罪集团利用人工智能技术从事非法活动的行为。

——推动国际合作防范人工智能重大风险。推进国际科学界同各国政府间的合作，在风险和红线方面进一步建立科学共识，对可能给人类带来灾难性或生存性风险的人工智能系统进行预警。鼓励各国加大对人工智能安全技术的研发投入，确保人工智能技术应用安全、可控、可靠。

——推动构建透明可问责的人工智能安全治理合作机制。加强各国科研机构、企业、政府在人工智能安全标准、监管体系建设和技术研发等领域的信息共享与合作，促进标准化组织间的合作以及国际多边平台的共建共享工作。

——加强人工智能安全技术保障的国际合作。呼吁各国政府加强合作，从科学、技术、工程等各方面持续提升人工智能的安全可控可靠水平，支持建设国际人工智能安全技术研究合作网络和科学计划，成立全球人工智能风险识别和测试评估专家组，推进国际安全技术研究交流和技术工具供给。

——加强全球人工智能安全能力建设。呼吁人工智能领先国家通过人工智能基础设施建设合作、搭建人工智能能力建设国际合作平台、共建人工智能安全评测验证平台、加强人工智能战略对接和政策交流等多种形式帮助发展中国家提高人工智能安全治理能力。

About the Authors

Scott Singer is a visiting scholar in the Technology and International Affairs Program at the Carnegie Endowment for International Peace.

Karson Elmgren is a researcher focused especially on China and U.S.-China relations on AI. He has also worked on topics ranging from compute governance to standards and evaluation.

Oliver Guest is a researcher at the Institute for AI Policy and Strategy. His research focuses on the international governance of advanced AI.

Acknowledgments

We are grateful to Renan Araujo, Jon Bateman, Corey Hinderstein, Arthur Nelson, Matt Sheehan, and Sam Winter-Levy for their feedback on previous drafts.

Notes

1 Bill Bishop, "April Politburo Study Session on AI is Bad News for Nvidia," *Sinocism* (blog), April 26, 2025, https://sinocism.com/p/april-politburo-study-session-on.

2 Meaghan Tobin, Paul Mozur, and Alexandra Stevenson, "How Chinese A.I. Start-Up DeepSeek Is Competing With Silicon Valley Giants," New York Times, January 28, 2025, https://www.nytimes.com/2025/01/28/business/deepseek-owner-china-ai.html.

3 Alexandra Stevenson, "Xi Jinping Meets With Jack Ma and Other Business Leaders in Beijing," New York Times, February 17, 2025, https://www.nytimes.com/2025/02/17/business/china-xi-jinping-jack-ma.html.

4 Renan Araujo et al., "Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges," Institute for AI Policy and Strategy, October 7, 2024, https://www.iaps.ai/research/understanding-aisis.

5 Ryan Greenblatt, et al., "Alignment Faking in Large Language Models," preprint, arXiv, December 18, 2024, https://arxiv.org/abs/2412.14093.

6 Bowen Baker et al., "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation," OpenAI, March 10, 2025, https://cdn.openai.com/pdf/34f2ada6-870f-4c26-9790-fd8d-ef56387f/CoT_Monitoring.pdf.

7 Peter S. Park, et al., "AI Deception: A Survey of Examples, Risks, and Potential Solutions," *Patterns* 5, no. 5 (2024), https://www.cell.com/patterns/fulltext/S2666-3899(24)00103-X.

8 "Alignment Faking in Large Language Models," Anthropic, December 18, 2024, https://www.anthropic.com/research/alignment-faking.

9 Rishi Bommasani, Scott Singer, et al., "Draft Report of the Joint California Policy Working Group on AI Frontier Models," March 18, 2025, https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf.

10 Caroline Meinhardt and Graham Webster, "What Do We Know About China's New AI Safety Institute?," DigiChina, February 6, 2025, https://digichina.stanford.edu/work/what-do-we-know-about-chinas-new-ai-safety-institute/.

11 Nan Boyi, "傅莹：未来AI安全、负责任应用的最佳出路需要中美两股力量", 澎湃新闻, February 12, 2025, https://m.thepaper.cn/newsDetail_forward_30139070.

12 Fu Ying, "Cooperation for AI Safety Must Transcend Geopolitical Interference," *South China Morning Post*, February 12, 2025, https://www.scmp.com/opinion/china-opinion/article/3298281/cooperation-ai-safety-must-transcend-geopolitical-interference.

13 "Pre-Deployment Evaluation of OpenAI's o1 Model," AI Security Institute, December 18, 2024, https://www.aisi.gov.uk/work/pre-deployment-evaluation-of-openais-o1-model.

14 Marie Buhl, et al., "How Can Safety Cased be Used to Help with Frontier AI Safety," AI Security Institute, February 10, 2025, https://www.aisi.gov.uk/work/how-can-safety-cases-be-used-to-help-with-frontier-ai-safety.

15 Joshua Clymer, et al., "Safety Cases: How to Justify the Safety of Advanced AI Systems," preprint, arXiv, March 15, 2024, https://arxiv.org/abs/2403.10462.

16 Oliver Guest and Kevin Wei, "Bridging the Artificial Intelligence Governance Gap," RAND Corporation, December 2024, https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3700/PEA3703-1/RAND_PEA3703-1.pdf.

17 Beijing Academy of Artificial Intelligence (BAAI), "FlagEval," uploaded October 18, 2024, https://perma.cc/KNA9-UEAN.

18 Karson Elmgren and Oliver Guest, "Chinese AISI Counterparts," Institute for AI Policy and Strategy, October 2024, https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/672261e1cb8fe024f2d10f9c/1730306539334/Chinese+AISI+Counterparts.pdf#page=41.

19 "U.S.-China Track II Dialogue on the Digital Economy Consensus Agreement," National Committee on U.S.-China Relations and China Institute for Innovation and Development Strategy, November 17, 2021, https://www.ncuscr.org/wp-content/uploads/2022/05/DE-Consensus-Agreement-Nov2021-FINAL-English-2022-05-06.pdf.

20 "中国电子信息产业发展研究院," Baidu Baike, accessed June 6, 2025, https://baike.baidu.com/item/中国电子信息产业发展研究院/8101155.

21 CCID Research Institute, "中国电子信息产业发展研究院," https://baike.baidu.com/item/%E4%B8%AD%E5%9B%BD%E7%94%B5%E5%AD%90%E4%BF%A1%E6%81%AF%E4%BA%A7%E4%B8%9A%E5%8F%91%E5%B1%95%E7%A0%94%E7%A9%B6%E9%99%A2/8101155.

22 Graham Webster et al., "Forum: Analyzing an Expert Proposal for China's Artificial Intelligence Law," DIGICHINA, August 23, 2023, https://digichina.stanford.edu/work/forum-analyzing-an-expert-proposal-for-chinas-artificial-intelligence-law/.

23 "IDAIS_Beijing, 2024," IDAIS, March 10, 2024, https://idais.ai/dialogue/idais-beijing/.

24 Zhou Zhihua, "关于强人工智能," China Computer Federation, https://www.ccf.org.cn/upload/resources/file/2018/01/16/51070.pdf.

25 Matt Sheehan, "China's Views on AI Safety Are Changing—Quickly," Carnegie Endowment for International Peace, August 27, 2024, https://carnegieendowment.org/research/2024/08/china-artificial-intelligence-ai-safety-regulation?lang=en.

26 "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," China Law Translate, January 4, 2022, https://www.chinalawtranslate.com/en/algorithms/.

27 "Provisions on the Administration of Deep Synthesis Internet Information Services," China Law Translate, December 11, 2022, https://www.chinalawtranslate.com/en/deep-synthesis/.

28 Matt Sheehan, "Tracing the Roots of China's AI Regulations," Carnegie Endowment for International Peace, February 27, 2024, https://carnegieendowment.org/research/2024/02/tracing-the-roots-of-chinas-ai-regulations?lang=en.

29 "Interim Measures for the Management of Generative Artificial Intelligence Services," China Law Translate, July, 13, 2023, https://www.chinalawtranslate.com/en/generative-ai-interim/.

30 Sarah Zheng and Jane Zhang, "Beijing Tries to Regulate China's AI Sector Without Crushing It," *Bloomberg,* August 14, 2023, https://www.bloomberg.com/news/articles/2023-08-14/china-tries-to-regulate-ai-with-state-control-support-for-tech-companies.

31 Interim Measures for the Management of Generative Artificial Intelligence Services," China Law Translate, July, 13, 2023, https://www.chinalawtranslate.com/en/generative-ai-interim/.

32 "The State of AI Safety in China Spring 2024 Report," Concordia AI, May 14, 2024, https://concordia-ai.com/wp-content/uploads/2024/05/State-of-AI-Safety-in-China-Spring-2024-Report-public.pdf.

33 Jiaming Ji, et al., "AI Alignment: A Comprehensive Survey," preprint, arXiv, October 20, 2023, last updated April 4, 2025, https://arxiv.org/abs/2310.19852.

34 Lijun Li, et al., "SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models," preprint, arXiv, February 7, 2024, last updated June 7, 2024, https://arxiv.org/abs/2402.05044.

35 "Statement on AI Risk," Center for AI Safety, https://www.safe.ai/work/statement-on-ai-risk.

36 "IDAIS-Oxford, 2023," IDAIS, October 31, 2023 https://idais.ai/dialogue/idais-oxford/.

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

38 Vincent Manancourt, Tom Bristow, and Laurie Clarke, "China Expected at UK AI Summit Despite Pushback from Allies," *Politico,* August 25, 2023, https://www.politico.eu/article/china-likely-at-uk-ai-summit-despite-pushback-from-allies/.

39 "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," UK Government, February 13, 2025, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

40 "Introducing the AI Safety Institute," UK Government, November 2023, https://assets.publishing.service.gov.uk/media/65438d159e05fd0014be7bd9/introducing-ai-safety-institute-web-accessible.pdf.

41 Department for Science, Innovation and Technology and AI Safety Institute, "AI Safety Institute: Approach to Evaluations," UK Government, February 9, 2024, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.

42 "International AI Safety Report 2025," UK Government, January 29, 2025, https://www.gov.uk/government/publications/international-ai-safety-report-2025.

43 Concordia AI, "About," Accessed June 6, 2025, https://concordia-ai.com/#about.

44 "Global AI Governance Initiative," Embassy of the People's Republic of China in Grenada, October 24, 2023, http://gd.china-embassy.gov.cn/eng/zxhd_1/202310/t20231024_11167412.htm.

45 "COPU会议纪要," China Open Source Software Promotion Union, July 23, 2024, https://cosspu.org.cn/osinfo/shownews.php?id=103.

46 "Explainer: What is China's 'Third Plenum'?," *Reuters*, July 15, 2025, https://www.reuters.com/world/china/what-is-chinas-third-plenum-2024-07-15/.

47 Matt Sheehan, "China's Views on AI Safety Are Changing—Quickly," Carnegie Endowment for International Peace, August 27, 2024, https://carnegieendowment.org/research/2024/08/china-artificial-intelligence-ai-safety-regulation?lang=en.

48 "What Does the Chinese Leadership Mean by "Instituting Oversight Systems to Ensure the Safety of AI?," *AI Safety in China* (blog), Concordia AI, August 2, 2024, https://aisafetychina.substack.com/p/what-does-the-chinese-leadership.

49 Zhu Bulou, "提高人工智能安全治理水平（有的放矢）," *People's Daily*, http://paper.people.com.cn/rmrb/html/2024-08/05/nw.D110000renmrb_20240805_4-09.htm.

50 "High-level Meeting on International Cooperation on Capacity-building of Artificial Intelligence," United Nations, September 25, 2024, https://webtv.un.org/en/asset/k18/k188rsfh7d.

51 Yi Zeng (@yi_zeng), "Introducing the Chinese AI Safety Network," X (formerly Twitter), June 16, 2025, https://archive.ph/xYcMA.

52 "People," Beijing AI Safety and Governance Institute, accessed June 10, 2025, https://beijing.ai-safety-and-governance.institute/people.

53 Renana Araujo, Kristina Fort, and Oliver Guest, "Understanding the First Wave of AI Safety Institutes," Institute for AI Policy and Strategy, October, 2024, https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6705a4a7fd9bf94a706d10ce/1728423090604/IAPS+-+Understanding+the+First+Wave+of+AI+Safety+Institutes.pdf.

54 "Le Gouvernement Annonce la Création de L'Institut National pour P'évaluation et la Sécurité de L'intelligence Artificielle (INESIA)," Le Ministère de L'Économie et des Finance, January 31, 2025, https://www.entreprises.gouv.fr/espace-presse/le-gouvernement-annonce-la-creation-de-linstitut-national-pour-levaluation-et-la.

55 "With Robust and High End Common Computing Facility in Place, India All Set to Launch Its Own Safe & Secure Indigenous AI Model at Affordable Cost Soon," Ministry of Electronics & IT, January 20, 2025, https://pib.gov.in/PressReleasePage.aspx?PRID=2097709.

56 "Introducing the AI Safety Institute," Department for Science, Innovation & Technology, November 2023, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

57 Fu Ying: The Best Path to the Safety and Responsible Application of Future AI Requires Both Chinese and American Efforts," The Paper, February 12, 2025, https://m.thepaper.cn/newsDetail_forward_30139070.

58 Fu Ying, "Cooperation for AI safety must transcend geopolitical interference," *South China Morning Post*, February 12, 2025, https://www.scmp.com/opinion/china-opinion/article/3298281/cooperation-ai-safety-must-transcend-geopolitical-interference.

59 Matt Sheehan, "China's AI Regulations and How They Get Made," Carnegie Endowment for International Peace, July 10, 2023, https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

60 Christina Criddle and Anna Gross, "UK's Ambitions to Police AI Face Trump's 'Starly' Different Approach," *Financial Times*, December 13, 2024, https://www.ft.com/content/c9f6067c-3faa-4e6a-bc0f-2ed7290a8476.

61 Fu Ying, "Cooperation for AI Safety Must Transcend Geopolitical Interference," *South China Morning Post*, February 12, 2025, https://www.scmp.com/opinion/china-opinion/article/3298281/cooperation-ai-safety-must-transcend-geopolitical-interference.

62 Interim Measures for the Management of Generative Artificial Intelligence Services," China Law Translate, July 13, 2023, https://www.chinalawtranslate.com/en/generative-ai-interim/.

63 Matt O'Brien, "Trump Rescinds Biden's Executive Order on AI Safety in Attempt to Diverge From His Predecessor, *AP News,* January 22, 2025, https://apnews.com/article/trump-ai-repeal-biden-executive-order-artificial-intelligence-18cb6e4ffd1ca87151d48c3a0e1ad7c1.

64 J.D. Vance, "Remarks by the Vice President at the Artificial Intelligence Action Summit in Paris, France," American Presidency Project, February 11, 2025, https://www.presidency.ucsb.edu/documents/remarks-the-vice-president-the-artificial-intelligence-action-summit-paris-france.

65 "FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco," U.S. Department of Commerce, November 20, 2024, https://www.commerce.gov/news/fact-sheets/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international.

66 Jeffrey Dastin, "U.S. AI Safety Institute Director Leaves Role," *Reuters*, February 6, 2025, https://www.reuters.com/technology/us-ai-safety-institute-director-leaves-role-2025-02-06/.

67 Sammy Westfall, "After 100 days, the Toll of Trump's foreign Aid Cuts Has Begun to Sink In," *Washington Post*, May 1, 2025, https://www.washingtonpost.com/world/2025/05/01/trump-aid-cuts-hunger-hiv-100-days/.

68 "Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation," U.S. Department of Commerce, June 3, 2025, https://www.commerce.gov/news/press-releases/2025/06/statement-us-secretary-commerce-howard-lutnick-transforming-us-ai.

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

70 "Hawley Introduces Legislation to Decouple American AI Development from Communist China." Josh Hawley U.S. Senator for Missouri, January 29, 2025, https://www.hawley.senate.gov/hawley-introduces-legislation-to-decouple-american-ai-development-from-communist-china/.

71 "Additions to the Entity List," Federal Register, March 28, 2025, https://www.federalregister.gov/documents/2025/03/28/2025-05427/additions-to-the-entity-list.

72 Lily Ottinger, "Where's China's AI Safety Institute?," ChinaTalk (blog), November 20, 2024, https://www.chinatalk.media/p/wheres-chinas-ai-safety-institute.

73 "Singapore AI Safety Red Teaming Challenge: Evaluation Report," Infocomm Media Development Authority of Singapore (IMDA) and Humane Intelligence, February 2025, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/singapore-ai-safety-red-teaming-challenge-evaluation-report.pdf.

74 Matt Sheehan, "Mapping Chinese AI Regulation with Matt Sheehan," *AI Policy Podcast*, posted April 2, 2025, by Center for Strategic and International Studies, YouTube, https://www.youtube.com/watch?v=-pf5XlFYhrQ.

75 Xudong Pan, et al., "Frontier Ai Systems Have Surpassed the Self-Replicating Red Line," GitHub, December 10, 2024, https://github.com/WhitzardIndex/self-replication-research/blob/main/AI-self-replication-fudan.pdf; Haochen Zhao et al., "ChemSafetyBench: Benchmarking LLM Safety on Chemistry Domain," preprint, arXiv, November 23, 2024, https://arxiv.org/abs/2411.16736.

76 Zhao Ziwen, "China and US Need to Cooperate on AI or Risk 'Opening Pandora's Box', Ambassador Warns," *South China Morning Post,* March 2, 2025, https://www.scmp.com/news/china/diplomacy/article/3300738/china-and-us-need-cooperate-ai-or-risk-opening-pandoras-box-ambassador-warns.

77 Bill Bishop, "April Politburo Study Session on AI is Bad News for Nvidia," *Sinocism* (blog), April 26, 2025, https://sinocism.com/p/april-politburo-study-session-on?r=6s32&utm_medium=ios&triedRedirect=true.

78 Kristy Loke, et al., "Forum: Xi's Message to the Politburo on AI," *DigiChina*, April 30, 2025, https://digichina.stanford.edu/work/forum-xis-message-to-the-politburo-on-ai/.

79 "Frontier AI Safety Commitments, AI Seoul Summit 2024," UK Government, February 7, 2025, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

80 Scott Singer, "DeepSeek and Other Chinese Firms Converge with Western Companies on AI Promises," Carnegie Endowment for International Peace, January 28, 2025, https://carnegieendowment.org/research/2025/01/deepseek-and-other-chinese-firms-converge-with-western-companies-on-ai-promises?lang=en.

81 Ngor Luong and Zachary Arnold, "China's Artificial Intelligence Industry Alliance," Center for Strategic and International Studies, May 2021, https://cset.georgetown.edu/publication/chinas-artificial-intelligence-industry-alliance/.

82 "守护AI安全，共建行业自律典范——首批17家企业签署《人工智能安全承诺》," China Academy of Information and Communications Technology, https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng.

83 Matt Sheehan and Scott Singer, "What DeepSeek Revealed About the Future of U.S.-China Competition," *Foreign Policy*, February 3, 2025, https://foreignpolicy.com/2025/02/03/deepseek-china-ai-artificial-intelligence-united-states-tech-competition/.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Technology and International Affairs Program

The Technology and International Affairs Program develops insights to address the governance challenges and large-scale risks of new technologies. Our experts identify actionable best practices and incentives for industry and government leaders on artificial intelligence, cyber threats, cloud security, countering influence operations, reducing the risk of biotechnologies, and ensuring global digital inclusion.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org