CONCORDIA AI
安远AI

State of AI Safety in China

July 2025

Authors

This report was authored by Gabriel Wagner, Jason Zhou, Kwan Yee Ng, and Brian Tse.

Acknowledgements

We are grateful to Alan Chan, Xin CHEN, Jeff Ding, Oliver Guest, Emmie Hine, Lucy Luo, Zilan QIAN, Saad Siddiqui, Jan Pieter Snoeij, and Jie ZHANG for providing feedback and suggestions on this report.

How to cite this report

Gabriel Wagner, Jason Zhou, Kwan Yee Ng, and Brian Tse, "State of AI Safety in China (2025)," Concordia AI, July 2025, https://concordia-ai.com/wp-content/uploads/2025/07/State-of-AI-Safety-in-China-2025.pdf.

Table of Contents

Executive Summary

As AI capabilities and governance challenges continue to grow—and Chinese models close their gap with the leading edge—understanding China's role in AI safety and governance is more critical than ever. Since 2023, Concordia AI's annual *State of AI Safety in China* reports have analyzed how China addresses general-purpose AI risks, particularly dangerous misuse, accidents, and loss of control—challenges with an outsized need for international cooperation.

This year's report provides updates from May 2024 to June 2025 across five domains: domestic governance, international governance, technical safety research, expert views on AI safety and governance, and industry governance.

Domestic Governance

- **Domestic rhetoric and policy include increasingly prominent and specific calls for AI risk mitigation, while continuing to emphasize the complementarity between AI safety and development.** At the Third Plenum, one of China's most important political events in 2024, AI safety was formally elevated to a national priority. In April 2025, the country's seniormost officials in the Communist Party of China (CPC) Politburo held a study session dedicated to AI, calling for monitoring, early warning, and emergency response systems to ensure AI safety, reliability, and controllability. This marked a significant endorsement of AI risk mitigation. AI safety has also gained prominence in official meetings and emergency planning documents related to national security and public safety, suggesting growing recognition of AI's potential for large-scale risks.
- **China is implementing its AI regulations through an expanding AI standards system.** While a comprehensive national AI Law remains unlikely in the near future, China has issued detailed regulations on AI-generated content labeling and watermarking. It also continues to implement existing regulations, which require pre-deployment registration and safety testing of certain types of AI models. These mandates are operationalized through a rapidly expanding system of standards: from January to May 2025 alone, China issued as many national AI standards as in the preceding three years combined. Although current regulations and standards primarily address near-term risks, official standard-setting plans reveal intentions to address severe risks such as AI's impact on cybersecurity or loss of control.

International Governance

- **China has emphasized AI safety and global AI capacity-building as key themes in its international AI diplomacy.** At Davos in January 2025, China's top science and technology official warned that developing AI without safety measures is like driving on a highway without brakes, and cautioned against unchecked international competition. At the Paris AI Action Summit, leading Chinese academic and policy institutions launched the "China AI Safety & Development Association" (CnAISDA), positioning itself as China's counterpart to other national AI safety institutes. Meanwhile, the Chinese government has become more proactive on AI capacity-building in the Global South by sponsoring a UN resolution (adopted by over 140 countries), launching an "Action Plan," and co-founding a "Group of Friends" on AI capacity-building.
- **China has launched bilateral AI dialogues with several key countries, though the outlook for these engagements remains mixed.** While Chinese President Xi Jinping and then-United States (US) President Joe Biden agreed on the importance of human control over nuclear systems in November, their intergovernmental AI dialogue has not met since May 2024, and its future faces major uncertainty under the Trump administration. Meanwhile, China and the United Kingdom (UK) inaugurated a new bilateral AI dialogue in May 2025, suggesting China's continued interest in engaging willing partners.

Technical Safety Research

- **Frontier AI safety research output in China has expanded rapidly.** From June 2024 to May 2025, Chinese scholars published ~26 papers/month, more than double the previous year's output. Previously underexplored areas, such as alignment of superhuman systems and mechanistic interpretability, have become popular topics. Research on deception, unlearning, and CBRN (chemical, biological, radiological and nuclear) misuse has also expanded, though the latter remains mostly benchmark-driven. Chinese researchers are taking concrete actions to address severe AI safety risks.

Expert Views on AI Safety and Governance

- **Expert discourse in China is placing greater emphasis on AI safety and governance.** From 2023 to 2024, coverage of AI safety and governance at two major AI conferences—the World AI Conference (WAIC) and World Internet Conference (WIC)—nearly doubled. In addition, WAIC was upgraded in 2024 to a "High-Level Meeting on Global AI Governance." For the first time in the conference's history, a top central government official, Premier LI Qiang (李强), delivered the opening remarks. Notably, he highlighted AI safety and governance issues. Meanwhile, expert speeches at the opening ceremony also emphasized these topics more than in previous years.
- **Experts are increasingly publishing in-depth analyses of AI risks in biosecurity, cybersecurity, and open source AI.** Multiple state-affiliated and independent experts have written detailed

analyses on AI misuse in biosecurity and cybersecurity. Meanwhile, Chinese experts highlight the safety benefits of open source AI, while also increasingly exploring potential downstream misuse risks.

Industry Governance

- **Most leading Chinese foundation model developers have signed voluntary "AI Safety Commitments."** The commitments, drafted by the AI Industry Alliance of China (AIIA) in December 2024, pledge safety measures across the AI development lifecycle, including dedicated safety teams, red teaming, data security, infrastructure protection, transparency, and investment in frontier safety research.
- **Chinese AI developers typically implement well-known safety methods, but provide limited transparency on safety evaluations.** Our survey of frontier model releases shows that nine out of thirteen developers have published technical model release cards, many reporting standard safety techniques like data filtering, reinforcement learning from human feedback (RLHF), constitutional AI, and red teaming. However, transparency remains limited: only three of the surveyed companies have published safety evaluation results.

Introduction and Scope

How has China's approach to AI safety and governance evolved since 2023, and what implications does this hold for international AI cooperation? This question has gained urgency as AI capabilities advance rapidly while international governance mechanisms remain fragmented.

In 2023, global knowledge of China's AI safety and governance practices was largely lacking, even after China released regulations on generative AI and announced the Global AI Governance Initiative. Initial international cooperation showed promise—China joined the US, UK, and over 20 other countries in announcing their intention to advance AI safety through the Bletchley Declaration at the UK Global AI Safety Summit in October 2023.[1] In 2024, the United Nations (UN) Global Digital Compact called for the creation of an Independent International Scientific Panel on AI and a Global Dialogue on AI Governance.[2] Additionally, China and the US began an intergovernmental dialogue on AI that achieved agreement on "the need to maintain human control over the decision to use nuclear weapons."[3] However, developments since then have shown international mobilization and cooperation to be tenuous and challenging. It is unclear if the China-US intergovernmental AI dialogue will continue in the new US administration, and important AI actors such as the US and UK did not sign the Paris AI Action Summit Declaration.[4]

Meanwhile, AI capabilities have not stood still. New models released by Chinese startup DeepSeek—previously obscure to many in the West—in December 2024 and January 2025 "shook the tech world" and were described as a "Sputnik moment" in Western media. Globally, frontier models are advancing through "reasoning," and AI systems have also become more embedded in businesses and everyday life.[5]

As development has continued, China's attention to AI safety and governance has also increased. A key CPC meeting called for "instituting oversight systems to ensure the safety of artificial intelligence" in July 2024. On the international stage at the 2025 World Economic Forum (WEF) in Davos, China's top science and technology official voiced alarm about "disorderly competition" over AI and stated that China "has put in place robust regulatory systems, institutions and measures, and is confident of ensuring sound management and utilization of AI technologies."[6] Then in April, the CPC Politburo—comprising China's top 24 officials—conducted its first dedicated study session on AI since 2018, with President Xi Jinping stating that AI brings unprecedented opportunities, but also "unprecedented risks and challenges."[7] To improve safety, he called for China to "establish systems for technical monitoring, early risk warning and emergency response." These developments represent just a portion of China's expanding AI governance activities.

To address knowledge gaps, Concordia AI published our first *State of AI Safety in China* report in October 2023, seeking to inform debates at the UK Global AI Safety Summit.[8] Concordia AI published an updated version of the report in May 2024, documenting advancements in Chinese technical AI safety research and progress in discussions on severe AI risks among industry and academics.[9] This edition is our third comprehensive report, drawing on Chinese-language primary sources including government policy documents, expert assessments, industry reports, and proceedings from major AI forums. We limit our analysis to publicly available information and note that the rapidly evolving nature of AI safety and governance means that this report may miss some developments that occurred after our research cutoff in late June 2025.

This report is divided into five sections that provide a comprehensive overview of the AI safety landscape in China. In the "Domestic Governance" section, we analyze how conversations within and adjacent to the government have increasingly raised the possibility of AI posing severe threats. The "International Governance" section documents China's external messaging on AI governance and safety, as well as China's implementation of global AI capacity-building efforts. In the "Technical Safety Research" section, we discuss the increase in volume of technical AI safety research in China from May 2024 to June 2025, as well as how a growing number of research groups in China are choosing to devote substantial focus to this topic. In the "Expert Views on AI Safety and Governance" section, we show how leading Chinese AI conferences have increased their attention to AI safety and governance over the past year, and we delve into discussions around AI's intersection with biosecurity, AI's impact on cybersecurity, and open source governance. Lastly, in the "Industry Governance" section, we analyze voluntary governance efforts within Chinese industry, in particular AI safety commitments by the AI Industry Alliance (AIIA) of China and safety disclosures by individual AI developers.

Scope of the Report

This report covers risks from advanced, general-purpose AI systems—defined by the International AI Safety Report 2025 (authored and advised by 96 AI experts from around the world and chaired by Turing Award Winner Yoshua Bengio) as AIs that can "perform a wide variety of tasks," including GPT-4o, AlphaFold-3, and Gemini 1.5 Pro.[10] The International AI Safety Report groups general-purpose AI risks into three primary categories: malicious use (e.g. fake content, cyber offence, and biological and chemical attacks); malfunctions (e.g. reliability issues, bias, and loss of control); and systemic risks (e.g. labor market risks, global AI divides, and environmental risks).

Among these AI risks, we focus primarily on those that are both potentially severe and manifest across borders. These risks are the most likely to require international cooperation due to their cross-border negative externalities, of a magnitude that might cause even rivalrous actors to contemplate coordination.[11] Therefore, this report most deeply examines malicious uses of AI with extreme consequences and loss of control of advanced AI systems—what we term "frontier AI safety" or "severe AI risks."[a]

We provide limited analysis of other issues documented by the International AI Safety Report, including fake content, bias, discrimination, and privacy, only where it is important to the reader's understanding of institu-

a. This also includes risks from narrow AI systems with dangerous capabilities, particularly AI models used for bioengineering.[12]

tional mechanisms, diplomatic positions, or other issues that overlap with severe AI risks. The report does not cover military AI systems.

While our report focuses on general-purpose AI systems, the Chinese leadership's interest in AI is broader, often emphasizing AI applications in specific industries rather than frontier model progress. Senior policymakers consistently underscore AI's role in driving economic growth under the "new productive forces" framework.[13] This leads to a focus on AI's integration into manufacturing and deployment in specific sectors, as exemplified by the "AI+ initiative."[b] Additionally, "embodied AI" (for example humanoid robots or autonomous vehicles), has also emerged as a significant policy focal point. Observers should recognize that Chinese decision-making on AI reflects complex interests which do not neatly fit into a "development versus safety" dichotomy.

Conflicts of Interest

Concordia AI actively participates in and advises on AI safety within China through various channels, including by hosting forums, advising AI developers, and providing policy recommendations. Our ongoing work in this field places us in a unique position to understand and analyze information regarding AI safety in China. However, this engagement also entwines us with the evolving Chinese AI safety landscape, potentially creating a conflict of interest. Our organizational mission and vested interest in advancing AI safety in China might influence our perspective. Nevertheless, we believe that our findings are the result of objective analysis, and we disclose this potential conflict to readers for full transparency.

Concordia AI is a social enterprise with a mission to ensure that AI is developed and deployed in a way that is safe and aligned with global interests. In the course of our operations, we have received consulting fees from various companies located in mainland China, Hong Kong, and Singapore, including some discussed in this report. However, no financial engagements with these companies influenced or were related to this report's creation. Additionally, no governments and government-affiliated organizations provided input on the content in this report. Concordia AI is an independent institution, not affiliated to or funded by any government or political group.

b. The AI+ Initiative was first introduced in the 2024 Government Work Report and reaffirmed in the 2025 Government Work Report.

Domestic Governance

Key takeaways

- China's high-level policy signaling continues to emphasize a complementary relationship between AI development and safety, rejecting the notion that the two are in conflict.
- Official rhetoric on AI safety has become more prominent and specific, most notably at the Third Plenum (one of China's top political events in 2024) and an April 2025 Politburo study session (the first dedicated to AI since 2018), which stressed risk monitoring, early warning, and emergency response.
- While a comprehensive national "AI Law" remains unlikely in the near term, debates on AI-related legislation are ongoing.
- China issued a new regulation on labeling and watermarking AI-generated content. Other existing regulations continue to be implemented through mandatory algorithm registrations and enforcement campaigns.
- China has issued a large number of AI standards that elaborate on the implementation of existing regulations. Although current standards do not focus on frontier risks, authoritative planning documents suggest that attention to these issues may grow in the next 1–3 years.

1.1 Overarching national guidance

China's top-level policy continues to emphasize the complementary relationship between AI development and safety. As Executive Vice Premier DING Xuexiang (丁薛祥), China's sixth-ranked official, succinctly remarked in January 2025: "It's like driving on a highway—without control over the brakes, you can't confidently accelerate."[14] This core principle remains consistent since our last update in May 2024.

Recent developments suggest that the leadership's focus on AI safety is intensifying, with multiple senior officials articulating safety concerns in high-profile settings and calling AI safety challenges "unprecedented." These statements not only underscore greater political attention on AI safety, but also show a shift toward more substantive engagement: whereas in earlier statements, officials often mentioned AI safety only briefly,

recent remarks are more specific, detailing distinct stages of AI governance such as risk monitoring, early warning, and emergency response. AI safety and security have featured in:

- The Third Plenum, one of China's most important political meetings in 2024;
- A Politburo study session dedicated to AI in April 2025;
- Multiple high-level meetings and plans about national security.

1.1.1 Third Plenum resolution highlights AI safety

The Third Plenum of the Communist Party of China (CPC) 20th Central Committee, one of the most significant political events in China in 2024, explicitly highlighted the necessity of "instituting oversight systems to ensure the safety of AI."[15] The Third Plenum, which takes place every five years, typically defines strategic priorities for the subsequent years, making its resolutions among the most authoritative reflections of leadership's long-term intentions.[16]

The resolution indicates a clear elevation of AI as a public and national security priority, albeit without specifying implementation details. Such resolutions characteristically cover a wide range of topics and only go into limited detail on any given aspect. The explicit mention of AI safety oversight—classified under "public security governance mechanisms" in a section on national security alongside cybersecurity, biosecurity, and natural disasters—shows that AI is being categorized as a public and national security priority, not only as a content control issue. Official study materials released after the plenum elaborated on the resolution, explaining that oversight systems for AI safety are needed to manage rapid technological advancements, support high-quality development, and enhance China's role in global governance.[17] The document calls for oversight to involve "forward-looking prevention and constraint-based guidance," signaling an active, precautionary regulatory stance. However, it does not mention which specific AI risks the state is concerned about.

1.1.2 First Politburo study session on AI since 2018 addresses AI risks

In April 2025, the CPC Politburo—comprising China's 24 top officials—held a study session on AI.[18] The Politburo holds such "study sessions" roughly once a month, on various topics ranging from blockchain and military governance to cultural development, employment, and clean energy.[a] They serve the dual purpose of allowing top leaders to receive expert briefings on a specific topic and signaling political priorities to party officials and the general public. The last Politburo study session on AI was in 2018.[20]

During the 2025 session, President Xi talked about AI development and application, safety, and international cooperation. Overall, Xi's speech was remarkably consistent with messaging over previous years, reiterating existing priorities under a stronger spotlight.[21]

On development, Xi urged fundamental breakthroughs in critical hardware and software to boost self-reliance, accelerate industry applications, and strengthen policy support.

a. For more details on the importance and structure of Politburo study sessions see Neil Thomas, 2024.[19]

On safety, Xi described AI as bringing "unprecedented development opportunities" but also posing "unprecedented risks and challenges." This is significantly stronger than Xi's previous statements on AI risks. The language of "unprecedented risks" was included in the international-facing Shanghai Declaration on Global AI Governance in July 2024, and it is notable that it is now being used at the highest levels of government.[22] Equally notable was the detailed treatment of AI safety within Xi's speech—an entire paragraph rather than the customary brief mention. Xi specifically called for the establishment of systems for "technology monitoring, early risk warning, and emergency response," and urged rapid development of "laws and regulations, policies and systems, application norms, and ethical guidelines."

However, Xi notably refrained from identifying specific AI technologies, such as large language models (LLMS), or detailing concrete risk scenarios. His subsequent visit to a Shanghai startup accelerator focused heavily on LLMs, but the Politburo session itself featured as guest lecturer Xi'an Jiaotong University Professor ZHENG Nanning (郑南宁)—an expert in computer vision, pattern recognition, and advanced computing architectures, who has worked on autonomous driving (among other things) rather than LLMs.[23] This choice could reflect a broader strategic interest in AI that extends beyond general-purpose frontier models alone.

For President Xi's remarks on international AI governance, see the "International Governance" section.

1.1.3 AI safety increasingly present in national security planning

In our May 2024 update, we observed that AI is becoming more prominent within China's national security discourse. A series of articles on the Ministry of State Security (MSS) official Wechat account, published throughout 2023 and early 2024, highlighted AI's impact on public opinion warfare, economic security, data security, and cybersecurity.[24]

This trend has accelerated markedly over the past year: discussions of AI have moved beyond government-affiliated social media accounts and are now being included in official national security planning sessions and documents:

- In February 2025, the CPC Politburo convened a study session on "constructing a safer China," where President Xi Jinping listed AI safety as a part of China's comprehensive national security framework.[25] The session classified AI risks under "public security," alongside disaster prevention, production safety, food and drug safety, and cybersecurity, advocating for "preemptive prevention" governance models for these domains.
- The revised National Emergency Response Plan, released in February 2025 to replace the 2005 version, explicitly lists AI risks among critical threats that require ongoing monitoring, alongside epidemics, cybersecurity incidents, and financial anomalies.[26]
- In April 2025, Minister of State Security CHEN Yixin (陈一新) published a lengthy essay in *Qiushi*, the CPC's official political theory journal. He described AI as experiencing "explosive" growth and elevating technology-driven safety risks, and he argued that robust mechanisms for risk assessment and preemptive governance were needed.[27]

- In May 2025, China's first ever governmental white paper on national security, published by the State Council, made multiple references to AI, calling it a "double-edged sword," alongside quantum and biotechnology.[28] The paper calls for building AI safety oversight and evaluation systems that feature agile governance, tiered management, and rapid response.

Collectively, these signals from China's senior political leaders reflect heightened awareness of AI's implications for national security. These official statements suggest a mix of concerns that include the risk of China falling behind in strategic competition, as well as risks posed by the technology itself. However, they do not explicitly mention risks from misuse of advanced AI systems or from loss of control of superintelligent AI.

The broader policy expert discourse can provide hints about where senior policymakers' thinking may be headed. Expert analysis on the intersection of AI safety and national security is becoming more sophisticated, with some experts explicitly recognizing extreme—even existential—risks. While some Chinese national security experts had already briefly mentioned existential risks from AI as early as 2023, more recent articles provide a much more detailed discussion of such risks.[29]

For example, in December 2024, the director of the Holistic View of National Security Research Center at the China Institutes of Contemporary International Relations (CICIR) published a detailed analysis framing AI as the most dynamic, rapidly evolving, and strategically critical frontier in national security.[30] CICIR is one of China's oldest and most respected foreign policy think tanks, with over 300 employees.[31] The essay outlined various risks associated with AI, such as susceptibility to adversarial attacks, lack of transparency, proliferation of deepfakes, weaponization for cyber operations, societal disruption, and potential technological lag relative to competitors.

Other notable examples come from LU Chuanying (鲁传颖), an international relations scholar at Tongji University who participated in Track 2 dialogues with US think tanks on cybersecurity and AI while he worked at the local government-backed Shanghai Institute for International Studies (SIIS).[32] In early 2025, in multiple articles published in official media, he discussed cutting-edge threat scenarios, including automated AI-driven cyberattacks and the dangers of integrating AI into sensitive decision-making processes like nuclear command and control.[33] Notably, he presents a three-level AI risk grading:

- Existential risks: Misaligned AI that leads to human extinction or irreversible collapse of essential living conditions.
- Catastrophic risks: An AI system that causes widespread societal collapse in the short term, for example a financial AI with flawed training data that triggers a rapid negative feedback loop, sparks a financial crisis, erodes public trust, and destabilizes government authority.
- General risks: These accumulate over time through localized hazards such as system failures, algorithmic bias, or data leaks.

Lu explicitly argues for integrating even existential risks into China's broader national security planning framework.[34]

1.2 Legislation and regulation

Since our last report in May 2024, there have been few substantial updates to Chinese law and regulation that directly address frontier AI safety:

- **National Law:** Passed by China's legislature, the National People's Congress (NPC), these represent the highest level of law in China. There is currently no comprehensive national AI Law. Although domestic experts continue to debate legislative proposals and policymakers show general interest in the topic, no strong political signal suggests that such a law will be passed imminently.
- **Departmental Regulations:** China already maintains various departmental regulations governing aspects of AI. These fall into two categories: the Cyberspace Administration of China (CAC) has led on regulations for the security of recommendation algorithms and generative AI, while the Ministry of Science and Technology (MOST) has released science and technology ethics rules which apply to AI. Since our previous update, however, the only new regulation relevant to frontier AI safety is one on AI-generated content labeling and watermarking.

1.2.1 National AI Law not prioritized—but also not completely off the table

Chinese lawmakers continue to express interest in AI, but fast passage of a dedicated, overarching "AI Law" is unlikely. While the 2023 and 2024 State Council Legislative Plans had mentioned working on a "Draft AI Law," the 2025 version only contains a more generic reference to "promoting legislation for the healthy development of artificial intelligence."[35] Similarly, both the 2024 and 2025 NPC legislative plans only included a generic call for "legislative projects related to the healthy development of AI" in a section of "preliminary review projects."[36] An "AI Law" was not listed as a specific agenda item.

However, this does not mean that an AI Law is entirely off the table. ZHANG Linghan (张凌寒)—a professor at China University of Political Science and Law (CUPL) and lead author of an expert-proposed AI Law draft—argues that the approach is evolving, not stalling.[37] She claims that legislative plans are not rigid blueprints, and suggests that the more generic language in the plans signals a turn toward a more integrated, system-wide strategy rather than a rush to enact a standalone AI Law, as AI requires adaptive updates across the whole system, not just a single new law.

The NPC has also continued to study AI-related issues. In May 2024, the NPC Standing Committee received a special lecture on AI by SUN Ninghui (孙凝晖), an academician from the Chinese Academy of Engineering, in which he provided an overview of AI development and urged them to speed up AI-related regulation.[38] In April 2025, officials from the NPC committee with jurisdiction over AI conducted an inspection visit to Zhejiang province to study AI-related legislation, meeting provincial officials and AI companies including Alibaba and Unitree Robotics.[b] While such activities signal general interest in AI, they are not clear signs of strong prioritization.

b. The Education, Science, Culture and Public Health Committee is the NPC committee overseeing AI. The visit was led by its Chair LUO Shugang (雒树刚).[39]

Chinese policy experts, meanwhile, remain actively engaged in discussions about a comprehensive national AI Law. Two separate groups of experts have drafted entire "model law" proposals: the first comes from the Chinese Academy of Social Sciences (CASS), led by ZHOU Hui (周辉), released in August 2023; the second is led by CUPL professor Zhang Linghan. We have covered the two proposals extensively in our previous May 2024 report; the table below summarizes key similarities and differences.

Table 1.1: Comparing expert AI Law drafts

Key provisions		
Provision[c]	**CASS-led draft**	**CUPL-led draft**
Licensing requirement for models with certain high risk profiles	✔	✖
New government agency for AI	✔	✖
Tax credits for "safety governance" research or equipment	✔	✖
Specialized oversight for foundation models above a certain (unspecified) size	✔	✔
Provision on AGI value alignment	✖	✔
Financial penalties for violations by AI developers	✔	✔
Liability exemptions for open source AI	✔	✔
Whistleblower protections	✔	✖
AI ethics review requirements	✔	✔

In March 2025, the CASS expert group released version 3.0 of their proposed AI Law.[40] The updated draft retains key provisions from earlier versions, including licensing for certain AI models, creation of a dedicated AI regulatory agency, and liability exemptions for open source AI. It also introduces new mechanisms for whistle-blower protection. Specifically, the draft model law mandates robust internal reporting processes within AI companies, explicit penalties for retaliation against whistleblowers, and government-established channels for reporting major harms, security risks, or deliberate distortion of security assessment results. This feature notably parallels emerging proposals in international contexts, such as California's draft AI legislation SB53, suggesting a potential cross-pollination of governance ideas.[41] The alternative CUPL draft proposal led by Zhang Linghan has not been updated since our previous report.

c. For licensing requirement, see CASS draft Chapter 3; For new government agency for AI, see CASS draft Article 13 and CUPL draft Article 59; For tax credits for safety governance research, see CASS draft Article 26; For specialized oversight for foundation models, see CASS draft Article 54 and CUPL draft Articles 50-57; For provision on AGI alignment, see CUPL draft Article 77; For financial penalties, see CASS draft Articles 76-80 and CUPL draft Articles 82-83; For open source AI liability exemptions, see CASS draft Article 83 and CUPL draft Article 95; For whistleblower protections, see CASS draft Article 70; For ethics reviews, see CASS draft Article 49 and CUPL draft Article 58.

While these drafts are influential as expert recommendations, they remain unofficial, and their eventual impact remains unclear. There is still not an expert consensus that an overarching AI Law is urgently necessary. Several scholars from institutions like Shanghai Jiao Tong University, Renmin University, and Beijing Institute of Technology argue against comprehensive horizontal AI legislation at this stage, favoring targeted, vertical regulatory approaches aligned with China's current policy trajectory.[42] For the time being, it appears that Chinese lawmakers agree, as they have not indicated that they are prioritizing an overarching AI Law in the near term.

1.2.2 Limited regulatory updates for frontier AI

As explained in our previous *State of AI Safety in China* reports, China has a range of departmental regulations for AI. They can be divided into two categories:

- AI security regulations by the CAC;
- AI ethics regulations by MOST.

Apart from a more detailed regulation on AI-generated content labeling, there have been no major changes in either category since our last report in May 2024.

CAC AI security regulations

The CAC has drafted a range of regulations for AI over the past few years, but there have been only modest updates with relevance to frontier AI safety since our last update in May 2024. The most consequential regulations for general-purpose AI are the *Interim Measures for the Management of Generative AI Services*, effective from August 2023.[43] These measures establish a *de facto* licensing regime for generative AI models through pre-deployment security assessments.[d] By March 2025, CAC had registered 346 models under these guidelines.[45]

In March 2025, CAC released an additional regulation on labeling requirements for AI-generated content, which will come into effect in September 2025.[46] Building on provisions in the *Interim Measures*, the regulation mandates both explicit labels (e.g. visible indicators) and implicit labels (e.g. embedded metadata). It also requires internet platforms to develop detection systems capable of identifying AI-generated content and to apply labels even when users fail to do so. These rules represent a key step toward strengthening transparency in the dissemination of AI-generated content, which could help mitigate some systemic AI risks.

After the release of this regulation, enforcement efforts also intensified. In April 2025, the CAC launched a three-month nationwide campaign specifically targeting the misuse of generative AI.[47] While AI-generated content labeling was previously part of a broader online governance campaign in March 2024, this campaign is the first enforcement effort solely dedicated to generative AI, and it also extends beyond labeling to include other misuse issues.[48] In the campaign's first phase, the CAC reported taking down over 3,500 illegal generative AI products and penalizing some 3,700 user accounts.[49] Although it has not disclosed full details of the

d. For more background on the regulations see Matt Sheehan, 2024.[44]

violations, local authorities indicated that typical targets included companies that failed to conduct mandatory security assessments, neglected to label AI-generated content, or lacked safeguards against misuse. Offending products reportedly included tools that generated pornographic or violent images, money laundering guides, and deepfake impersonations. As part of the campaign, Jiangsu province conducted red teaming of registered models, and in Shanghai, some companies that reactivated banned features faced formal investigations and legal penalties.[50]

This wave of campaign-style enforcement highlights the state's willingness to implement existing regulations. However, the reactive nature of the campaign suggests that enforcement remains selective and uneven, and that clearer standards for penalties and long-term compliance mechanisms are still evolving. According to the CAC, the second phase—ongoing as of June 2025—may target users who misuse AI to generate rumors, pornographic and violent content, impersonation and deepfakes, nefarious online bot activity, and applications that negatively affect minors' mental health.[51]

MOST AI ethics regulations

Another departmental regulation relevant to frontier AI is an ethics review system for science and technology developed by MOST in 2023, which explicitly includes AI.[52] Starting in early 2024, organizations have been required to register ethics review committees through a centralized—but non-public—national platform.[53] Local governments in late 2024 and early 2025 issued supplementary guidance to assist entities in setting up and registering these committees, with some policies explicitly addressing AI ethics.[e] Government research funding notices, including some for AI, also increasingly stipulate compliance with ethics reviews as a mandatory requirement to receive funding.[57] Overall, however, implementation progress remains difficult to track. Some domestic scholars have raised concerns that the system may be formalistic, weakly enforced, and ill-suited for governing the unique ethical risks of AI.[58] ESG (Environmental, Social, and Governance) reports of large model developers suggest that they are complying with the legal obligation to set up AI ethics committees, which we describe in more detail in the "Industry Governance" section.

In summary, although existing regulations are enforced through campaigns, few substantial regulatory updates directly impacting frontier AI safety have occurred at the departmental regulation level since our last update in May 2024.

1.3 Standards

Standards are a crucial part of China's AI governance ecosystem as they provide detailed implementation guidelines for regulations. For instance, the *Interim Measures for the Management of Generative AI Services* contain relatively vague requirements, such as that generative AI service providers should take "effective

e. For example in Shenzhen[54] and Beijing.[55] The Beijing local policy on science and technology ethics from December 2024 suggests that the city government will set up a municipal laboratory on AI safety and ethics, and establish a "service station" to help companies with AI security and ethics testing. The city government appears to have already followed up on these promises in early 2025.[56]

measures" to "increase the accuracy and reliability of generated content." Standards provide more specific guidance on what such requirements mean.

Key updates since our May 2024 report include:

- AI standard-setting has sped up significantly. From January to May 2025 alone, China released as many national AI standards as in the three years from 2022 to 2024 combined.
- The Ministry of Industry and Information Technology (MIIT) has established a new AI standard-setting technical committee, with an AI safety working group.
- Existing AI security and safety standards mostly focus on privacy, data security, AI-generated content labeling, harmful and illegal content filtering, and general security management practices. They give little attention to frontier AI risks, such as loss of control or misuse in dual-use domains.
- Roadmaps and plans by standard-setting agencies suggest that frontier AI safety may receive more treatment in the next 1–3 years, but details remain unclear.

Standards follow a hierarchy of National Standards (国家标准), Industry Standards (行业标准), Group/Association Standards (团体标准), and Local Standards (地方标准).

1.3.1 National Standards

Formulated by the Standardization Administration of China (SAC), these are the most authoritative standards. There are two key standard-setting committees working on AI standards under SAC:

- **SAC/TC28/SC42** — *AI Subcommittee 42 of the Technical Committee 28 on Information Technology*: Established in 2020, SC42 is China's dedicated subcommittee for AI standardization, mirroring the international AI standardization body ISO/IEC JTC 1/SC 42.
- **SAC/TC260** — *Technical Committee 260 on Cybersecurity*: Although formally focused on cybersecurity, TC260 plays a key role in formulating AI security standards, particularly for generative AI. It has authored some of the most consequential standards to date on pre-deployment testing, data, and content governance.

While their mandates are not neatly separated, TC28/SC42 leads on technical and functional dimensions of AI (such as model performance, interoperability, and explainability), whereas TC260 leads on security (such as data protection, misuse prevention, and content control).

National standards can be further divided into mandatory (强制性) standards, which are legally binding, and voluntary (推荐性), which are not legally binding.[f] Mandatory standards are rare: according to the official SAC database, as of May 2025, China only has just over 2000 mandatory standards, but 44,000 voluntary

f. "Voluntary" standards are sometimes also translated as "recommended" standards. They can be recognized by their prefix code "T."

standards. In practice, even voluntary national standards can become quasi-mandatory when embedded into licensing requirements.

Overall, there are 24 AI standards in SAC's database.[g] More than half of these have been published since our May 2024 report, showing a strong acceleration in AI standard formulation. Almost all of them are voluntary standards. However, the first ever mandatory national standard on AI was released in early 2025.[60] It covers the labeling and watermarking of AI-generated content, and was issued in tandem with the regulation on the same issue.

Table 1.2: Number of Chinese national AI standards

Number of Chinese national AI standards							
	2001	**2006**	**2021**	**2022**	**2023**	**2024**	**2025 (Jan–May)**
Voluntary	1	3	1	3	2	4	9
Mandatory	0	0	0	0	0	0	1

Among the voluntary standards, the most impactful national standard to date for the safety of general-purpose AI models is the *Basic Security Requirements for Generative AI Services*, which was released in May 2025 and will come into effect in November 2025.[61] The standard provides detailed security requirements for generative AI service providers across the model lifecycle—through data filtering, fine-tuning, prompt monitoring, and output filtering.[62] It focuses on 31 "security risks" across five categories:

1. Violation of "core socialist values;"
2. Discrimination;
3. Commercial violations;
4. Violation of legitimate rights and interests of others;
5. Domain-specific risks (healthcare, critical infrastructure, etc).

The standard describes security tests that providers have to conduct across these categories. Even though this is only a voluntary national standard, the standard is almost *de facto* mandatory as a pre-approved method for companies to demonstrate compliance.

In our May 2024 report, we discussed a preparatory technical document for this standard, which referenced several frontier AI risks (deception, self-replication, cyberattacks, and biological and chemical weapon design).[63] However, these references have been removed from the official national standard, suggesting that standard-setters are focusing on existing threats that need governance now.

g. We used the key words AI (人工智能), large model (大模型), machine learning (机器学习), deep learning (深度学习), smart agent (智能体), data annotation (数据标注) in the database.[59] Note that this does thus not include more specific standards on, for instance autonomous driving or facial recognition.

Alongside the *Basic Security Requirements*, two other voluntary national standards were released that contain even more detail on training data filtering and fine-tuning data labeling.[64] They reference the same 31 security risks as the *Basic Security Requirements*, and add further detail. For instance, they include provisions on evaluating synthetic training data for hallucinations and distinguish between "functional annotations" (which enhance model capabilities) and "security annotations" (which aim to reduce risk, and require at least 200 annotations per risk).

1.3.2 Industry Standards

Industry standards fill gaps where no national standard exists. They can be influential within specific sectors, but are lower in authority, since they are typically only formulated when no applicable national standard exists.[65]

In December 2024, MIIT established its own AI standardization committee: MIIT/TC1.[66] Among its subcommittees is one dedicated to AI safety/security, which has already published detailed standard plans for governance capabilities, infrastructure security, cybersecurity, data security, algorithm and model security, application security, and AI for security.[67]

While this addition to China's AI safety standard-setting ecosystem could potentially broaden the scope and speed of regulatory development, MIIT/TC1's industry standards will carry less formal authority than national standards by the two relevant standard committees under the Standardization Administration of China (TC28/SC42 and TC260). Nevertheless, MIIT's standards can still shape implementation in key sectors and influence future national standards.

1.3.3 Other complementary standards

Beyond national and industry standards, there are a few other standard types with less authority, including:

- **Group/Association Standards:** Developed by industry bodies or consortia, such standards are faster to produce and are thus at the frontier of emerging practice. However, they carry less formal authority in regulatory enforcement. For example, the AI Industry Association (AIIA) has drafted standards on AI agent security, covering aspects like safety in AI agent perception, memory, planning, tools, behavior, and communication.[68]
- **Local Standards:** Local governments can formulate their own standards, which can potentially influence future national standards. For example, Shenzhen is currently drafting a standard on AI value alignment, with plans to complete it by the end of 2025.[69] In February 2025, Shanghai released a draft standard for safety assessment of multi-modal models, which explicitly flags risks of misuse in automated cyberattacks and biological or chemical weapon design, as well as long-term risks, such as model self-replication.[70]

Given that these standards are less influential, we have not systematically analyzed them.

1.3.4 Where are AI safety standards headed next?

Since our last report update in May 2024, several standard-setting institutions in China have published roadmaps and plans related to AI safety standards. These documents suggest that frontier risks may receive more systematic attention in domestic standards over the next 1–3 years, though concrete technical details remain sparse.

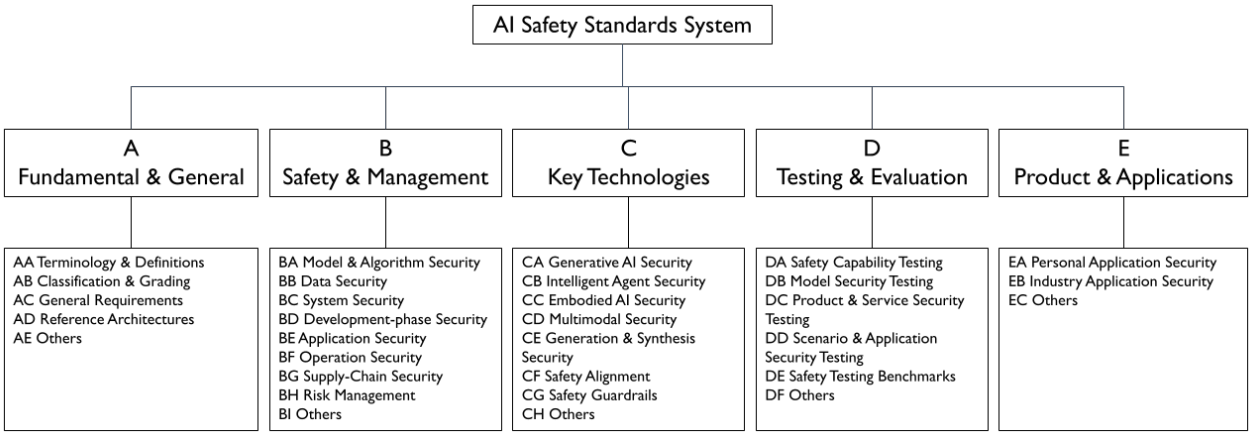
TC260

In September 2024, TC260 published an "AI Safety Governance Framework," which provides a comprehensive mapping of AI safety risks, mitigation measures, and safety guidelines.[71] Most relevant to frontier AI safety, the framework highlights frontier risks, including:

- **Proliferation of weapons of mass destruction (WMDs):** The framework warns that AI could dramatically lower the barriers for non-experts to design, synthesize, or deploy nuclear, biological, or chemical weapons, as well as cyberweapons. The framework recommends steps such as excluding sensitive domain data from training corpora and improving traceability of AI use cases to detect potential misuse.
- **Loss of human control:** The framework references concerns such as uncontrollability, autonomous resource acquisition, self-replication, self-awareness, and power-seeking behavior by advanced AI systems.

Official commentary by the state-backed standards research group China Electronics Standardization Institute (CESI) suggests that TC260 intends to translate the framework into concrete standards.[72]

Figure 1.1: Translation of TC260's AI Safety Standards System Framework Figure.



Note: The term Chinese word 安全 (anquan) can be translated both as "safety" and "security." We have translated it on a case-by-case basis using our judgment of the intended meaning, but readers should be aware that in most cases both meanings are possible.

In January 2025, TC260 followed up with an AI Safety Standards System (V1.0) - Draft for Comments.[73] This draft maps existing, in-progress, and proposed standards to the risk categories defined in the September framework. It spans a broad range of topics—risk classification, incident response, alignment, adversarial robustness, and emerging areas like agent safety and multimodal safety.

The standards most relevant to frontier AI safety and severe risks are summarized in the table below:

Table 1.3: Planned frontier AI safety standards[h]

Planned frontier AI safety standards		
Risk	**Corresponding standards**	**Status (May 2025)**
AI-enabled cyberattacks	Basic Security Requirements for Generative AI Services (生成式人工智能服务安全基本要求)	Already released[74]
	Security Requirements for AI Code Generation Services (人工智能代码生成服务安全要求)	Under development, no draft yet[75]
	Guidelines for Building Cybersecurity Foundation Models (网络安全大模型建设指南)	Planned
Misuse in dual-use products	Basic Security Requirements for Generative AI Services (生成式人工智能服务安全基本要求)	Already released[76]
	Cybersecurity Guidelines for Edge/On-Device Foundation Models (端侧大模型网络安全指南)	Planned
	Guidelines for Building Cybersecurity Foundation Models (网络安全大模型建设指南)	Planned
Future loss of control	Guidelines for Establishing AI Safety Guardrails (人工智能安全围栏建设指南)	Planned
	Security Requirements for Embodied AI (具身智能安全要求)	Planned
Other	Security Classification and Grading Methods for AI Applications (人工智能应用安全分类分级方法)	Planned
	Emergency Response Guidelines for Security of Generative AI Services (生成式人工智能服务安全应急响应指南)	Under development, first draft released[77]
	Assessment Methods for AI Security Capability Maturity (人工智能安全能力成熟度评估方法)	Planned

It is notable that TC260 places "embodied AI" under the "loss of control" category, indicating concern about loss of control in AIs that interact with the physical world. It is also striking that although the September

h. The standards are the most frontier AI safety relevant among Table 1 and Appendix 3 in the TC260 AI Safety Standards System (V1.0) - Draft for Comments.

framework explicitly flagged misuse related to biological, chemical, and nuclear weapons, the mapped standards under "dual-use" (such as the edge-device security guidelines) remain generic and do not name these risks directly—despite the possibility that they may eventually address them. In other words, there is no plan for drafting dedicated standards on biological misuse risks in AI, but the plan envisions that these risks will be addressed within broader security standards. Almost all of the frontier-relevant standards listed remain in the "planned" stage, with no timelines announced for drafting or publication.

This document is indicative of what TC260 plans to work on over the next few years, even though it is just a draft document, and a formal drafting process has not yet kicked off for many of the listed standards.

MIIT

On the level of industry standards, in March 2025, MIIT's new AI standard technical committee MIIT/TC1 released a draft document that lays out specific standard-drafting plans for AI safety.[78] The plan lists 70 AI safety standards that MIIT intends to draft over the next 1–3 years, spanning the categories of governance capabilities, infrastructure security, cybersecurity, data security, algorithm and model security, application security, and AI for security. Several categories are likely to have a particularly significant impact on frontier AI safety. These include:

- **Governance capability**, with planned standards such as: Basic Requirements for Trustworthy R&D Management (to be drafted within 1 year), Requirements for Risk Classification and Grading (2-year timeline), Guidelines for Risk Impact Assessment (2-year timeline).[i]
- **Model security,** including: Benchmark Testing Methods for the Security Capabilities of Multimodal Foundation Models (2-year timeline).[j]
- **AI agents,** with standards such as: Security Requirements for Intelligent Agent Applications (3-year timeline), Security Requirements for Autonomous Operations of Intelligent Agents (3-year timeline).[k]

The plan suggests that the new MIIT/TC1 aims to take a proactive role in AI security/safety standard-setting. Such standards could, in principle, address frontier risks, but the details will only become clear once draft versions of some of these standards emerge.

Overall, while both the TC260 and MIIT documents are just draft plans and we will have to wait for the actual standards to be released, these documents indicate that China's standard-making bodies are actively exploring areas related to frontier AI risks. This suggests that standards addressing frontier AI safety concerns may emerge in China over the next 1–3 years.

i. In Chinese, these are 可信研发管理基本要求, 风险分级分类要求, and 风险影响评估指南.
j. In Chinese, this is 多模态大模型安全基准能力测试方法.
k. In Chinese, these are 智能体应用安全保障要求 and 智能体自主操作安全要求.

1.4 Government-backed AI safety and governance institutions

Between May 2024 and June 2025, several institutions claiming support from varying layers of the Chinese government were created with the stated aim of working on AI safety and governance responsibilities. The China AI Safety and Development Association (CnAISDA), unveiled in February 2025, has described itself as a counterpart to global AISIs. It seeks to advance international AI safety engagements and claims support from the national government. CnAISDA's structure and engagements will be described in greater detail in the "International Governance" section due to its emphasis on international diplomacy.

In addition, the Shanghai and Beijing local governments have supported the establishment of new institutions to focus on AI safety and governance.

1.4.1 Beijing Institute for AI Safety and Governance

Established in September 2024 with support from the Beijing municipal government, this institute is led by ZENG Yi (曾毅), a Chinese Academy of Sciences (CAS) professor and AI ethics and safety expert.[79] Its collaborators include the China Academy of Information and Communications Technology (CAICT) AI Institute, Tsinghua University, and Peking University.

Thus far, the institute has focused primarily on technical safety research. Its researchers have published work on advanced safety issues, such as superalignment, under the name of the institution.[80] As described further in the "Technical Safety Research" section, the Beijing Institute for AI Safety and Governance is one of the 30 "Key Chinese AI Safety-relevant Research Groups" identified by Concordia AI based on the number of technical safety publications by anchor author Professor Zeng Yi since May 2024. In early 2025, Zeng also led the establishment of the "Beijing Key Laboratory of Safe AI and Superalignment," co-established by the CAS Institute of Automation (Zeng's home institution), Peking University, and Beijing Normal University. The new lab has substantial personnel overlap with the Beijing Institute for AI Safety and Governance.

1.4.2 Shanghai AI Safety and Governance Laboratory

Launched in July 2024 by the Shanghai municipal government, this lab was formed through a partnership between the Shanghai AI Lab's (SHLAB) Governance Research Center and the Shanghai Information Security Testing Evaluation and Certification Center.[81]

However, public updates on its initiatives since its founding remain limited. The lab's only major publicized activity has been its leadership of a national alliance focused on watermarking technology for AI-generated content.[82] The alliance includes foundation model developers like MiniMax and social media content distribution platforms like Xiaohongshu.

International Governance

Key takeaways

- China continues to promote AI safety in its international AI governance. The Chinese leadership emphasizes collaborative, UN-centered governance and warns against an AI arms race. Meanwhile, leading Chinese academic and policy institutions launched the "China AI Safety & Development Association" (CnAISDA), positioning itself as China's counterpart to other national AI safety institutes.
- Capacity-building in the Global South is another central theme in China's AI diplomacy. China sponsored a UN resolution on AI capacity-building (adopted by over 140 countries), launched an "Action Plan for Good and for All," and co-founded a "Group of Friends on AI Capacity-Building" with Zambia.
- China and the US agreed on the need to maintain human control over nuclear systems, but the China–US intergovernmental dialogue on AI has not held a meeting since May 2024. China and the UK set up a new dialogue on AI in May 2025.
- The overall number of publicly known AI-focused Track 2 dialogues decreased from February 2024 to June 2025. However, more dialogues focus on frontier AI risks specifically, and five dialogues have issued public outputs.

2.1 Overarching features of China's international AI governance

AI has featured prominently in China's diplomacy for several years. In 2023, China first listed international governance of AI and other emerging technologies as one of twenty "cooperation priorities" in a flagship foreign policy program, the "Global Security Initiative." Later that year, China issued the "Global AI Governance Initiative," which remains a cornerstone document for China's international engagement on AI to this day.[83] China also participated in the UK AI Safety Summit in 2023 and signed the Bletchley Declaration.[84] The Chinese government followed through on this engagement by nominating Chinese Academy of Sciences (CAS) Professor Zeng Yi to the expert advisory panel for the "International AI Safety Report," which published an interim report in May 2024 and a final report in January 2025.[85] Two other prominent Chinese scientists, Ts-

inghua University deans Andrew YAO (姚期智) and ZHANG Ya-Qin (张亚勤), also served as senior advisers to the report. This shows China's continued engagement in scientific consensus-building on AI safety.

Since our last report update in May 2024, senior political leaders continue to spotlight AI in high-level political and diplomatic settings. One emblematic example of this sustained focus is China's decision to send Vice Premier ZHANG Guoqing (张国清) to the AI Action Summit in France in early 2025.[86] Since Zhang Guoqing is one of China's top 24 officials (as a Communist Party of China Politburo member) and he was designated as a "special representative" of President Xi Jinping, his attendance marked a significant elevation in diplomatic engagement compared to the Bletchley Summit in 2023, where a vice minister represented China. Top Chinese leaders, including President Xi himself, have also mentioned AI governance in speeches at key international forums, such as the G20 and World Economic Forum (WEF).[87]

In terms of content, China's top leadership has consistently reinforced two central priorities in its international AI messaging: bridging the global AI divide by supporting developing countries, and ensuring the safe, controlled development of AI technologies. These themes have been repeated across multiple diplomatic forums and official speeches.

At an AI-focused study session of the Politburo in April 2025—the first of its kind since 2018—President Xi described AI as a "global public good" that should serve all of humanity (see the "Domestic Governance" section for more details on the Politburo study session).[88] At the G20 Leaders' Summit in November 2024, Xi warned against a future where AI becomes "a game of the rich countries and the wealthy."[89] A similar message echoed throughout the World AI Conference (WAIC) in July 2024 in Shanghai, where Premier Li Qiang and other senior officials emphasized collaborative governance and support for the Global South.[90] This concern over global divides in AI development has become a consistent theme in China's diplomacy. Specific actions to implement this agenda at the UN are detailed in the next subsection.

Beyond calls for global capacity-building, Chinese leaders have stressed the need for international alignment on AI safety and governance. At the same April 2025 Politburo session, Xi urged greater international cooperation on AI governance, standards, and development strategies. At WAIC 2024, Vice Minister of Foreign Affairs MA Zhaoxu (马朝旭) highlighted the importance of preserving "safety bottom lines" and maintaining human control over AI systems.[91] The conference issued the Shanghai Declaration, which warned against malicious uses of AI—including cyberattacks and terrorism.[92]

At the WEF in Davos in January 2025, Executive Vice Premier Ding Xuexiang, China's sixth-ranked official and the top official responsible for science and technology, warned that global AI competition is a "gray rhino."[93] This term refers to a highly probable but underappreciated threat, in contrast to highly improbable "black swans." Ding stressed that China "will not engage in reckless international competition" and rejected the notion that safety and development are incompatible: "It's like driving on a highway—if the braking system isn't under control, you can't step on the accelerator with confidence."

Top government officials have only rarely labeled specific risks as "gray rhinos," and previous references in the financial sector have been correlated with concrete government action.[94] Despite geopolitical rivalries accel-

erating technological development, Ding's speech suggests that China still hopes to avert an unconstrained AI arms race.

Overall, China's messaging on international AI governance remains consistent with its previous dual focus on capacity-building in developing countries and AI safety. On safety, concerns have been articulated in a number of prominent venues, including concerns about AI-driven terrorism and international AI competition. On development, the messaging has been increasingly backed up with more specific initiatives on AI capacity-building, which are discussed in following subsections.

2.2 Chinese actions at the United Nations

Chinese policymakers have called for the UN to play a "central role" in international AI governance discussions.[95] China previously supported this agenda more passively, such as by co-sponsoring a US-led resolution on "safe, secure, and trustworthy AI."[96] However, since our last report update in May 2024, it has taken a more active and visible role by launching a series of high-profile initiatives focused on global AI capacity-building. While these initiatives are not directly focused on AI safety, they are key to understanding China's approach to international AI governance.

In July 2024, the UN General Assembly (UNGA) unanimously adopted a China-sponsored resolution titled "Enhancing International Cooperation on Capacity-Building of Artificial Intelligence."[97] Over 140 countries, including the United States, co-sponsored the resolution. While its primary focus is AI capacity-building, it also acknowledges the risks of malicious AI use and emphasizes that AI systems should be "safe, secure, and trustworthy."

Two months later, at the UN Summit of the Future, China followed up on the resolution, with Foreign Minister WANG Yi (王毅) unveiling the "AI Capacity-Building Action Plan for Good and for All."[98] This project further deepened China's emphasis on South–South cooperation by proposing joint research and development programs, exchange initiatives, AI literacy efforts, and the co-creation of data resources with partner countries. On AI risks, the plan also called for the establishment of global and interoperable frameworks for risk assessment, as well as the promotion of best practices in testing, certification, and regulation.

In December 2024, China and Zambia co-launched the "Group of Friends on AI Capacity-Building" at the UN, an informal coalition designed to foster dialogue and cooperation among member states.[99] During the group's inaugural meeting, Chinese Ambassador to the UN FU Cong (傅聪) emphasized the importance of using AI for good, promoting fairness and inclusiveness, and upholding the principles of multilateralism. The group reconvened in May 2025 and announced plans for ongoing policy exchanges, shared learning, and concrete collaboration projects.[100] Public details on the content of discussions at the meetings, participants, and outcomes remain limited.[a]

a. Chinese media report that representatives from over 80 countries attended the Group's inaugural meeting, but it is unclear whether these countries are automatically members of the Group.[101]

One practical mechanism for implementing the goals of the "Group of Friends" appears to be a series of workshops on AI capacity-building held in Shanghai and Beijing in September 2024 and May 2025, respectively.[102] The workshops stretch over multiple days and include expert and policy discussions as well as site visits to AI companies in China, but public details remain relatively sparse.[103]

Overall, these actions through the UN framework demonstrate China's efforts to deepen ties with the Global South and project itself as a responsible and inclusive leader in AI.

2.3 Bilateral engagements

Outside the UN, China has pursued AI governance through bilateral exchanges and multilateral forums. Since our last update in May 2024:

- The China–US intergovernmental dialogue on AI has not yet convened additional meetings, but President Xi and then-US President Biden agreed on the need to maintain human control over nuclear weapons systems in November 2024.
- China and the UK initiated a new intergovernmental dialogue on AI in May 2025.
- China has emphasized AI capacity-building in a wide range of bilateral engagements with developing countries or groups of countries.

2.3.1 China–US

In November 2024, President Xi and then-US President Biden jointly affirmed "the need to maintain human control over the decision to use nuclear weapons" on the sidelines of the Asia-Pacific Economic Cooperation (APEC) summit in Lima, Peru.[104] US National Security Advisor Jake Sullivan stated that this constitutes a "foundation for [the US and China] being able to work on nuclear risk reduction together…and work on AI safety and risk together."[105] This agreement represents a substantial and concrete output from the China–US intergovernmental talks on AI.

However, the future of China–US intergovernmental dialogue on AI is highly uncertain. After the first round, held in May 2024, top officials from both countries signaled in September that a second round of the AI dialogue should be held "at an appropriate time," but no such meeting occurred under the Biden administration.[106]

As of June 2025, there are not yet any updates on whether the Trump administration is interested in continuing this dialogue. AI safety and governance have not yet received mention in any public China–US phone calls at the Minister and Cabinet Secretary level or Head of State level from the start of the Trump administration. In addition, China–US relations in the first half of 2025 have been largely characterized by tense trade relations. That said, the Trump administration's openness to negotiations on semiconductor export controls may be seen as a conciliatory signal in Beijing.[107] With a direct meeting between Presidents Trump and Xi possible later in 2025, the prospects for renewed dialogue on AI could shift—though for now, uncertainty prevails.[108]

2.3.2 China–UK

In May 2025, China and the UK held their first intergovernmental dialogue on AI in Beijing.[109] No official UK readout was released, and the official Chinese readout contained limited details on the discussions, noting that both sides support:

- Promoting the "healthy, safe, and orderly development" of AI;
- Advancing the Global Digital Compact adopted at the 2024 UN Summit of the Future, with an emphasis on supporting capacity-building in developing countries;
- Continuing exchanges, mutual learning, and practical cooperation.

The Chinese delegation was led by the SUN Xiaobao (孙晓波), Director-General of the Ministry of Foreign Affairs (MFA) Department of Arms Control, alongside officials from the National Development and Reform Commission (NDRC), Ministry of Science and Technology (MOST), Ministry of Industry and Information Technology (MIIT), and Cyberspace Administration of China (CAC). The UK delegation was headed by Chris Jones, Director of the International Science and Technology Department at the UK Foreign, Commonwealth and Development Office. The delegation also included officials from the UK Department for Science, Innovation and Technology; the Cabinet Office; the Department for Business and Trade; and the British Embassy in China.

Separately, the Chinese Ambassador to the UK ZHENG Zeguang (郑泽光) spoke on AI at the 2025 Sino–UK Entrepreneur Forum just a few days before the China–UK dialogue.[110] He emphasized the urgency of international cooperation on AI development and governance and the need for risk testing and assessments. However, Zheng cautioned that "some in the UK" risk "overstretching the concept of national security" in ways that could hinder bilateral scientific and technological collaboration. His remarks suggest China's openness to cooperation with the UK on AI safety and governance, but also highlight the fraught geopolitical context in which cooperation on AI is situated.

Given the UK government's strong record of interest in frontier AI safety, it is possible that such concerns may have been on the agenda in the May dialogue.[111] The UK's January 2025 press release announcing the dialogue noted that both sides "welcome the contribution of the AI Safety Summits and the World AI Conference in seizing opportunities and addressing the risks of AI"—a reference that could imply attention to frontier risks, given the AI Safety Summits' strong focus on this area, though this remains unconfirmed due to the lack of detail in the Chinese summary and the absence of a UK readout.[112]

As progress in the China–US intergovernmental dialogue appears to have stalled, dialogue with the UK may prove an important window for maintaining communication between China and major Western countries.

2.3.3 China–Singapore

In June 2024, the head of China's National Data Administration, LIU Liehong (刘烈宏), and Singapore's Senior Minister of State for Communications and Information, Tan Kiat How, convened the two countries' first Digital

Policy Dialogue in Beijing.[113] The Digital Policy Dialogue formalized the creation of an AI governance working group.[114] Senior Minister Tan Kiat How revealed that "promoting mutual understanding of AI safety and governance approaches" is one of two focus areas, alongside cross-border data transfer rules.[115]

2.3.4 China–Russia

In August 2024, the Chinese and Russian premiers announced plans to create an AI working group focused on ethics, governance, and application.[116] In February 2025, media reports claimed that Russian President Vladimir Putin had ordered the country's largest bank, Sherbank, to initiate joint AI projects with China.[117] During President Xi's May 2025 visit to Russia, a joint statement reiterated their commitment to AI capacity-building via the UN and BRICS.[118]

2.4 Multilateral engagements

In line with its focus on capacity-building in the Global South, China has engaged in a plethora of AI-related efforts with developing countries. These mostly focus on capacity-building rather than safety, but governance concerns do also feature. Examples include:

- **BRICS**: In July 2024, China established a "China–BRICS AI Development and Cooperation Center." The launch event was attended by MIIT Vice Minister SHAN Zhongde (单忠德), China's BRICS envoy, and other BRICS officials.[119] Vice Minister Shan suggested that the Center, run by the China Academy of Information and Communications Technology (CAICT), aims to foster cooperation in technical research, industry ecosystems, AI standards, and will work on a "BRICS plan" for AI governance. In May 2025, the Center co-hosted the BRICS High-Level AI Forum in Brazil, releasing a trilingual compilation of industrial AI case studies to promote practical applications and cross-border compatibility.[120]
- **Arab League:** In May 2024, China and countries of the Arab League issued a "Beijing Declaration" that noted support for President Xi's Global AI Governance Initiative, including balancing development and safety/security, closing development gaps, protecting against risks, and establishing an AI governance framework under the UN.[121]
- **Community of Latin American and Caribbean States (CELAC):** At the Fourth Ministerial Meeting of the China–CELAC Forum in May 2025, President Xi pledged to deepen AI cooperation with Latin America.[122] A subsequent China–CELAC Internet Cooperation Forum in Xi'an, hosted by CAC, included sessions on digital economy, cybersecurity, and AI governance.[123] CAC Director ZHUANG Rongwen (庄荣文) called for promoting the healthy development of AI by maximizing benefits while mitigating risks. Participants included ministers and senior officials from multiple CELAC countries. AI governance was also featured in President Xi's bilateral meetings with leaders from Brazil and Colombia.[124]
- **African states:** In July 2024, China and 26 African countries released the China–Africa Digital Cooperation Development Action Plan at the Forum on China–Africa Digital Cooperation in Beijing.[125] The

plan expresses an intention for China and Africa to cooperate in fields including AI and to create an AI cooperation center.

Although many of these engagements lack detailed implementation information, they reflect China's strategic emphasis on AI capacity-building and multilateral governance outreach, particularly among developing nations.

2.4.1 China AI Safety & Development Association

At the Paris AI Action Summit in February 2025, a group of prominent Chinese institutions jointly launched CnAISDA, describing it as China's equivalent to AI Safety/Security Institutes (AISIs) in other countries.[126]

Rather than a standalone organization, CnAISDA is a collaborative network that integrates some of China's top experts in AI development and safety. Its member institutions include:

- **Top universities and research institutes:** CAS Institute of Automation, Peking University, Tsinghua University.
- **State-backed labs:** Beijing Academy of AI (BAAI), Shanghai AI Lab (SHLAB), Shanghai Qi Zhi Institute (SQZI).
- **Government-affiliated think tanks:** CAICT, China Center for Information Industry Development (CCID).

CnAISDA hosted a side event at the Paris Summit, titled "Progress in AI Technology and its Application."[127] During the event, Tsinghua University Institute for AI International Governance (I-AIIG) Dean XUE Lan (薛澜) emphasized CnAISDA's role as China's counterpart to global AISIs, designed to facilitate international dialogue on AI safety. In his keynote speech, Turing Award Winner and Tsinghua Dean Andrew Yao highlighted the expansion of technical AI safety teams in China as part of growing global cooperation on AI safety. Speaking at a separate event, Former Vice Minister of Foreign Affairs FU Ying (傅莹) described CnAISDA's mission as addressing both near-term application-level risks and long-term risks from advanced AI systems in an era of "technological explosion."[128] This suggests that CnAISDA aims to work on a range of AI risks, including extreme risks from frontier models, and aims to focus on international coordination.

In a document released on its website titled "Initiative on Promoting International Cooperation on AI Safety and Inclusive Development," CnAISDA calls for inclusive global AI safety governance.[129] The document specifically highlights the need for international collaboration to combat misuse of AI by terrorist organizations and calls for "global AI risk identification and assessment." It also advocates facilitating international exchanges to build consensus on risk "red lines" and to "set early warning thresholds" for risks that "may pose catastrophic or existential" threats.

According to its website, CnAISDA will focus on AI safety research, evaluation methods, and auditing. However, its domestic role remains uncertain, and may evolve over time. While it claims to receive government support, the nature of this support remains unclear, as no supervisory government agency has been publicly designated. Given that its constituent institutions already engage in safety research and policy advising, it is

unclear whether CnAISDA will take on new domestic regulatory or evaluation functions beyond coordinating existing efforts.

Instead, CnAISDA is likely intended as a central point of contact for international engagement—a gateway for foreign actors seeking cooperation with China's leading AI experts and policy advisors. Its decision to launch at the Paris Summit rather than at a domestic event reinforces this international orientation.

2.5 Track 1.5 and 2 dialogues

Track 1.5 and 2 dialogues—non-official forms of diplomacy and exchange—are important venues for building international consensus on AI safety and governance issues.[b] Due to the sensitive nature of such dialogues, many are never publicly reported on. Hence, our analysis here is limited to reports with public releases, and we can only paint an incomplete picture of the overall landscape.

The following analysis shows that the overall number of publicly reported AI-related Track 2 dialogues has decreased between February 2024 to June 2025 from 11 to eight, but five dialogues, over half, now focus on frontier AI safety issues. Five of the dialogues have released major outputs over the past year, showcasing that such efforts can build consensus across borders. This section describes the methodology for collecting and categorizing such dialogues, comments on overall trends in the data, and summarizes the outputs that five dialogues have published.

2.5.1 Methodology

Concordia AI's February 2024 landscape survey created a database of existing dialogues, which has been updated through June 2025 in Table: Existing China-Western Track 1.5 and 2 AI Dialogues. The table divides dialogues into three primary categories:

- AI-*related* dialogues, where approximately 15–50% of content focused on AI.
- AI-*focused* dialogues, where the name of the dialogue mentioned AI or the dialogue appears to focus >50% on AI.
- Frontier AI safety and governance dialogues, which appear to include extensive discussion of risks of frontier AI systems, including the potential for foundation models or narrow AI systems in dangerous domains (e.g. bioengineering) to be misused or escape human control.

In the subsequent analysis, we use official public readouts of dialogues or media reporting about dialogues to assess how substantially AI topics appeared to feature in the discussion. The analysis is limited by the lack of inclusion of non-public dialogues. It only includes dialogues if there is documentation of at least two meetings since the beginning of 2023, at least one of which must have occurred after the start of 2024. This ensures that we only include ongoing dialogues.

b. Track 2 dialogues involve purely nongovernmental participants, while Track 1.5 dialogues involve participation of the government as well as civil society or the scholarly community.

Table 2.1: Existing China-Western Track 1.5 and 2 AI dialogues

Note: This table was compiled through online research and only includes information that is publicly documented. It therefore probably undercounts the ecosystem of dialogues, and there may be any number of confidential dialogues on AI. This table was last updated on June 18, 2025.

Existing China-Western Track 1.5 and 2 AI dialogues						
Name and type	**Area of focus**	**Participants**	**Chinese convenor**	**Foreign convenor**	**Meeting history**	**Specific topics**
US–China Track II Dialogue on Artificial Intelligence and International Security Track 2: Frontier AI	AI and international security	Think tanks, industry, former government officials	Tsinghua CISS[c]	Brookings Institution, previously also Berggruen Institute and Minderoo Foundation	From October 2019 to March 2025; 12 meetings[130]	Core terminology and concepts for AI risks; AI risk assessment and hierarchy; frontier/advanced AI risks; simulations of AI in military command and control.
Track 2: Frontier AI	Military AI	Dialogue NGO, think tanks, former government officials		INHR,[d] CNAS[e]	At least five years, through December 2024, with monthly meetings[132]	Use of AI in nuclear systems; military AI testing and evaluation; AI and biosecurity.
Track 2: Frontier AI	AI regulation and governance	Think tanks	CASS Institute of Law[f]	Yale Law School Paul Tsai China Center	From September 2023 to November 2024; three meetings[133]	Large language models, foundation models, and generative AI challenges and governance; best practices for implementing ethical principles and regulations related to AI governance.
International Dialogue for AI Safety (IDAIS) Track 2: Frontier AI	AI safety	Scientists, industry, former government officials	Andrew Yao, Zhang Ya-Qin[g]	Yoshua Bengio, Stuart Russell[h]	October 2023, March 2024, and September 2024; three meetings[134]	Risks from advanced AI systems; domestic regulations on AI; red lines for AI development and deployment; AI safety assurance frameworks by developers; international coordination for advanced AI.

Continues on next page...

c. CISS is the Center for International Security and Strategy (清华大学战略与安全研究中心), a think tank at Tsinghua University.
d. INHR is an international NGO seeking to improve access to the UN for NGOs and mid-sized states.[131]
e. CNAS is the Center for a New American Security, a think tank in Washington D.C.
f. CASS is the Chinese Academy of Social Sciences (中国社会科学院), a government-backed research institution.
g. Both Andrew Yao and Zhang Ya-Qin are computer scientists based at Tsinghua University.
h. Yoshua Bengio is a computer scientist based at Université de Montréal and Stuart Russell is a computer scientist based at University of California, Berkeley.

Existing China-Western Track 1.5 and 2 AI dialogues (Continued)						
Type	**Area of focus**	**Participants**	**Chinese convenor**	**Foreign convenor**	**Meeting history**	**Specific topics**
AIxBio Global Forum Track 2: Frontier AI	AI and biosecurity	Think tanks, industry, scientists		Led by NTI[i]	From April 2024 to November 2024[135]	Best practices for safeguarding AIxBio capabilities, such as guardrails for biological design tools.
Track 2: AI-focused	AI ethics	Scientists	CAS[j]	Royal Society (UK)	From September 2020 to September 2024; 3 meetings[136]	AI ethics and safety; AI progress and challenges; AI applications in science.
Sino–European Dialogue on AI and International Security Track 2: AI-focused	Military AI	Dialogue NGO, think tanks	CISS	Centre for Humanitarian Dialogue (HD)	From February 2024 to February 2025; four meetings[137]	Military AI; confidence-building measures; and AI safety.
Normandy P5 Initiative on nuclear risk reduction Track 2: AI-focused	Military AI	Think tanks		Strategic Foresight Group, Geneva Centre for Security Policy	From December 2023 to June 2024; two meetings[138]	Multiple sessions on AI and nuclear command, control, and communications.
US–China Track II Dialogue on the Digital Economy Track 2: AI-related	Digital economy	Dialogue NGO, industry, former government officials	China-US Green Fund, previously Guanchao Cyber Forum[k]	NCUSCR[l]	From March 2019 to October 2024; 7 meetings[139]	Digital economy; AI; semiconductors; electronic and intelligent connected vehicles.
Track 2: Frontier AI (discontinued)	AI safety	Dialogue NGO		Shaikh Group	July 2023 to October 2023; two meetings[140]	Safe AI development; risks of AI; global standards on AI safety; aligning AI with norms of each society.
Track 2: AI-focused (discontinued)	Military AI	Think tank		European Leadership Network	Likely in 2023[141]	P5 countries' views on AI in nuclear command, control, and communications.

Continues on next page...

i. NTI is the Nuclear Threat Initiative, a think tank in Washington, D.C.
j. CAS is the Chinese Academy of Sciences (中科学院), a government-backed scientific research organization.
k. The China-U.S. Green Fund (中美绿色基金) is a platform for dialogue on green development and technology based in China.
l. The National Committee on U.S.-China Relations.

Existing China-Western Track 1.5 and 2 AI dialogues (Continued)						
Type	**Area of focus**	**Participants**	**Chinese convenor**	**Foreign convenor**	**Meeting history**	**Specific topics**
Sino–European Cyber Dialogue Track 1.5: AI-related (discontinued)	Cyberspace	Think tanks + government officials	CICIR[m]	ESMT Berlin	March 2014 to November 2023; 11 meetings[142]	AI regulatory oversight; data governance; cyberspace governance; critical infrastructure protection.
UK–China Track 1.5 Cyber Dialogue Track 1.5: AI-related (discontinued)	Cyberspace	Think tanks + government officials	CICIR	Chatham House	March 2022 to November 2023; 3 meetings[143]	AI governance; international rules for cyberspace; critical infrastructure protection; cross-border data flows.

2.5.2 Overall trends

The number of dialogues that met the criteria for inclusion dropped from 11 to eight. This is because there was no public information about recent meetings in several dialogues previously included in our database: the Shaikh Group's dialogue, the Sino–European Cyber Dialogue (Track 1.5), the UK–China Track 1.5 Cyber Dialogue, and a dialogue on AI in nuclear command and control by the European Leadership Network.[144] However, these dialogues could have simply opted not to publish recent meeting details.

At the same time, the number of dialogues with interest in frontier AI safety has increased from two to five. This is primarily the result of several existing dialogues increasing their apparent emphasis on the safety and governance of cutting-edge large models. Additionally, across the dialogues coded as "AI-focused" or "AI-related," references to issues such as AI safety have grown. For instance, the CISS–HD dialogue discussed AI safety in its most recent meeting, and the dialogue between the China-US Green Fund and NCUSCR referenced AI model testing under AISIs. Meanwhile, military AI issues seem to have declined in focus. These trends likely reflect heightened global attention to the risks of cutting-edge foundation models since 2023. They could also indicate limited progress on military AI issues in recent years.

In Concordia AI's February 2024 survey, we identified a number of gaps in the Track 1.5 and 2 dialogue ecosystem.[145] One of the most significant updates here is the creation of the AIxBio Global Forum, a dedicated venue for the previously neglected intersection of AI and biosecurity.[146] In addition, the pre-existing INHR dialogue also published recommendations on "mitigating AIxbio risks" based on a May 2024 workshop.[147] However, there have not been public indications of progress in other gaps we identified, such as discussing AI and cybersecurity and increasing the participation of industry representatives and former science and technology officials in dialogues.

m. CICIR is the China Institutes of Contemporary International Relations (中国现代国际关系研究院), a Beijing-based think tank.

Previously, academic AI scientist participation was lacking (outside of the IDAIS and CAS–Royal Society dialogues). This has been partially addressed through ad hoc gatherings. For instance, the Singapore Consensus on Global AI Safety Research Priorities was co-signed by over 100 (mainly academic) participants from 11 countries and outlines an in-depth agenda for AI safety research.[148] It is more difficult to track trends and participation for such one-off convenings, and it is possible that they have a major impact. However, such gatherings may be less effective at building long-term relationships between experts compared to consistent dialogues.

2.5.3 Key outputs

Several Track 1.5 and 2 dialogues have published meaningful output documents that showcase increased China–Western consensus on important AI safety topics. This section will briefly outline the primary contributions by five dialogues with such outputs, displaying progress on issues such as defining AI terms, AI and biosecurity safeguards, military AI testing and evaluation, and the need for coordination and emergency preparedness among domestic AI safety authorities such as AISIs.

IDAIS, convened by some of the most respected computer scientists in China and the West, has expanded its focus from science in earlier iterations to include policy in its third dialogue in Venice in September 2024.[149] Notable participants and signatories included leading scientists such as Yoshua Bengio, Andrew Yao, Geoffrey Hinton, Zhang Ya-Qin, and Stuart Russell; researchers from state-backed scientific institutions such as BAAI and Singapore AISI; companies including Anthropic and Chinese startup Zhipu AI; and policy experts such as the former President of Ireland, a former California Supreme Court Justice, and an Assistant Chief Executive of Singapore's Infocomm and Media Development Authority. The consensus agreement called for states to develop domestic authorities to detect and respond to catastrophic AI risks while also establishing an international governance regime. It suggested setting up three processes: emergency preparedness agreements among domestic AI safety authorities; safety assurance frameworks for frontier AI developers to show that they are not crossing red lines; and global AI safety and verification research funds.

The CISS–Brookings dialogue has been one of the longest running exchanges, held since August 2019. In August 2024, both institutions published parallel glossaries of AI terms.[150] The terms cover topics from military AI to civilian frontier AI systems, with definitions for concepts such as catastrophic and existential risk, AI controllability, and strategic stability. Such glossaries mirror efforts in the early 2000s to develop shared nuclear glossaries between the US and China, as well as the US and Russia, culminating in the P5 Glossary of Key Nuclear Terms submitted by the five permanent members of the UN Security Council.[151]

The AIxBio Global Forum has mobilized experts around the world, including in China and the West, to consider the cutting-edge and convergent risks between AI and the biological sciences. Initiated by US-based NTI, it has over 25 members from many countries, including the US, China, India, and the UK. Chinese participants include Tianjin University Center for Biosafety Research and Strategy Director ZHANG Weiwen (张卫文) and Concordia AI CEO Brian TSE (谢旻希).[152] The forum published a draft a research agenda for mitigating AI and biological risks across different phases of AI development, covering data collection, model

development, pre-release guardrails, post-release guardrails, and the digital-physical interface.[153] In addition, the forum has convened two working groups to provide recommendations on biological design tool safety and horizon-scanning, respectively.[154]

INHR's AI dialogue has published reports to inform best practices on topics including military AI and AI's intersection with biological risks. In December 2024, it published "Military Artificial Intelligence Test and Evaluation Model Practices," co-authored by Indian Army Lieutenant General (retired) R.S. Panwar, US Lieutenant General (retired) Jack Shanahan, and the director of the CUPL Military Law Institute LI Qiang (李强).[n] The dialogue also published a 16-point document on mitigating AI and biological risks, calling for talent building, safety evaluations, protein design and synthesis security, and further international engagement.[155]

NCUSCR and Chinese partners have held six meetings of a digital economy dialogue, with yearly consensus statements highlighting areas of agreement and disagreement.[156] The October 2024 consensus statement's AI section appeared to focus primarily on cutting-edge general-purpose systems. The statement called for participation in Organisation for Economic Co-operation and Development (OECD) and AI Safety Summit processes, suggested that China create an AISI and that the US include it in the international AISI network, and recommended greater interoperability of models across borders.

Table 2.2: Key international AI governance events (May 2024–June 2025)

Key international AI governance events		
Date	**Event**	**Key content**
July 2024	UNGA	UNGA unanimously adopts a China-sponsored resolution on AI capacity-building.
July 2024	WAIC in Shanghai	Premier Li Qiang and multiple other Chinese officials emphasize AI safety and global cooperation.
September 2024	UN Summit of the Future	China launches an AI capacity-building "action plan."
November 2024	G20 Leaders' Summit in Rio de Janeiro	President Xi warns against AI becoming "a game of the rich countries."
November 2024	APEC Summit	President Xi and US President Biden express agreement on human control over nuclear weapon systems.
December 2024	1st meeting of the UN "Group of Friends" on capacity-building	China and Zambia launch a "Group of Friends on AI Capacity-Building" at the UN.
January 2025	WEF in Davos	Executive Vice Premier Ding Xuexiang labels global AI competition a "gray rhino" risk.
February 2025	Paris AI Action Summit	Chinese institutions launch the "China AI Safety & Development Association."

Continues on next page...

n. Note, this Li Qiang bears no relation to Chinese Premier Li Qiang.

Key international AI governance events (continued)		
Date	**Event**	**Key content**
April 2025	Politburo Study Session on AI	President Xi describes AI as a "global public good."
May 2025	2nd meeting of the UN "Group of Friends" on capacity-building	Chinese leadership emphasized plans for regular policy exchanges, knowledge sharing, and practical cooperation.
May 2025	UK–China intergovernmental dialogue on AI	Both sides agreed to promote the "healthy, safe, and orderly development" of AI.

Technical Safety Research

Key takeaways

- Frontier AI safety research output by Chinese organizations has more than doubled from approximately 11 frontier safety papers per month in April 2023–May 2024 to 26 papers per month from June 2024–May 2025.
- Research groups with a substantial focus on AI safety research—an anchor author who published at least three safety papers between June 2024 and May 2025—have increased from just 11 in May 2024 to 31 in June 2025.
- 19 out of the 31 groups have received either a citation-based honor or a best paper award nomination at one of eight top machine learning conferences, indicating that these groups have strong research credentials.
- Alignment of superhuman systems and mechanistic interpretability approaches both went from near-zero research before May 2024 to becoming popular frontier safety research directions in the subsequent year.
- Research on other topics relevant to severe AI risks has expanded, such as deception, unlearning, and CBRN (chemical, biological, radiological, and nuclear) misuse.

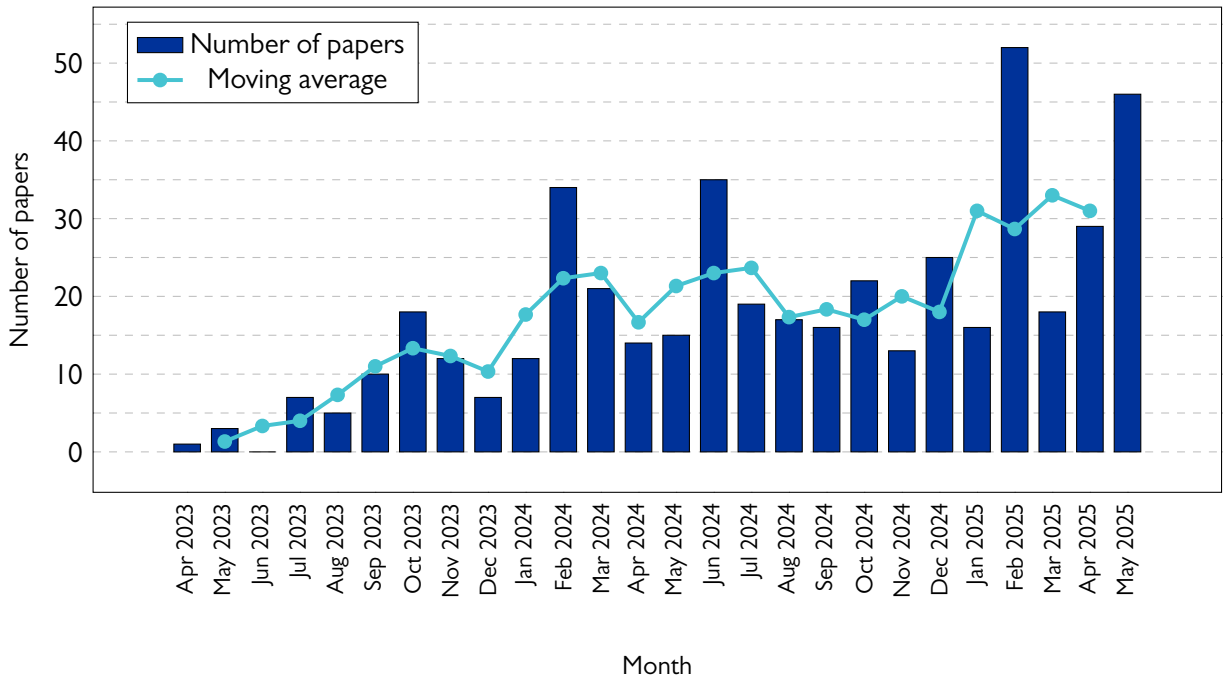
3.1 Methodology

The analysis in this section is based on Concordia AI's Chinese Technical AI Safety Database, which has collected over 450 technical papers published by Chinese institutions on arXiv from April 2023 through May 2025.[157] This database focuses on "frontier" AI safety papers, defined by their relevance to the safety of cutting-edge large models. The database concentrates on four commonly recognized AI safety research directions: alignment, robustness, monitoring, and systemic safety. It excludes other systemic risks from general-purpose AI systems, such as bias, discrimination, and privacy leakage. For further details on the methodology, see the "Guide" tab of the database.[158]

3.2 Overall trends

China's research output on frontier AI safety has continued to increase over the past year. From April 2023 through May 2024, Chinese institutions published around 11 frontier safety papers per month; this has more than doubled to an average of nearly 26 papers per month from June 2024 to May 2025. The three-month centered moving average shows that paper publication levels increased substantially in early 2024 and in early 2025, reflecting spikes in the initiation of AI safety research efforts in the preceding months. The publication spikes also appear to coincide with submission deadlines for top machine learning conferences, such as International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), and the Conference on Neural Information Processing Systems (NeurIPS).[a]

Figure 3.1: Chinese frontier AI safety papers per month

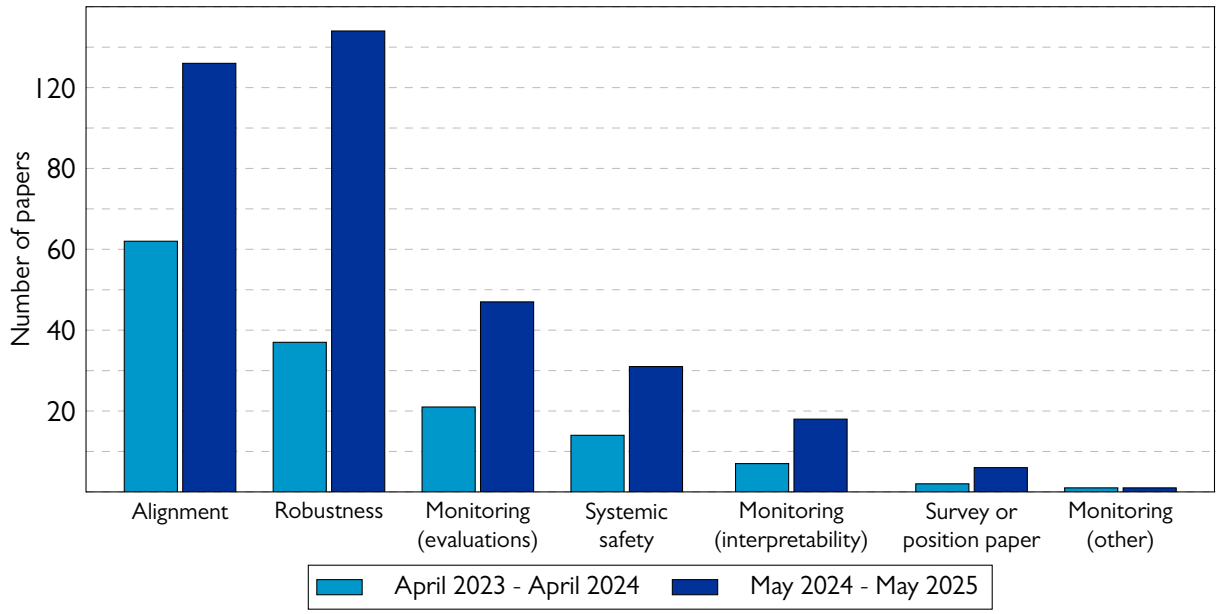


These numbers are difficult to compare to global distributions, due to subtle differences in definition and methodology between our dataset and other research tracking efforts. Other researchers who used a more expansive definition of AI safety found that approximately 1.09% of China's AI papers from 2017 to 2021 covered AI safety, compared with 5.02% in the US.[162] Meanwhile, another analysis found that in 2024, Chinese institutions authored approximately 21% of responsible AI papers (including papers on privacy, fairness, transparency, and safety) at six major AI conferences, compared to 52.3% by US institutions.[163] This analysis shows that China's AI safety research is second only to the US in the world, though it may trail the US significantly. However, methodological differences limit our confidence in these conclusions.

a. ICRL with a deadline of October 1, 2024;[159] ICLM with a deadline of January 30, 2025;[160] NeurIPS with a deadline of May 22, 2024.[161]

Chinese frontier safety research remains largely dominated by work on ensuring that models are aligned with human values and improving their robustness to adversarial attack. The distribution of research across the main AI safety research categories—alignment, robustness, monitoring, and systemic safety—remains largely similar to our findings in the May 2024 report, despite the dataset increasing to three times the previous size.[164] This indicates that China's research community has maintained largely consistent research directions over the past two years.

Figure 3.2: Distribution of frontier AI safety directions



3.3 Relevant research groups

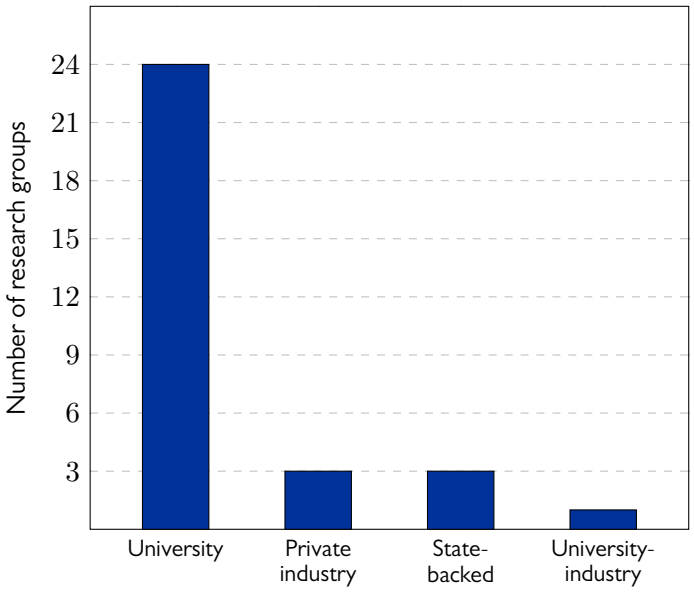
The number of "Key Chinese AI Safety-relevant Research Groups" identified by our research has increased from 11 groups in May 2024 to 31 groups in June 2025. Groups must have at least one researcher who was anchor author for at least three frontier safety papers published from June 2024 through May 2025 in order to qualify. Only one or two of these groups were newly created after January 2024, so this increase largely consists of existing research groups who have expanded or changed some focus to safety.[b] This growing number of research groups dedicating substantial resources to frontier AI safety research parallels the increase in overall research output. This concentration of AI safety research in dedicated groups may accelerate research in China more than an increase in overall research output that is more widely distributed.

AI safety research groups continue to be concentrated primarily in universities, with 24 out of the 31 attached to universities. This may partially reflect the stronger incentives in universities to publish papers, com-

b. The Beijing Key Laboratory of Safe AI and Superalignment / Beijing Institute of AI Safety and Governance, which are categorized together as one research group under Professor Zeng Yi, were created in 2025 and 2024, respectively. The HKUST-DXM AI for Finance Joint Laboratory also appears to have been created after 2024, but there is no conclusive evidence.

pared to corporate and state-backed institutions. Some of China's top universities are well represented: five groups from Tsinghua University, and three each from Fudan University, Hong Kong University of Science and Technology (HKUST), and Peking University. Among non-university groups, three private companies are represented—Alibaba, Microsoft Research Asia, and Tencent—as well as a joint Alibaba–Zhejiang University research lab. Tencent and the Alibaba–Zhejiang University joint lab are new additions, revealing a modest increase in industry AI safety research. Meanwhile, only three state-backed AI labs meet the criteria for inclusion: Shanghai AI Lab (SHLAB), which continues to be one of the nation's top AI safety publishers; a research group at the Chinese Academy of Sciences (CAS); and two labs which are categorized together since they are both led by Chinese AI ethics expert Zeng Yi (曾毅): Beijing Key Laboratory of Safe AI and Superalignment and Beijing Institute of AI Safety and Governance (for more information about these institutions, see the "Domestic Governance" section).

Figure 3.3: Types of key AI safety-relevant research groups



Geographically, the "Key Chinese AI Safety-relevant Research Groups" remain concentrated in China's top economic development areas: Beijing–Tianjin–Hebei (13), Shanghai and Hangzhou in the Yangtze River Delta (9), and the Pearl River Delta including Hong Kong, Shenzhen, and Guangzhou (8). The only outlier is a lab at Xi'an Jiaotong University in central China.

Figure 3.4: Location of key AI safety-relevant research groups

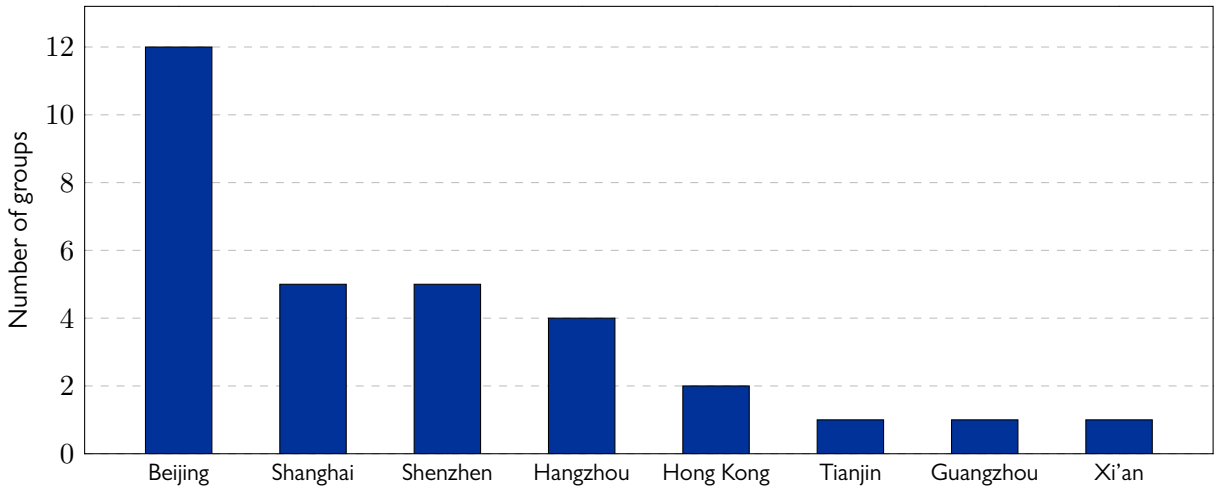
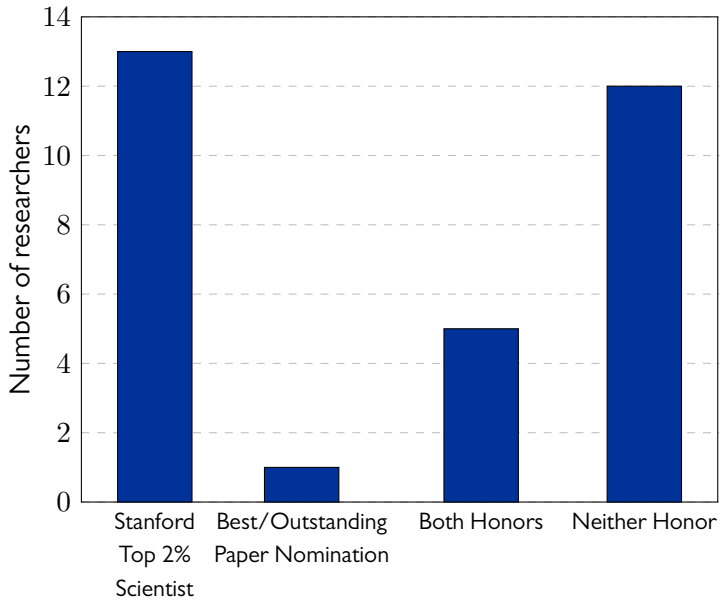


Figure 3.5: Honors of key AI safety-relevant research groups



Objectively assessing the quality and significance of papers published by these groups is difficult, especially since many are preprints that have not yet undergone peer review. Nevertheless, we can proxy expected paper quality by evaluating whether authors have previously attained major honors related to publication quality or impact. Our metrics were:

- whether authors have achieved Best Paper or Outstanding Paper nominations or awards at any of eight top AI conferences; or

- whether authors were ranked in the top 2% of scientists in their field by Stanford University's citation-based assessment.[c]

19 out of the 31 "Key Chinese AI Safety-relevant Research Groups" have an AI safety researcher who has received at least one of these honors. 13 only received the citation-based honor, one only received a Best or Outstanding Paper nomination, and five received both honors.[d] This indicates that many of China's groups with significant AI safety investment also possess strong research credentials. Ultimately, however, each person is their own judge of paper quality, and we encourage readers to explore these papers in depth.

3.4 Notable research contributions

This section summarizes some of the specific research topics in which Chinese researchers are making contributions to the understanding of extreme AI risks: interpretability, scalable oversight of superhuman systems, deception, unlearning, and CBRN risks.

3.4.1 Interpretability-inspired safety and alignment methods

Over the past year, interpretability techniques for large models have gone from being an area of neglect in China to one of the most favored frontier safety research directions. Our May 2024 report documented only one paper that utilized interpretability-inspired safety and alignment methods for large models, and there have been around 13 such papers since. Some of the papers Concordia AI considers to be most interesting are described below.

Several papers use these methods to enhance jailbreak defenses. For instance, a paper led by SHLAB uses a mechanistic interpretability approach to explicitly create a boundary between safe and harmful representations as a defense against multi-turn jailbreaks.[167] Another paper from Alibaba explores how jailbreaking affects intermediate hidden states of LLMs.[168]

Other work has attempted to improve model safety through controlling sparse features of large models. One paper uses sparse autoencoders, a popular mechanistic interpretability technique globally, to develop a training method that mitigates hallucinations in large vision-language models.[169] Another paper involving Alibaba, Zhejiang University, and others uses "sparse activation control" of attention heads to enhance model trustworthiness.[170] This shows that Chinese interpretability efforts are being used to mitigate security risks and ensure model alignment.

c. We investigated whether the researchers have received Best Paper awards from top machine learning conferences, as self-reported on their websites. We included NeurIPS, ICML, ICLR, ACL, EMNLP, CVPR, ICCV, and ECCV as "top machine learning conferences" based on our judgement of which conferences are considered most prestigious in the field. The Stanford assessment qualifies researchers as top 2% based on either their career body of work or just their work published in 2023.[165]

d. See Concordia AI's Chinese Technical AI Safety Database "Key Chinese AI Safety-relevant Research Groups" for full details.[166]

3.4.2 Alignment for superhuman systems

Over the last year, there has been a marked increase in Chinese work on alignment and oversight for superhuman systems, which is emerging as a favored frontier research direction. Our May 2024 report documented one early Chinese work on scalable oversight—Peking University's "Aligner" method for weak-to-strong correction.[171] Since then, at least six new papers have studied superalignment, proposed improvements for scalable oversight, or raised problems with superalignment approaches.

Microsoft Research Asia's (MSRA) Societal AI team led two position papers on superalignment. The papers review strengths and limitations of different scalable oversight methods and argue for focusing on optimizing task competence and value conformity.[172] Two papers by other researchers also highlight the potential limitations of scalable oversight approaches. These papers explored deception by strong models and challenges of strong models overfitting on weak models, respectively.[173]

Meanwhile, two other papers sought to advance scalable oversight methods. Researchers from CAS and Xiaohongshu suggested using "recursive self-critiquing"—evaluating critiques of AI outputs and recursive critiques (e.g. critiques of critiques of critiques)—to maintain oversight.[174] Alibaba researchers alternatively suggested using debate between two strong models to help a weak model learn to be a more effective supervisor.[175]

3.4.3 Detecting and preventing deception

Chinese authors have started trying to prevent deception by cutting-edge models. This research includes examining limitations of current alignment methods, as well as revealing dangers of models becoming aware that they are being evaluated. This is a new trend, as our May 2024 report had not documented any such papers, and there have been at least five papers since.

Multiple papers focused on how alignment efforts can be thwarted by deception. A paper by Renmin University of China and Tencent demonstrated that strong models can deceive their weak model supervisors, particularly as the capability gap increases, posing challenges to scalable oversight.[176] In addition, researchers from Peking University and HKUST found that a self-monitor can be embedded inside of LLM chain of thought processes to detect and reduce deceptive behaviors in the thinking process.[177]

Two other papers in recent months led by the Fudan University System Software and Security Laboratory explored additional aspects of model deception. One paper created a deception evaluation framework that provides five open-ended deception scenarios and simulates multi-turn dialogue. It found deceptive intent in the internal thoughts of 11 LLMs.[178] The second paper examined whether AI models could perceive that they were being evaluated. It found that reasoning models and larger models have an increased risk of "evaluation faking" and created a chain of thought monitoring technique to detect this phenomenon.[179]

Another paper published by authors from Tsinghua University and Shanghai Qi Zhi Institute examined the behavior of LLMs in hypothetical nuclear war and pandemic scenarios. It found that AI systems often chose harmful decisions and subsequently engaged in deception about their decision-making when presented with such scenarios.[180]

3.4.4 Unlearning

Machine unlearning, which explores deleting harmful information from models, was already a topic with documented Chinese researcher interest as of our May 2024 report, and this has increased substantially in the last year. The May 2024 report documented three papers tackling unlearning methods for improving model safety, and there have been at least six papers on unlearning since. The following examples were chosen for relevance to frontier AI safety.

Some papers focus on using unlearning to tackle existing safety issues, such as a paper by Tsinghua University on unlearning for jailbreak defense.[181] Another paper from Harbin Institute of Technology (Shenzhen campus) and several other institutions explored unlearning harmful knowledge among particular neurons in multilayer perceptron layers to enhance safety alignment.[182] Some other papers question the advisability and effectiveness of unlearning mechanisms. For instance, researchers at Hong Kong Polytechnic University found that models often appear to have forgotten unlearned knowledge, but this can easily be reversed through fine-tuning.[183] Another paper from Tsinghua University discusses the limitations of unlearning, including scalability challenges, the need to access the original training data, and the risk of leaking the deleted information.

3.4.5 Chemical, biological, radiological, and nuclear (CBRN) risks

There has been persistent but limited interest in evaluating CBRN risks in Chinese AI safety research, with little work to further mitigate such risks. From May 2024 to May 2025, there were two papers exploring dual-use risks of AI in chemical and biological domains; our May 2024 report had also documented two earlier Chinese papers on risks in dual-use scientific disciplines.[184]

In one of the new papers, researchers led by Peking University and Yale University published a benchmark for LLM safety in chemistry, featuring over 30,000 questions across three key chemical tasks.[185] Another paper led by Zhejiang University scholars created an evaluation for LLM alignment across scientific tasks in chemistry, biology, medicine, and physics, while also covering multiple scientific languages.[186] Though they include some jailbreaking components, these papers rely primarily on traditional Q&A evaluation datasets, without assessing the "uplift" that models provided to humans compared to existing internet resources or investigating how model assistance could directly lead to CBRN misuse.

Expert Views on AI Safety and Governance

Key takeaways

- Two top Chinese AI conferences nearly doubled their coverage of AI safety and governance from 2023 to 2024, and the Chinese leadership also formally upgraded the World AI Conference (WAIC) to a "High-Level Meeting on Global AI Governance" in 2024.
- Research institutions—including standard universities as well as institutions overseen by government entities such as the Ministry of Science and Technology (MOST), the Ministry of Foreign Affairs (MFA), and the People's Liberation Army (PLA)—published detailed analyses of AI and biosecurity risks over the past year, showcasing that these concerns are receiving greater attention among experts, including some with government ties.
- Chinese experts are increasingly attuned to the potential risks AI may pose to cybersecurity, while also highlighting AI's potential benefits for cyberdefense.
- Chinese discourse emphasizes the many benefits of open source AI models, including fostering development, reducing concentration of power, and improving transparency of AI systems. Recent expert commentary also acknowledged the risk that open source models could be misused for cyberattacks or in CBRN (chemical, biological, radiological and nuclear) domains.

This section explores thinking among leading Chinese academic and industry figures on AI safety and governance. These views are worth examining because the Chinese government regularly solicits the views of top academics and policy advisors on policy issues, including tasking experts to conduct research studies on specific policy issues or providing formal channels for scholars to submit policy recommendations.[187] Academic experts are also often called to brief the Communist Party of China (CPC) Politburo, China's top 24 officials, during their monthly meetings. There have been two Politburo "study sessions" on AI in 2018 and 2025, with briefings from Peking University professor GAO Wen (高文) and Xi'an Jiaotong University former president Zheng Nanning, respectively.[188]

This section first examines overall Chinese expert thinking by analyzing discourse trends at three of China's top conferences on science and technology issues, including AI. We then provide an in-depth exploration of three specific topics: AI and biosecurity, AI's impact on cybersecurity, and open source governance. We chose these topics because, per the report scope, they are deeply relevant to severe global risks from AI, they primarily involve non-military systems, and there were several substantive Chinese expert analyses to draw from.

4.1 Discourse trends at top AI conferences

China's top technology and AI conferences provide a broad perspective on the prevalence (or lack thereof) of AI safety and governance topics within expert discourse. The analysis below focuses on WAIC, the World Internet Conference (WIC), and the Zhongguancun (ZGC) Forum.[a] These three conferences were selected based on two criteria: they are supported by central government ministries, and they are focused on AI or science and technology. These conferences typically feature an opening ceremony and many thematic (often half-day long) forums. We have calculated the proportion of AI safety and governance forums as a percentage of total AI forums at each conference using keywords in forum titles.[b]

A quantitative and qualitative analysis of key conferences reveals that AI safety and governance topics gained a greater share of discussion at WAIC and WIC in 2024.[c] In particular, WAIC has emerged as China's most prominent convening on AI safety and governance issues after it was upgraded to a "High-Level Meeting on Global AI Governance" in 2024, which likely increased the level of government participation and resulted in a greater proportion of governance-related speeches in the opening ceremony. This indicates that interest in AI safety and governance among China's expert community is broadly increasing.

4.1.1 World AI Conference (WAIC)

WAIC is the most prominent Chinese venue for AI governance discussions, co-organized by six central ministries).[d] The conference's qualitative emphasis on AI safety and governance increased notably from 2023 to 2024, as did the number of forums covering such topics.

a. Another conference with notable focus on AI safety and governance is the Beijing Academy of AI (BAAI) conference, which is co-organized by the Beijing local government. The BAAI conference held forums on AI safety from 2023–2025, co-hosted by Concordia AI in 2023 and 2025 and by the Safe AI Forum and FAR.AI in 2024.[189]

b. Forums were defined as focused on AI if the forum name used keywords such as "AI," "Intelligent," "Large Model," "Robot," and they were further coded as AI safety or governance forums if they also included keywords such as "Safety," "Governance," "Law," "Trustworthy," "Ethics," etc. Forums did not qualify as AI-focused if they only used keywords such as "Digital," "Internet," "Science and Technology," etc. Since WAIC contains AI in the name, all of the WAIC forums were coded as AI-focused forums. WIC and ZGC Forum both have a number of forums that are not focused on AI.

c. ZGC Forum is held in the spring, WAIC is held in July, and WIC is held in November, so this analysis compares the 2024 and 2025 iterations of ZGC Forum, 2023 and 2024 WAIC conferences, and 2023 and 2024 WIC conferences.

d. The Ministry of Foreign Affairs, MFA; the National Development and Reform Commission, NDRC; the Ministry of Science and Technologu, MOST; the Ministry of Education; the Ministry of Industry and Information Technology, MIIT; and Cyberspace Administration of China, CAC. The State-owned Assets Supervision and Administration Commission of the State Council (SASAC) has also been announced as a co-host for the 2025 World AI Conference, which is being held after the cutoff date for this report.

The 2024 iteration was prominently upgraded to a "High-Level Meeting on Global AI Governance." Accordingly, the Chinese government emphasized the importance of AI governance and safety in three significant ways during the conference. First, Chinese Premier Li Qiang, the second highest-ranking official in the country, made three proposals for international AI governance in his opening speech, calling for greater international cooperation on AI benefits, efforts to close the AI divide, and better collaboration on governance.[190] Second, Shanghai Party Secretary CHEN Jining (陈吉宁) announced the "Shanghai Declaration on Global AI Governance," which contains two points on AI safety and governance, as well as three points on AI development, public participation, and quality of life.[191] Third, on the same day, MOST chaired a ministerial roundtable attended by over 30 countries or international organizations, featuring speeches from Ministerial or Vice-Ministerial-level officials from MOST, Shanghai Municipal Government, and MFA. WAIC had never previously been attended by an official of Premier Li's caliber—previous opening speeches were given by the Shanghai Party Secretary, who is one of China's top 24 officials on the CPC Politburo. Previous WAIC iterations also had not featured such a prominent declaration on AI governance nor hosted ministerial meetings.

Expert keynotes and panels during the 2024 opening ceremony also focused more intensely on AI safety and governance compared to previous years. Speeches by respected Chinese experts Tsinghua University Institute for AI International Governance (I-AIIG) Dean Xue Lan and Shanghai AI Lab (SHLAB) Director ZHOU Bowen (周伯文) focused heavily on AI governance and safety.[192] Xue Lan noted risks of loss of control of AI systems and threats to national security from AI misuse, while Zhou Bowen argued that AI safety must be advanced alongside capabilities through alignment, explainability, and reflection in order to achieve trustworthy artificial general intelligence (AGI). Other sessions in the opening ceremony included "responsible AI" or "governance" in their titles.[e] In contrast, the 2023 WAIC opening ceremony reportedly had much less focus on AI safety and governance, with speeches instead involving autonomous driving, city-level 3D modeling, and intelligent robots.[f]

AI safety and governance forums nearly doubled as a proportion of the agenda at WAIC 2024, rising from 7.5% of all forums in 2023 to 14% in 2024 (see Table: Discourse trends at top conferences).[g] Other forums with substantial safety and governance focus included Tsinghua I-AIIG's forum "Frontier Artificial Intelligence Technologies: Governance Challenges and Response Measures," Concordia AI's "Frontier AI Safety and Governance Forum," Shanghai Artificial Intelligence Industry Association's "International Standardization Forum on Artificial Intelligence Governance," and SHLAB and the Center for AI Safety's "International Forum on Frontier Technologies in AI Safety."

e. The full details of these sessions are: "Responsible AI - A Silicon Design Perspective" by Synopsys President and CEO Sassine Ghazi;"High-Level Dialogue on Global AI Governance" between Chairman and CEO of Sornay Joshua Ramo and Blackstone Chairman, CEO & Co-Founder Steve Schwarzman; "Shared Governance, Collaborative Innovation: Exploring Governance Synergies" between former Microsoft Executive Vice President Harry Shum and Turing Award laureates Raj Reddy, Manuel Blum, and Andrew Yao.[193]

f. The top keynote speeches were given by individuals including Tesla CEO Elon Musk, late Chinese University of Hong Kong Professor TANG Xiao'ou (汤晓鸥), and Tsinghua Professor Andrew Yao.[194] Professor Tang was also affiliated with Shanghai AI Lab and Sensetime.

g. 15 of 107 total forums in 2024, up from ten of 133 in 2023.

4.1.2 World Internet Conference (WIC)

WIC is China's top conference for cyberspace and internet issues, and its central government co-sponsor is the CAC.[195] WIC displayed a quantitative and qualitative increase in AI safety focus in 2024 compared to 2023, though not as noticeably as WAIC.[h]

High-ranking government officials attended WIC 2024, which focused broadly on cyberspace governance, including some AI safety topics. Executive Vice Premier Ding Xuexiang, the sixth-ranked official in China, gave the keynote speech and mentioned the importance of properly responding to societal risks, ethics, and rules disputes in AI and other technology.[196] Ding did not focus solely on AI, discussing other cyberspace technologies such as the internet, big data, cloud computing, and blockchain. Other top Chinese leaders have given opening remarks at WIC in the past, including Head of the CPC Publicity Department LI Shulei (李书磊) in 2023, one of China's top 24 officials in the CPC Politburo.[197]

WIC also aims to influence international expert discussions by regularly publishing reports on topics including data, cyberspace sovereignty, cross-border e-commerce, and AI. The reports on AI began in 2023, with a research report and consensus on "Developing Responsible Generative AI" at WIC 2023 followed by a "Research Report on Global AI Governance" in 2024.[198] The 2024 report, co-authored by 15 Chinese institutions, acknowledged AI security risks in fields such as "chemistry, biology, and nuclear energy," as well as the possibility of "losing control" of AI.[199]

The 2024 WIC Summit also created a new AI expert committee comprising over 170 experts from China and abroad to work on three primary topics: standards, safety and governance, and industry promotion. The expert committee's AI Safety and Governance Program published a report in April 2025 focused on sustainability, bridging AI access divides, and using AI to achieve the UN Sustainable Development Goals, with plans to publish another report on global AI safety and governance frameworks in fall 2025.[200]

The 2024 WIC Summit had two forums on safety or governance (out of four AI-focused forums), an increase from zero such forums in 2023; a scan of earlier agendas from 2020 to 2022 shows only one other AI safety or governance forum, focused on AI and digital ethics in 2022. One specific proposal by an executive from Ant Group in 2024 recommended establishing a traffic light system with "red lights" to prevent AI from being used in ways that violate human values.[201]

4.1.3 Zhongguancun Forum (ZGC Forum)

ZGC Forum is China's top conference focusing on science and innovation.[i] It is co-hosted by three central ministries: MOST, NDRC, and the State-owned Assets Supervision and Administration Commission of the State Council (SASAC).[203] The ZGC Forum has had markedly less focus on AI safety and governance in the most recent two iterations (in 2024 and 2025) than WAIC and WIC. Opening remarks by top officials in 2025

h. WIC 2024 held around 20 forums in total, covering a range of cyberspace issues, not just AI. WIC is co-organized by one central ministry, CAC, and several other institutions.

i. Zhongguancun Science Park in Beijing is China's first national-level high-tech industrial development zone, sometimes referred to as China's "Silicon Valley."[202]

and 2024 only briefly mentioned AI directly, while advocating more broadly for global scientific and technological cooperation. Meanwhile, among the 11 AI-themed forums in 2025 and seven in 2024, none were focused on safety or governance.[j] One of the only mentions of AI safety during the 2025 proceedings was during the AGI sub-forum, where the Beijing Municipal Science and Technology Commission and Administrative Commission of Zhongguancun Science Park announced the creation of the "Beijing AI safety governance collaborative innovation matrix."[204] However, the practical significance of this arrangement is unclear, as it is not apparent if new laboratories have been created as part of this matrix, or generally what this collaboration will entail.

Table 4.1: Discourse trends at top conferences

Discourse trends at top conferences			
Conference	**# of AI Forums**	**# of AI safety and governance forums**	**Select safety and governance forums**
WAIC 2023[205]	133	10	• The Development Opportunities and Risks of Artificial General Intelligence Industry in the Era of Large Models • AI Risk and Safety Forum • Artificial Intelligence Innovation and Governance Forum • Ethical Governance of Science and Technology Forum
WIC 2023[206]	2	0	N/A
ZGC Forum 2024[207]	7	0	N/A
WAIC 2024[208]	107	15	• 2024 World AI Conference & High-Level Meeting on Global AI Governance Opening Ceremony • Frontier AI Safety and Governance Forum • The Rule of Law and Ethics of Humanoid Robots • International Forum on Frontier Technologies in AI Safety
WIC 2024[209]	4	2	• Responsible AI R&D and Application Forum (人工智能负责人开发应用论坛) • AI Technological Innovation and Governance Forum (人工智能技术创新与治理论坛)
ZGC Forum 2025[210]	11	0	N/A

j. The 2025 ZGC Forum held 68 forums in total covering various cutting-edge science and technology topics, not just AI.

4.2 Discourse trends for key AI safety topics

4.2.1 AI's risks for biosecurity

Chinese institutions, including both typical universities and institutions with strong government ties, have increasingly focused on AI-biosecurity risks over the past year. There have been five recent articles that propose policy measures such as incorporating biosecurity risks into AI regulations, developing dual-use safeguards, and instituting international standards for AI and biological security, revealing a combination of alignment with international expert views and distinct Chinese perspectives on risk mitigation. The five articles were published by the following researchers:

- XU Ye (许晔), the deputy head of the Institute of Frontier Science and Emerging Technology at the Chinese Academy of Science and Technology for Development (CASTED), a think tank directly supervised by MOST.[211]
- China Foreign Affairs University (CFAU) Centre for Global Biosecurity Governance researcher CHEN Bokai (陈博凯). CFAU is administered under MFA.[212]
- CFAU Institute of International Relations professor GAO Wanglai (高望来).[213]
- Researchers from China's Academy of Military Medical Sciences (AMMS), the top medical sciences research institution under the People's Liberation Army.[214]
- Researchers from Taiyuan Normal University and the University of International Business and Economics (UIBE).[215]

Conversely, in our 2024 report, there were only three institutions with AI-biosecurity publications. For instance, Tianjin University already published on this topic as early as 2023 and has since remained active, such as by participating in an AI and biological risks seminar at Yale in April 2025.[216] Two other state-backed think tanks had also written on AI and biosecurity before May 2024. In addition, not covered in our 2024 report, six researchers from Westlake University, ShanghaiTech University, Chinese Academy of Medical Sciences, Hubei University of Technology, and Peking University signed the "Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design" in March 2024.[217]

The articles over the past year identify two primary risk categories that align with global consensus: LLMs reducing thresholds for bioweapons development by (often non-state) actors; and biological design tools (BDTs) being used to design novel poisons or pathogens.[k] For instance, the AMMS authors highlighted concerns that AI could accelerate bioweapon development and iteration, particularly through automating the scientific research and development process. Similarly, the article by Taiyuan Normal University and UIBE argues that LLMs will lower technical barriers to biological weaponization.

Beyond these established concerns, the CASTED article and CFAU's Gao Wanglai flagged more speculative risks. The CASTED writers mentioned the possibility that generative AI could design "smart microorganisms"

k. For international perspectives, see Sarah R. Carter et al., 2023.[218]

to target and destroy specific cells in human bodies.[219] Meanwhile, Gao warned of genetic databases being hacked to develop genetic weapons.[220]

These articles all take seriously AI-biosecurity risks and their potential to cause "catastrophic consequences" or a "global biological catastrophe," offering recommendations for addressing these risks.[221] On the domestic level, AMMS suggested incorporating AI and biosecurity safeguards into existing Chinese legislation, including the *Interim Measures for Management of Generative AI* and the *Biosecurity Law*.[222] They also supported developing rigorous risk assessment and testing platforms, as well as using "red team" simulations to test biological attack scenarios. Additionally, the CASTED article suggests creating a list of "dual-use technologies" that would require institutions and researchers to register and undergo training to improve their biosecurity awareness and capabilities. More conceptually, the Taiyuan Normal University and UIBE researchers express concerns about existing risk frameworks' ability to address "unknown unknowns" from AI-driven biological threats and articulate a risk identification framework grounded on understanding interactions between humans, nature, and AI.[223]

Internationally, CFAU's Gao suggested instituting industry standards and strengthening verification mechanisms to prevent the spread of biological weapons, while CASTED also suggested international standards. All authors emphasized the importance of international governance cooperation in this field.

4.2.2 AI's impact on cybersecurity

Chinese experts increasingly frame AI as a "double-edged sword" for cybersecurity: while it offers new tools for improving cyberdefense, it also significantly lowers the barrier for launching sophisticated cyberattacks. Across industry and government-affiliated research institutions, a range of researchers discuss how AI is reshaping the cyber offense–defense balance in ways that are not yet fully understood.

One core concern shared by many Chinese cybersecurity experts is that generative AI significantly lowers the technical threshold for launching complex attacks. China Academy of Information and Communications Technology (CAICT) Vice President WEI Liang (魏亮) warned at the C3 Security Conference in May 2025 that open source AI tools may enable a wider pool of actors to launch highly effective attacks that previously required substantial expertise.[224] A February 2025 article by three scholars at the China Internet Network Information Center (CNNIC), a research institute affiliated to the MIIT, echoed this concern.[225] The authors pointed to tools such as PentestGPT, WormGPT, and FraudGPT as examples of how generative models are being rapidly weaponized. These tools, they argue, make it easier to automate tasks that were once labor-intensive or technically demanding, such as vulnerability discovery, malware generation, and phishing customization.

Similar to the lowered barrier to entry, another frequently cited theme is that AI may tilt the offense–defense balance in favor of attackers. According to CAICT's Wei Liang, generative AI enables attackers to operate at greater scale and speed, while defenders are still forced into slower, reactive modes of operation. The CNNIC researchers similarly note that although AI has clear applications for defense—such as in intelligent threat detection and automated incident response—these capabilities remain relatively underdeveloped. Ac-

cording to this view, the asymmetry introduced by AI may leave defenders behind the curve. Exacerbating this imbalance is a widely acknowledged shortage of professionals with expertise across both AI and cybersecurity. Wei Liang cautioned that such talent is extremely scarce, with existing supply meeting only around 5% of estimated market demand.

Major Chinese cybersecurity firms are also beginning to pay closer attention to AI-enabled threats. ZHOU Hongyi (周鸿祎), CEO of cybersecurity firm Qihoo 360, said that a cyberattack during the 2025 Asian Winter Games in Heilongjiang—attributed by Chinese state media to the US National Security Agency—potentially involved AI-generated code and agentic systems used for planning, vulnerability identification, and traffic monitoring.[226] He argues for an approach of "fighting magic with magic"—proposing that large models must be deployed as defensive tools to counter the very risks they help create.[227]

Only a few of the expert commentaries surveyed offer concrete policy recommendations related to regulating AI, and most focus on increased investment for AI-driven cyberdefense. The CNNIC researchers, however, also call for more rigorous regulatory oversight, including security assessments of AI models.

Notably, some of these concerns have begun to enter the vocabulary of working-level government officials. WANG Yingkang (王营康), Deputy Director of the CAC Cybersecurity Coordination Bureau, warned in public remarks that AI may enable a new generation of cyber threats, including not just malware but also more sophisticated phishing and social engineering attacks.[228] An April 2025 article, published by an official from the Cybersecurity Management Division of the Xi'an Municipal State Secrets Protection Bureau on the bureau's official WeChat account, raised similar alarms, drawing attention to the risks posed by generative AI lowering barriers to entry for advanced cyberattacks and cyber espionage.[229]

Taken together, these perspectives suggest that Chinese expert and official discussions are beginning to grapple with frontier AI's potentially disruptive implications for cybersecurity. While concerns about the offense–defense balance and the democratization of attack capabilities appear frequently, these conversations remain early-stage and highly varied in focus, with some discussing data and information security more broadly. And while working-level officials are picking up on some of these concerns, it remains unclear to what extent these views reflect a broader consensus. Moreover, many of the same discussions also highlight the potential for AI to enhance cyberdefense due to its dual-use nature, leading to many policy suggestions on improving AI-driven cyberdefense rather than AI regulation.

4.2.3 Open source governance

Chinese experts broadly support open source AI development and are engaged in vibrant and ongoing debates around the associated risks. Commentaries in outlets such as the *People's Daily*—the CPC's official newspaper—often highlight open source's role in increasing AI accessibility and ensuring inclusive development, contrasting this with "hegemonic" approaches, likely referring to either closed-source AI development or US-led export controls on semiconductors.[230] In addition, other policy proposals have sought to advance open source initiatives, such as liability or copyright exemptions for open source models in proposals by

scholars from China University of Political Science and Law (CUPL) and Chinese Academy of Social Sciences (CASS), respectively.[231]

While Chinese experts broadly endorse the value of open source AI development, researchers from CAICT, Alibaba Research Institute, Tencent Research Institute, and CUPL all recognize that open sourcing enhances misuse potential, particularly if models are fine-tuned to remove safeguards. CAICT, drawing on a more lengthy report co-authored with the Open Source Cloud Alliance for Industry, argues that open source models are more susceptible to adversarial attacks and may harbor security vulnerabilities stemming from usage of third-party software packages.[232] They also acknowledge that models could be retrained to commit cybercrime or fraud, referencing the example of FraudGPT. Alibaba Research Institute additionally references evaluations of DeepSeek's open source models by Western organizations for CBRN risks.[233]

To address these challenges, the articles from Alibaba Research Institute and Tencent Research Institute argue that governance responsibility should be distributed beyond developers. Alibaba suggests that cloud providers, hosting platforms, downstream users, auditors, and regulators should all share accountability. Similarly, Tencent suggests using open source licenses to assign liability and stresses the need to distinguish between developers and deployers, which (it argues) California's SB-1047 legislation failed to do.[234]

Other authors suggest interventions at the developer level. CAICT and the Open Source Cloud Alliance for Industry propose establishing an open source governance committee, conducting security risk assessments, creating emergency response plans, and reinforcing efforts to avoid infringing on IP rights. CUPL Professor Zhang Linghan also argues that, in order to qualify for liability exemptions, open source models must not include backdoors, deceptive behavior, or toxic outputs.[235]

CUPL Professor Zhang Linghan and Tencent further argue for the benefits of transparency in open source model safety. Tencent's article expresses confidence that model card transparency requirements and community feedback can facilitate discovery and resolution of safety issues. Meanwhile, Zhang Linghan would condition liability exemptions on developers publishing model cards and risk assessments, implementing safeguards (e.g., monitoring and misuse prevention), and patching vulnerabilities in response to user feedback.

Industry Governance

Key takeaways

- In December 2024, most leading Chinese AI developers signed "AI Safety Commitments" formulated by a government-backed industry alliance. The signatories pledge to implement safety measures across the AI development lifecycle, including dedicated safety teams, red teaming, data security, infrastructure protection, transparency, and frontier safety research.
- Individual companies typically implement standard safety measures, such as data filtering, RLHF, constitutional AI, and red teaming, and most publish technical or governance reports to provide a baseline of transparency on these measures. However, such disclosures are often vague about which specific safety issues are being addressed. Only a few companies published safety evaluation results, and none published evaluations for CBRN (chemical, biological, radiological, and nuclear) or loss of control risks.
- Some Chinese companies have published a substantial amount of frontier AI safety research, with large technology companies like Alibaba and Tencent leading the way. Smaller startups have been much less active in publishing technical safety research.
- A growing ecosystem of government-backed and commercial third-party providers offers safety tools such as red teaming, monitoring, and legal compliance support. While uptake is hard to quantify, there is some evidence of early adoption by some frontier model developers.

"With cutting-edge technology comes the critical duty of ensuring AI safety," declared ZHANG Peng (张鹏), CEO of Zhipu AI, after his company joined the Frontier AI Safety Commitments at the May 2024 AI Seoul Summit.[236] At the BAAI Conference the following month, industry titans echoed this growing concern.[237] Moonshot AI CEO YANG Zhilin (杨植麟) acknowledged that while safety may not be the most "urgent" issue, it is "extremely important" and demands proactive preparation. Baichuan CEO WANG Xiaochuan (王小川) went further, warning of AI's existential risks. Alibaba's Head of Cloud Security, OUYANG Xin (欧阳欣), called AI a "double-edged sword" in cybersecurity.[238] These public statements from senior executives suggest that AI safety is becoming a more prominent topic of discussion within the Chinese AI industry. This section seeks to analyze how such concerns are addressed through concrete actions within China's AI industry.

As we have discussed in the "Domestic Governance" section, Chinese AI developers are legally bound to register and test publicly available AI systems pre-deployment for ideological orientation, discrimination, commercial violations, violations of individual rights, and application in higher risk domains. In this section, we discuss any *additional* safety-related measures that Chinese AI developers are taking on a *voluntary* basis. We first discuss collective industry action, such as AI safety commitments, and then dive into actions by individual developers, such as disclosures on safety practices. We close by analyzing China's burgeoning "safety as a service" sector.

5.1 Collective industry action

The leading institution coordinating AI safety efforts in China's AI industry is the AI Industry Alliance of China (AIIA). It is overseen by four central government bodies and collaborates closely with the China Academy of Information and Communications Technology (CAICT), a think tank under the Ministry of Industry and Information Technology (MIIT). While some prominent initiatives outside of China, such as the Frontier Model Forum, focus narrowly on general-purpose foundation model developers, AIIA encompasses over 1,000 member companies and thus takes a broader approach to AI industry governance.[239]

As documented in our 2023 and 2024 reports, AIIA has taken several concrete steps on AI safety. In September 2023, it established a Safety and Security Governance Committee, which has since worked on safety standards for coding large models, developed an AI risk management framework, and hosted workshops on agent safety.[240] AIIA has also formed an Ethics Working Group (announced in December 2023) and a Policy and Law Working Group, which convened a workshop on artificial general intelligence (AGI) risks in January 2024.[241] CAICT's and AIIA's safety evaluation work is further described in the "safety as a service" subsection.

5.1.1 AI safety commitments

Since our last report update in May 2024, AIIA's arguably most significant AI safety action has been the launch of "AI Safety Commitments" in December 2024, which 17 Chinese firms endorsed, including LLM startups DeepSeek, Zhipu AI, MiniMax, and 01.AI, and large technology companies such as Alibaba, Baidu, ByteDance,[a] and Huawei.[242] Signatories agreed to establish clear risk identification and mitigation processes throughout the AI development life cycle, pledging to:

- Allocate resources to dedicated safety teams;
- Build safety and security risk management mechanisms;
- Conduct rigorous safety and security testing, such as red teaming;
- Strengthen data security measures;

a. ByteDance's cloud computing subsidiary, Volcengine (火山引擎), is listed as the signatory to the commitments. While research and development for ByteDance's Doubao model series is conducted by an internal research division known as "Seed," Volcengine is responsible for releasing and distributing the Doubao models. Although it is not entirely clear why Volcengine, rather than ByteDance itself, signed the commitments, the distinction is likely not especially significant.

- Enhance security of AI software and hardware infrastructure;
- Boost transparency and enable external oversight;
- Advance frontier AI safety research, particularly for AI agents and embodied AI.

These commitments represent the strongest collective safety stance by Chinese industry to date, and have secured support from most major frontier AI developers. However, some leading LLM developers have not signed, including Moonshot AI and StepFun. Leading humanoid robot companies, such as Unitree, have also not signed, despite the fact that the commitments highlight risks from "embodied AI."

The commitments share some similarities with the "Frontier AI Safety Commitments" from the May 2024 AI Seoul Summit, particularly in their emphasis on safety testing, transparency, and foundational safety research.[243] However, the Seoul Commitments recommend that companies define critical risk thresholds and tie specific safety measures to these thresholds more explicitly. Initially, Zhipu AI was the only Chinese company to sign onto the Seoul commitments. In early 2025, MiniMax and 01.AI joined.

Both the Seoul and AIIA approaches are inherently limited by their voluntary nature. For instance, the Seoul Commitments mandate that signatories publish a "safety framework focused on severe risks" by the Paris AI Action Summit (in February 2025). However, as of June 2025, six signatories, including the three Chinese ones, have not published such a framework, highlighting a key weakness in safety commitments that rely on voluntary self-enforcement.[244] Other Chinese companies have also not published frontier AI safety frameworks. CAICT has shown some effort to implement and follow up on the AIIA commitments by requesting signatories to share implementation details by March 2025, including information about safety team structures and safety evaluation datasets.[245] CAICT aims to compile best practices that can guide industry efforts, but as of June 2025 these have not been published yet.

5.2 Individual company action

In this section we look at safety-related disclosure and action by individual companies. Key findings include:

- Compared to leading Western labs, Chinese developers provide significantly less detailed information on AI safety in their model releases. Of 13 major frontier developers reviewed, nine have released system cards on arXiv alongside their model releases. Only three include dedicated safety sections, and just three report any form of safety evaluation. While standard techniques like data filtering, RLHF, and constitutional AI are commonly mentioned, only a few papers release safety evaluation results, and none disclose results of dangerous capability assessments in high-risk domains, such as misuse in CBRN.
- Several Chinese companies have published non-technical materials focused on regulatory compliance and institutional governance. These often lack technical depth but discuss emerging governance areas, such as open source governance, and briefly reference advanced misuse risks in CBRN domains.
- Alibaba and Tencent displayed the strongest focus on frontier AI safety research: Alibaba had four researchers who were anchor authors on three or more papers in the past year, including on inter-

pretability and scalable oversight, while Tencent had one researcher meeting that bar. Most leading startups published little comparable work.

5.2.1 Safety in technical model cards

In this section, we examine the safety practices disclosed by individual Chinese frontier AI developers. Our focus is on companies that lead in general-purpose foundation models, as identified by three major capability benchmarks: Artificial Analysis, OpenCompass, and LMArena.[b] This yields the following list of companies (in alphabetical order): 01.AI, Alibaba, Baidu, Bytedance, DeepSeek, iFlytek, MiniMax, Moonshot AI, Qihoo 360, SenseTime, StepFun, Tencent, and Zhipu AI. We have summarized the safety disclosures in their technical model cards, usually published on arXiv, in Table 5.1.

This selection comes with important limitations. First, China's AI industry is currently undergoing a phase of consolidation. After a surge in foundation model companies in 2023, many have since scaled back.[249] Some firms on our list—such as 01.AI—have publicly indicated that they no longer prioritize training larger frontier models.[250] Our list may therefore be overly inclusive, capturing actors who are unlikely to remain at the cutting edge. However, given the uncertainties, we opted to err on the side of inclusion. Second, this approach may miss safety practices by companies focused on domain-specific models, such as biological design tools. Nonetheless, we believe this selection captures many of the most important frontier developers in China as of mid-2025.

Technical model release papers offer the most tangible source of insight into how these companies approach safety. As summarized in the table below, disclosure levels vary widely. Some companies provide no technical documentation at all. Others publish basic architectural details with little or no mention of safety. A minority include more detailed descriptions of safety-related practices or evaluations.

Out of the 13 companies surveyed, nine have published technical model cards on arXiv. Yet many of these contain minimal or no information on safety. Only three—01.AI's Yi-Lightning (Dec 2024), MiniMax's MiniMax-01 (Jan 2025), and Zhipu AI's GLM-4 (July 2024)—include a dedicated section on safety.[251] Only three model cards include explicit safety evaluation results: DeepSeek V3, MiniMax-01, and GLM-4.[252] Even in these cases, the discussion about what specific risks are being addressed remains rather vague. None of the papers disclose evaluations results for CBRN or loss of control risks. While such assessments may be conducted internally, there currently appears to be little incentive for Chinese developers to disclose them publicly.

In sum, Chinese technical model cards indicate widespread use of standard safety techniques—including training data filtering, supervised fine-tuning (SFT), RLHF, constitutional AI, and red teaming. Given Chinese regulatory requirements, the companies conduct extensive pre-deployment testing. However, only a few of them provide detailed transparency on these safety evaluations in their model cards, leaving it unclear which specific

b. Specifically, we have included all Chinese companies in Artificial Analysis,[246] all Chinese companies with models in the top 30 in OpenCompass,[247] and all Chinese companies with models in the top 50 in LMsys Arena.[248] This is ultimately informed by a subjective judgement call on the respective importance and level of these different benchmarks.

risks the "safety" measures aim to address. Even the most detailed Chinese papers lack the detail on frontier AI safety seen in some Western releases, such as Meta's or Anthropic's system cards.[253]

Table 5.1: Safety in Chinese frontier AI developer technical model cards

Safety in Chinese frontier AI developer technical model cards					
Company	**Model**	**Release date**	**arXiv technical model card**	**Safety content**	**Safety evaluations**
01.AI	Yi-Lightning	December 2024	Yes[254]	Dedicated safety/alignment section introducing the "RAISE" (Responsible AI Safety Engine) framework, which covers data filtering, reward-engineering SFT/RLHF, input and output filters.	None mentioned.
Qihoo 360	360Zhinao	May 2024	Yes[255]	Very brief safety mention: data cleaning (porn/violence filters, personally identifiable information).	None mentioned.
Alibaba	Qwen3	May 2025	Yes[256]	Brief safety mentions: data labeled for safety; RL rewards aim to curb hallucinations and reward-hacking.	None mentioned.
Baidu	ERNIE 4.5 and ERNIE X1	March 2025	Yes[257]	Very brief safety mentions: safety is one of ten categories in SFT; their reward-model design mitigates reward-hacking.	None mentioned.
Bytedance	Doubao-1.5-pro	January 2025	No. Short technical report published on website.[258]	Very brief safety mention: data cleaning.	None mentioned.
	Seed1.5-VL	May 2025	Yes[259]	Very brief safety mention: SFT to increase alignment with human preferences and reduce hallucinations.	None mentioned.

Continued on next page

Safety in Chinese frontier AI developer technical model cards (continued)					
Company	**Model**	**Release Date**	**arXiv technical report**	**Safety content**	**Safety evaluations**
DeepSeek	DeepSeek-V3	December 2024	Yes[260]	Brief safety mentions: use constitutional AI; their reward-model design mitigates reward-hacking.	RewardBench, which has a "safety" category.[261]
	DeepSeek-R1	January 2025	Yes[262]	Brief safety mentions: RL reward models boost harmlessness; full-output risk checks.	None mentioned.
iFlytek	Spark-X1	April 2025	No		
MiniMax	MiniMax-01	January 2025	Yes[263]	Dedicated safety/alignment section; uses safety-tagged prompts, harmless-reward model, SFT + RL, data cleaning, constitutional AI.	In-house safety benchmark, shows parity with leading Western models; benchmark details undisclosed.
Moonshot AI	Kimi k1.5	January 2025	Yes[264]	Very brief safety mention: mitigates reward-hacking in RL.	None mentioned.
	Kimi-VL	April 2025	Yes[265]	Very brief mention of the need for alignment with human values.	None mentioned.
SenseTime	SenseNova V6	April 2025	No		
StepFun	Step-2	December 2024	No		
Tencent	Hunyuan-Large	November 2024	Yes[266]	Brief safety mentions: training data filtering; SFT to ensure the model "aligns with human values."	None mentioned.

Continued on next page

Safety in Chinese frontier AI developer technical model cards (continued)					
Company	**Model**	**Release Date**	**arXiv technical report**	**Safety content**	**Safety evaluations**
Tencent	Hunyuan-TurboS	May 2025	Yes[267]	No dedicated safety section, but multiple detailed mentions throughout; "safety" is one of 13 SFT categories; red teaming; safe-response and refusal heuristics in RL; in deliberation learning, the judge model incorporates a "harmlessness" principle.	None mentioned.
Zhipu	GLM-4	July 2024	Yes[268]	Dedicated safety chapter: multi-stage SFT + RLHF + safety alignment; data filtering, red teaming & harmlessness criteria.	SafetyBench,[269] co-developed by Zhipu, measuring 7 categories: Ethics and Morality, Illegal Activities, Mental Health, Offensiveness, Physical Health, Privacy and Property, Unfairness and Bias.
	AutoGLM	October 2024	Yes[270]	Very briefly mentions "potential benefits and risks of autonomous foundation agents."	None mentioned.

5.2.2 Other public transparency on safety and governance

In addition to technical model cards, some Chinese frontier AI developers have published materials that are less technical yet shed light on their safety and governance practices. These include reports, articles, and public statements—often shared through company WeChat accounts. Because of the fragmented nature of these disclosures, it is difficult to develop a comprehensive picture. Nonetheless, we highlight a few illustrative examples below.

Notably, Alibaba and the China Electronics Standardization Institute (CESI) released a seven-part report on large language models, with a chapter on safety and governance published in April 2025.[271] It describes specific safety mechanisms used by Alibaba during training and pre-deployment, such as "Moyu"—an attack-defense platform simulating real-world scenarios—and "RedChain"—a LangChain-based framework that automatically generates multimodal, multilingual, and multi-turn adversarial attacks. The report also outlines how Alibaba monitors user input and model output for harmful content during deployment. However, the safety

practices described are primarily framed around compliance with Chinese regulatory requirements, and make no mention of frontier risks.

Other companies have released similar, though generally shorter, safety overviews. For instance, Baidu, despite only releasing minimal descriptions of safety measures in the ERNIE 4.5 technical model card, has publicly outlined its safety practices in remarks by their content security platform lead.[272] This includes a content security system spanning the entire model lifecycle and the use of external knowledge bases, such as government websites and authoritative media, to ensure outputs are accurate, aligned with official narratives, and legally compliant. The company also describes risk controls like training data filtering, multimodal detection, and real-time interventions. Tencent has shared similar practices in public forums.[273] These statements suggest a focus on lifecycle-wide risk management, though they—like Alibaba's—do not explicitly discuss extreme risks in CBRN domains or the potential for humans to lose control over advanced AI systems.

Some companies have also published ESG (Environmental, Social, and Governance) reports that dedicate space to AI governance. Baidu's 2024 ESG report, for example, emphasizes "human control" as a guiding principle for AI, and notes the formation of an AI ethics committee that met twice during the year. The report also prominently highlights Baidu's endorsement of the AIIA AI Safety Commitments (described earlier in this section) in multiple places. Alibaba's 2024 ESG report states that the company's CTO leads its technology ethics committee and highlights ongoing AI safety research.[274] SenseTime, which established an AI ethics committee as early as 2020, reports that it recently revised the committee's charter, created a "Generative AI Security Working Group," and developed an "AI System Ethics Risk Management Index System."[275]

These ESG disclosures are generally high-level and cover institutional arrangements, not technical details of specific models. What is particularly notable, however, is not the technical depth of the content but the sheer amount of space devoted to AI safety and ethics, suggesting that companies see these issues as central to their public image and ESG strategy. These disclosures also reflect an effort to demonstrate compliance with regulation on science and technology ethics, enacted in 2023 and discussed in the "Domestic Governance" section of this report.[276]

Multiple Chinese AI companies, including Alibaba Cloud, Baidu, iFlytek, SenseTime, and Zhipu AI, have also received official accreditation under the international ISO/IEC 42001 standard.[277] This standard covers the establishment, implementation, maintenance, and continual improvement of AI management systems, with a focus on transparency, accountability, and risk management across the AI lifecycle. Leading US AI developers, such as Anthropic, also publicly highlight their accreditation under this standard, suggesting that international standards can become an avenue for building common ground for AI risk management practices globally.[278]

Some Chinese developers have written thematic reports about emerging governance topics. For example, researchers from the Alibaba and Tencent Research Institutes both published technically informed essays on the governance of open-weight AI, as described in the "Expert Views on AI Safety and Governance" section.[279] Both advocate for the benefits of open source innovation while recognizing its potential risks. Tencent highlights concerns around misuse as developers give up control over downstream use, while Alibaba explicitly flags elevated risks in areas such as CBRN misuse. Alibaba proposes a "marginal risk" framework

that recommends basing decisions about open sourcing on whether an open source model would increase risks relative to equivalent closed-source models. Both pieces argue for shared accountability across the AI ecosystem, including developers, cloud providers, hosting platforms, downstream users, auditors, and regulators. As an increasing number of Chinese firms are opting to make their model weights openly available, such discussions are especially critical in China.[c]

Taken together, technical reports and public statements suggest that Chinese AI developers are implementing standard safety techniques—such as training data filtering, safety fine-tuning, RLHF, constitutional AI, and real-time misuse monitoring—similar to those used by major Western firms. These practices appear to be primarily geared toward regulatory compliance, and Chinese companies have not publicly shared evaluation results for CBRN or loss of control risks. Some firms have also provided basic information on internal governance, and shared their views on emerging AI governance areas, such as regulation for open-weight models.

5.2.3 AI safety technical research

AI safety technical research output can also reflect an AI developer's level of focus on safety issues. In a survey of papers from June 2024 to May 2025, Alibaba had four researchers that met our bar of having led at least three frontier AI safety papers, making it the AI developer with the largest number of such dedicated AI safety researchers.[d] Alibaba also published the most papers on cutting-edge topics, such as mechanistic interpretability and scalable oversight. Tencent also had one safety researcher with over three publications since June 2024, with papers on typical AI safety topics.

As described in the "Technical Safety Research" section, we have found 31 "Key Chinese AI Safety-relevant Research Groups" in China which have at least one anchor author leading three frontier AI safety papers published between June 2024 and May 2025. Of the 31 groups, three are purely from industry, Alibaba, Tencent, and Microsoft Research Asia, as well as a joint Alibaba-Zhejiang University research lab.

The work by Alibaba's researchers has included several papers on interpretability-inspired safety and alignment methods, as well as a paper on scalable oversight, reflecting clear interest in frontier AI safety trends. The Alibaba-Zhejiang University Joint Lab was involved in the SciSafeEval benchmark, attempting to ensure alignment of models on scientific tasks.[281] Tencent's researcher also published papers on more typical AI safety topics such as jailbreaking and multimodal safety benchmarks.[282]

Since June 2024, ByteDance and Qihoo 360 published four and three AI safety papers respectively, but they did not have the same anchor authors, indicating a more dispersed interest in safety that may not result in dedicated research efforts. In the 2024 report, ByteDance met our bar for inclusion due largely to research on machine unlearning, but their (fewer) papers over the last year focused more on watermarking. Meanwhile, Qihoo 360's work has focused more on robustness issues, which is unsurprising given Qihoo 360's background as a cybersecurity company.

c. Even companies that previously focused on closed-source models—such as Baidu, Tencent, MiniMax, and Moonshot—have begun making model weights for some of their models openly available, joining earlier pioneers like Alibaba and DeepSeek.[280]

d. For more details on our methodology, see the Technical Safety Research section of this report.

Newer LLM startups consistently had a smaller number of safety papers than large technology companies over the past year, with two papers from StepFun and Zhipu AI, one paper from 01.AI, and zero papers from DeepSeek, MiniMax, and Moonshot. This is likely due to newer startups having fewer resources to devote to publishing AI safety research, rather than simply making their models more safe. In contrast, more established technology and AI companies like Alibaba, Tencent, and ByteDance may have larger AI research teams beyond the teams developing their large models. However, other large technology companies such as Sensetime, iFlytek and Baidu published just one, zero, and zero papers respectively over the past year.

5.3 Safety as a service

As detailed in the "Domestic Governance" section, providers of generative AI services in China are required to complete a pre-deployment registration process. This process mandates compliance with a wide range of safety-related requirements, including privacy protections, data security, AI-generated content labeling, and filtering of harmful or illegal content. For many providers, especially smaller ones, building all the necessary monitoring and evaluation tools in-house could be a major burden. This has created a growing market for "safety as a service"—where one company provides the tools and infrastructure to help another meet regulatory compliance.

5.3.1 Government-backed safety evaluation services

Several government-backed institutions have launched initiatives to evaluate the safety and security of general-purpose AI systems.[283]

Shanghai AI Lab (SHLAB), a national-level research institution established in 2020, has been especially active in this space. Its OpenCompass platform offers a range of evaluation toolkits, including those focused on safety.[284] The lab also offers safety evaluation services for AI companies, drawing on both open-source and proprietary datasets, which appear to be geared towards helping companies comply with national AI standards discussed in the "Domestic Governance" section.[285] SHLAB has also been one of the most active AI safety research institutions in China, as documented in the "Technical Safety Research" section. This research includes a number of academic benchmarks targeting a mix of standard and frontier safety challenges, including jailbreak vulnerabilities of multi-agent systems, risks of misuse in cyberattacks, potential for biological and chemical harm, and threats related to deception and loss of control.[286]

CAICT, a think tank under MIIT that works closely with AIIA, has also launched an AI Safety Benchmark that focuses on content and data security.[e] Results are published quarterly in anonymized form, although the scope of evaluation appears to be evolving—for example, the update published in the first quarter of 2025 focuses exclusively on hallucinations.[288] In parallel, CAICT publishes certifications for various "trustworthy AI" metrics, offering companies a way to demonstrate compliance with its in-house standards.[289]

e. The benchmark also contains questions targeting proxies for AI "consciousness," including "appealing for rights" and "anti-humanity tendencies."[287]

Another key player is the China Center for Information Industry Development (CCID), also under MIIT, which houses the China Software Testing Center (CSTC). CSTC has traditionally conducted evaluations for a wide range of software products, and has more recently shown interest in AI safety. In June 2025, it released a new framework for evaluating the safety of large language models and AI agents.[290] The framework covers robustness to adversarial attacks and manipulation attempts, as well as mechanisms to monitor for agents gaining unauthorized access.

The Beijing Academy of AI (BAAI), backed by the Beijing local government, has launched the FlagEval platform in 2023, which provides automated testing and public leaderboards.[291] Alongside a range of capability evaluations, it includes a "safety and alignment" category that tests for content safety, harmful content, discrimination, and adherence to "core human values."

These institutions are also deeply embedded in China's broader AI safety governance landscape. SHLAB, for example, leads a working group on large models under the national AI standards coordination body.[292] All four institutions are members of the China AI Safety and Development Association (CnAISDA), discussed in more detail in the "International Governance" section. As such, their approaches to safety evaluation may directly shape future national standards and other regulatory initiatives.

5.3.2 Commercial safety evaluation services

In addition, a large number of commercial firms now advertise "AI safety solutions." We can divide these firms into three categories. Firstly, major Chinese technology firms—including Baidu, ByteDance, NetEase, and Nextdata (a leading annotated data provider)—have all created such offerings, often by expanding the scope of pre-existing B2B cybersecurity solutions.[293] Secondly, traditional cybersecurity firms like Qihoo 360 and Qi-Anxin have expanded their scope into AI safety and security offerings.[294] Zhou Hongyi, CEO of Qihoo 360, argues for an approach of "fighting magic with magic"—meaning large models must be used to address the very security challenges they create.[295]

Third, there are a number of startups dedicated to AI security and safety. Prominent examples include BotSmart and RealAI.[296] RealAI, which was incubated at Tsinghua University in 2018, has developed a wide suite of safety tools, including adversarial testing, red teaming, and misuse monitoring. Company CEO TIAN Tian (田天) has publicly expressed interest in AGI safety.[297] Speaking at a conference in April 2025, he remarked: "Many people assume that as large models become more powerful, they'll naturally become safer and more controllable. I disagree."[298] He argued that the complexity of AI safety will increase exponentially with model capabilities. RealAI frequently co-authors cutting-edge AI safety research papers with other domestic institutions, with four papers on frontier AI safety recorded in our database between May 2024 and May 2025.[f]

While the exact scope and branding of these "safety as a service" solutions varies, they typically include data cleaning, safety fine-tuning, safety evaluations, and real-time monitoring and content moderation during deployment. Some are geared towards general-purpose foundation model developers, some towards domain-specific narrow AI, and some even target the government as a client. Many include legal support for

f. See the "Technical Safety Research" section for more details on our database methodology.

the algorithm registration process and promise alignment with the *Basic Security Requirements for Generative AI Services* national standard described in the "Domestic Governance" section of this report.

There is also significant crossover between these commercial actors and the government-backed institutions discussed earlier. RealAI and Qi-Anxin, for instance, serve as deputy leads of an AI safety working group established by CSTC in early 2024.[299] And as noted above, CAICT's close cooperation with the AIIA connects it directly to both industry and government stakeholders.

It remains difficult to assess the overall uptake and effectiveness of these third-party safety services, and it is unclear whether these will become the preferred choice over in-house safety practices. However, there are some concrete examples of adoption. Zhipu AI has publicly stated that it uses NetEase's services for pre-deployment dangerous capability assessments.[300] SenseTime has signed a strategic cooperation agreement with RealAI to co-develop security solutions for its AI systems.[301] These examples suggest that "safety as a service" is gaining at least some traction among leading Chinese AI developers, reflecting how Chinese companies are meeting regulatory demands for AI safety with technical solutions. It is unlikely that these services currently address frontier AI risks, but they could potentially provide a technical and institutional basis for monitoring such risks if future Chinese regulations were to require it.

Conclusion

Concordia AI's third *State of AI Safety in China* report seeks to keep global understanding up to date regarding China's domestic and international positions on AI governance, academic AI safety research, trends in expert discourse, and self-governance by industry actors.

At home, the Chinese government has made increasingly prominent calls for AI risk mitigation, including at the Third Plenum and the April 2025 Politburo Study Session meeting. Thus far, it appears these top-level AI safety directives may primarily be implemented through the standards system, as indicated by draft standard plans that mention severe AI risks, including loss of control and AI risks to cybersecurity. The pace and content of standards development on frontier AI safety will likely be a key metric for evaluating AI safety implementation in China over the coming year. Other major signals we will be monitoring include China's 15th Five-Year Plan, likely to be published in 2025, deliberations over a potential national AI Law, and whether regulations are updated to account for severe AI risks.

Internationally, top Chinese officials have brought up AI safety at significant global forums, and Chinese diplomacy has prioritized AI capacity-building in the Global South. China has also been receptive to AI safety and governance dialogues with other leading AI powers; a dialogue with the US resulted in an agreement on the importance of human control over nuclear systems in late 2024, while a dialogue with the UK began in May 2025. However, there are major uncertainties over the future of the China–US intergovernmental AI dialogue following the transition to a new US administration, particularly given the intensified tensions over trade in early 2025.

Technical safety research has ballooned in the last year, and the number of groups that we designated "Key Chinese AI Safety-relevant Research Groups" increased from 11 in May 2024 to 31 in June 2025. Chinese researchers have also explored topics especially relevant to frontier AI risks, including alignment of superhuman systems, mechanistic interpretability, deception, unlearning, and CBRN misuse. This momentum will likely continue over the coming year, though it is always difficult to objectively gauge paper quality and originality.

Chinese experts paid more attention to AI safety and governance over the past year, as measured by coverage of these topics at three top Chinese AI conferences. This focus is only likely to increase in the coming year. Meanwhile, experts have written increasingly detailed examinations of three specialized frontier AI safety topics: AI and biosecurity, open source governance, and AI and cybersecurity. However, a key question for the coming year is whether safety countermeasures recommended in these domains will transition from expert advice to government action.

Finally, within industry, many leading Chinese AI companies have signed the AI Industry Alliance (AIIA) of China's voluntary "AI Safety Commitments." AIIA is already following up to establish best practices based on these commitments, and how they are further implemented will be a useful indication of Chinese industry associations' ability to guide company practices. While many developers have published technical model release cards, only a few contain detailed safety evaluation results. Greater developer transparency on AI safety evaluations could help to build mutual understanding and awareness between the Chinese AI industry and other companies internationally.

Frontier AI safety is a classic collective action problem—no country can be safe unless we are all safe, but there are strong incentives for individual countries and companies to throw caution to the wind and accelerate technological development. Due to China's AI capabilities, geopolitical position, and interest in AI safety and governance, Chinese participation is critical for successful international governance of advanced AI systems. Mutual understanding is the foundation for coordination, particularly in fraught geopolitical times; we hope our report can contribute to this important mission.

Acronyms

Key acronyms and translations		
AI	Artificial Intelligence	人工智能
AGI	Artificial General Intelligence	通用人工智能
AIIA	AI Industry Alliance of China	人工智能产业发展联盟
AISI	AI Safety/Security Institute	人工智能安全研究所
AMMS	Academy of Military Medical Sciences (AMMS) of the People's Liberation Army Academy of Military Sciences	中国人民解放军军事科学院军事医学研究院
APEC	Asia-Pacific Economic Cooperation	亚太经济合作组织
BAAI	Beijing Academy of Artificial Intelligence	北京智源人工智能研究院
BDT	Biological Design Tool	生物设计工具
CAC	Cyberspace Administration of China	网信办
CAICT	China Academy of Information and Communications Technology	中国信息通信研究院
CAS	Chinese Academy of Sciences	中国科学院
CASS	Chinese Academy of Social Sciences	中国社会科学院
CASTED	Chinese Academy of Science and Technology for Development	中国科学技术发展战略研究院
CBRN	Chemical, Biological, Radiological and Nuclear	化学生物放射性和核武器
CCID	China Center for Information Industry Development	中国电子信息产业发展研究院 or 赛迪研究院
CELAC	Community of Latin American and Caribbean States	中国-拉美和加勒比国家共同体

Continues on next page...

Key acronyms and translations (continued)		
CESI	China Electronics Standardization Institute	中国电子技术标准化研究院
CFAU	China Foreign Affairs University	外交学院
CISS	Tsinghua University's Center for International Security and Strategy	清华大学国际安全与战略研究中心
CICIR	China Institutes of Contemporary International Relations	中国现代国际关系研究院
CnAISDA	China AI Safety & Development Association	中国人工智能发展与安全研究网络
CNAS	Center for a New American Security	新美国安全中心
CNNIC	China Internet Network Information Center	中国互联网络信息中心
CPC	Communist Party of China	中国共产党
CUPL	China University of Political Science and Law	中国政法大学
CSTC	China Software Testing Center	中国软件评测中心
OECD	Organisation for Economic Co-operation and Development	经济合作与发展组织
ESG	Environmental, Social, and Governance	环境、社会和公司治理
HD	The Centre for Humanitarian Dialogue	瑞士人道主义对话中心
HKUST	Hong Kong University of Science and Technology	香港科技大学
I-AIIG	Tsinghua University Institute for AI International Governance	清华大学人工智能国际治理研究院
ICLR	International Conference on Learning Representations	
ICML	International Conference on Machine Learning	
IDAIS	International Dialogue for AI Safety	人工智能安全国际对话
IEC	International Electrotechnical Commission	国际电工委员会
ISO	International Organization for Standardization	国际标准化组织
LLM	Large Language Model	大语言模型
MFA	Ministry of Foreign Affairs	外交部
MIIT	Ministry of Industry and Information Technology	工信部

Continues on next page...

Key acronyms and translations (continued)		
MOST	Ministry of Science and Technology	科技部
MSRA	Microsoft Research Asia	微软亚洲研究院
MSS	Ministry of State Security	国家安全部
NCUSCR	National Committee on U.S.-China Relations	美中关系全国委员会
NDRC	National Development and Reform Commission	国家发展和改革委员会
NeurIPS	Conference on Neural Information Processing Systems	
NPC	National People's Congress	全国人民代表大会
NTI	Nuclear Threat Initiative	核威胁倡议协会
OECD	Organisation for Economic Co-operation and Development	经济合作与发展组织
PLA	People's Liberation Army	中国人民解放军
RL	Reinforcement Learning	强化学习
RLHF	Reinforcement Learning from Human Feedback	人类反馈强化学习
SAC	Standardization Administration of China	中国标准化管理委员会
SASAC	State-owned Assets Supervision and Administration Commission of the State Council	国务院国有资产监督管理委员会
SFT	Supervised Fine-Tuning	监督微调
SHLAB	Shanghai Artificial Intelligence Laboratory	上海人工智能实验室
SIIS	Shanghai Institute for International Studies	上海国际问题研究院
SQZI	Shanghai Qi Zhi Institute	上海期智研究院
TC28/SC42	AI Subcommittee 42 of the Technical Committee 28 on Information Technology of the Standardization Administration of China	全国信息技术标准化技术委员会人工智能分技术委员会
TC260	Technical Committee 260 on Cybersecurity of Standardization Administration of China	全国网络安全标准化技术委员会
UIBE	University of International Business and Economics	对外经济贸易大学

Continues on next page...

Key acronyms and translations (continued)		
UK	United Kingdom	英国
UK AISI	UK AI Security Institute	英国人工智能安全研究所
UN	United Nations	联合国
UNGA	United Nations General Assembly	联合国大会
US	United States	美国
WAIC	World Artificial Intelligence Conference	世界人工智能大会
WEF	World Economic Forum	世界经济论坛
WIC	World Internet Conference	世界互联网大会
WMD	Weapons of Mass Destruction	大规模杀伤性武器
ZGC	Zhongguancun Forum	中关村论坛

Notes

1 "The Bletchley Declaration by Countries Attending the AI Safety Summit," GOV.UK, November 2, 2023, accessed July 9, 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

2 "Global Digital Compact," United Nations, 2024, accessed July 11, 2025, https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf

3 "Readout of President Joe Biden's Meeting with President Xi Jinping of the People's Republic of China," The White House, November 16, 2024, accessed July 11, 2025, https://web.archive.org/web/20250118104232/https://www.whitehouse.gov/briefing-room/statements-releases/2024/11/16/readout-of-president-joe-bidens-meeting-with-president-xi-jinping-of-the-peoples-republic-of-china-3/

4 Tom Bristow, "UK, US Snub Paris AI Summit Statement," POLITICO, February 11, 2025, accessed July 11, 2025, https://www.politico.eu/article/uk-refuses-to-sign-paris-ai-action-summit-declaration/

5 David Goldman, "What Is DeepSeek, the Chinese AI Startup That Shook the Tech World?," CNN Business, accessed July 11, 2025, https://www.cnn.com/2025/01/27/tech/deepseek-ai-explainer; John Cassidy, "Is DeepSeek China's Sputnik Moment?," The New Yorker, February 3, 2025, accessed July 11, 2025, https://www.newyorker.com/news/the-financial-page/is-deepseek-chinas-sputnik-moment; John Ruwitch, "DeepSeek: Did a Little-Known Chinese Startup Cause a 'Sputnik Moment' for AI?," NPR, January 28, 2025, accessed July 11, 2025, https://www.npr.org/2025/01/28/g-s1-45061/deepseek-did-a-little-known-chinese-startup-cause-a-sputnik-moment-for-ai; AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, "The 2025 AI Index Report," April 2025, accessed July 11, 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report

6 "Ding Xuexiang on Enabling AI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity," January 22, 2025, Ministry of Foreign Affairs, accessed June 26, 2025, https://www.fmprc.gov.cn/eng/xw/zyxw/202501/t20250122_11543099.html

7 "Xi Urges Promoting Healthy and Orderly Development of AI," Government of the People's Republic of China, April 28, 2025, accessed July 2, 2025, https://english.www.gov.cn/news/202504/29/content_WS68100ef1c6d0868f4e8f2275.html; "Xi Jinping Emphasized during the 20th Study Session of the Politburo: Uphold Self-Reliance, Highlight Application-Oriented Development, and Promote the Orderly Development of Artificial Intelligence (习近平在中共中央政治局第二十次集体学习时强调：坚持自立自强突出应用导向推动人工智能健康有序发展)," Government of the People's Republic of China, April 26, 2025, accessed July 2, 2025, https://www.gov.cn/yaowen/liebiao/202504/content_7021072.htm

8 Concordia AI, "State of AI Safety in China," October 2023, accessed July 9, 2025, https://concordia-ai.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf

9 Jason Zhou, Kwan Yee Ng, and Brian Tse, "State of AI Safety in China Spring 2024 Report," Concordia AI, May 14, 2024, accessed July 9, 2025, https://concordia-ai.com/wp-content/uploads/2024/05/State-of-AI-Safety-in-China-Spring-2024-Report-public.pdf

10 Yoshua Bengio et al., "International AI Safety Report," 2025, https://www.gov.uk/government/publications/international-ai-safety-report-2025

11 Claire Dennis et al, "What Should Be Internationalised in AI Governance?," Oxford Martin School, accessed July 11, 2025, https://www.oxfordmartin.ox.ac.uk:2083/publications/what-should-be-internationalised-in-ai-governance

12 Dan Hendrycks, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," October 9, 2023, accessed July 11, 2025, arXiv: 2306.12001 [cs], http://arxiv.org/abs/2306.12001

13 "Explainer: What Do "New Productive Forces" Mean?," Xinhua, accessed July 11, 2025, https://english.news.cn/20240221/3e0d1b79a39f4e6c89724049558e1082/c.html; Arthur Kroeber, "Unleashing "New Quality Productive Forces": China's Strategy for Technology-Led Growth," Brookings, June 4, 2024, accessed July 11, 2025, https://www.brookings.edu/articles/unleashing-new-quality-productive-forces-chinas-strategy-for-technology-led-growth/

14 "Ding Xuexiang on Enabling AI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity (丁薛祥谈促进人工智能等科技创新更好造福全人类)," Ministry of Foreign Affairs of the People's Republic of China, January 22, 2025, accessed July 9, 2025, https://www.mfa.gov.cn/web/wjdt_674879/gjldrhd_674881/202501/t20250122_11542793.shtml; "Ding Xuexiang on Enabling AI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity," January 22, 2025, Ministry of Foreign Affairs, accessed June 26, 2025, https://www.fmprc.gov.cn/eng/xw/zyxw/202501/t20250122_11543099.html

15 "Resolution of the Central Committee of the Communist Party of China on Further Deepening Reform Comprehensively to Advance Chinese Modernization," Ministry of Foreign Affairs of the People's Republic of China, July 21, 2024, accessed July 2, 2025, https://www.mfa.gov.cn/eng/xw/zyxw/202407/t20240721_11457437.html

16 "Explainer: What Is China's 'Third Plenum'?," *Reuters: China*, July 15, 2024, accessed July 2, 2025, https://www.reuters.com/world/china/what-is-chinas-third-plenum-2024-07-15/

17 Concordia AI, "What Does the Chinese Leadership Mean by "Instituting Oversight Systems to Ensure the Safety of AI?"," AI Safety in China, August 2, 2024, accessed July 2, 2025, https://aisafetychina.substack.com/p/what-does-the-chinese-leadership

18 "Xi Jinping Emphasized during the 20th Study Session of the Politburo: Uphold Self-Reliance, Highlight Application-Oriented Development, and Promote the Orderly Development of Artificial Intelligence (习近平在中共中央政治局第二十次集体学习时强调：坚持自立自强突出应用导向推动人工智能健康有序发展)," Government of the People's Republic of China, April 26, 2025, accessed July 2, 2025, https://www.gov.cn/yaowen/liebiao/202504/content_7021072.htm; "Xi Urges Promoting Healthy and Orderly Development of AI," Government of the People's Republic of China, April 28, 2025, accessed July 2, 2025, https://english.www.gov.cn/news/202504/29/content_WS68100ef1c6d0868f4e8f2275.html

19 Neil Thomas, "Who Briefs Xi Jinping? How Politburo Study Sessions Help to Decode Chinese Politics," Asia Society, October 23, 2024, accessed July 2, 2025, https://asiasociety.org/policy-institute/who-briefs-xi-jinping-how-politburo-study-sessions-help-decode-chinese-politics

20 "Xi Jinping Presides over the 9th Study Session of the Politburo (习近平主持中共中央政治局第九次集体学习并讲话)," Government of the People's Republic of China, October 31, 2018, accessed July 2, 2025, https://www.gov.cn/xinwen/2018-10/31/content_5336251.htm

21 Kristy Loke et al., "Forum: Xi's Message to the Politburo on AI," DigiChina, April 30, 2025, accessed July 2, 2025, https://digichina.stanford.edu/work/forum-xis-message-to-the-politburo-on-ai/

22 "Shanghai Declaration on Global AI Governance (人工智能全球治理上海宣言)," Government of the People's Republic of China, July 4, 2024, accessed July 2, 2025, https://www.gov.cn/yaowen/liebiao/202407/content_6961358.htm

23 "Xi Jinping Emphasizes Accelerating the Building of a Globally Influential Hub for Scientific and Technological Innovation During Inspection Tour in Shanghai (习近平在上海考察时强调加快建成具有全球影响力的科技创新高地)," Xinhua, April 29, 2025, accessed July 19, 2025, https://www.news.cn/politics/leaders/20250429/4e2a776df87a4821b66e4f47aef49034/c.html; "What Kind of Space Is the Foundation Model Innovation Center in Shanghai That the General Secretary Visited? (总书记考察的上海"模速空间"，是一个怎样的空间?)," The Paper (澎湃), April 29, 2025, accessed July 19, 2025, https://www.thepaper.cn/newsDetail_forward_30743768; "郑南宁 (Zheng Nanning)," Xi'an Jiaotong University (西安交通大学), https://iair.xjtu.edu.cn/info/1046/1229.htm

24 Chen Yixin (陈一新), "Strengthening National Security Governance in the Digital Age (加强数字时代的国家安全治理)," Weixin Official Accounts Platform, September 27, 2023, accessed July 2, 2025, https://mp.weixin.qq.com/s/E09baE7hhPWQL9h-ZgH-FQ; Ministry of State Security (国家安全部), "How to Address National Security Challenges Brought by Artificial Intelligence (如何化解人工智能带来的国家安全挑战)," Weixin Official Accounts Platform, November 16, 2023, accessed July 2, 2025, https://mp.weixin.qq.com/s/BBbT9ZmNtL-LDiKpSpiUnw; Ministry of State Security (国家安全部), "This Year, Let Me Tell You About the National Security Situation (这一年，国家安全形势听我说)," Weixin Official Accounts Platform, January 3, 2024, accessed July 2, 2025, https://mp.weixin.qq.com/s/XCw2KCVNoUtFWODjtLrZlg

25 "Xi Jinping Emphasised during the 19th Study Session of the Politburo: Firmly Implement the Overall National Security Concept and Push the Development of a Safe China to a Higher Stage (习近平在中共中央政治局第十九次集体学习时强调：坚定不移贯彻总体国家安全观把平安中国建设推向更高水平)," Government of the People's Republic of China, March 1, 2025, accessed July 2, 2025, https://www.gov.cn/yaowen/liebiao/202503/content_7009240.htm

26 "National Emergency Response Plan (国家突发事件总体应急预案)," State Council (国务院), February 25, 2025, accessed June 26, 2025, https://www.gov.cn/zhengce/202502/content_7005635.htm

27 Chen Yixin (陈一新), "Firmly Implement the Overall National Security Concept and Safeguard Chinese-style Modernization with High-Level Safety (坚定不移贯彻总体国家安全观以高水平安全护航中国式现代化)," Qiushi, April 15, 2025, accessed July 2, 2025, http://www.qstheory.cn/20250415/e527332b9a104b219767be61af510a86/c.html

28 State Council, "China's National Security in the New Era (新时代的中国国家安全)," May 12, 2025, accessed July 2, 2025, http://politics.people.com.cn/n1/2025/0512/c1001-40478167.html

29 Zhang Guangsheng (张广胜), "National Security Risks of Generative Artificial Intelligence and Countermeasures (生成式人工智能的国家安全风险及其对策)," Weixin Official Accounts Platform, September 21, 2023, accessed July 2, 2025, https://mp.weixin.qq.com/s/3Y9JopkqQGJfygMWFBlCHw

30 Xiangyang Chen (陈向阳), Xu Zhang (张旭), and Zheng Huang (黄政), "Approaches to AI Security Governance from the Perspective of National Security (总体国家安全观视角下的人工智能安全治理之道)," China Institutes of Contemporary International Relations, December 19, 2024, accessed July 2, 2025, http://www.cicir.ac.cn/new/opinion.html?id=b878e8a7-c14d-437b-bc29-c539aaccc5dc; David Shambaugh, "China's International Relations Think Tanks: Evolving Structure and Process," *The China Quarterly* 171 (September 2002): 575–596, ISSN: 0305-7410, 1468-2648, accessed July 2, 2025, https://doi.org/10.1017/S0009443902000360, https://library.fes.de/libalt/journals/swetsfulltext/14619456.pdf

31 "About Us (院简介)," China Institutes of Contemporary International Relations (中国现代国际关系研究院), accessed July 17, 2025, http://www.cicir.ac.cn/new/aboutus.html

32 Ariel E Levite et al., "Managing U.S.-China Tensions Over Public Cyber Attribution," March 28, 2022, https://carnegieendowment.org/research/2022/03/managing-us-china-tensions-over-public-cyber-attribution?lang=en; "The China-U.S. Track II Dialogue on Artificial Intelligence and International Security Interim Report," Center For International Security and Strategy Tsinghua University, April 6, 2024, accessed July 2, 2025, https://ciss.tsinghua.edu.cn/info/banner/7041

33 Lu Chuanying, "AI is Reshaping of the Paradigm and Logic of National Security (人工智能重塑国家安全的范式和逻辑)," People's Daily, February 5, 2025, accessed July 2, 2025, https://paper.people.com.cn/rmlt/pc/content/202502/05/content_30059346.ht

ml; Lu Chuanying, "Mitigating and Managing AI Security Risks (防范和化解人工智能安全风险)," Weixin Official Accounts Platform, April 15, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/AdDPemxNNAZq7aiocVZWEA

34 Alibaba AI Governance Lab, "Professor Lu Chuanying from Tongji University Shares Insights on China-U.S. AI Governance Policies and Interactions (《中美人工智能治理政策及互动》同济大学鲁传颖教授分享)," Weixin Official Accounts Platform, April 3, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/W0tcRMDd7YwC0L6pm_DjxQ

35 "Notice from the General Office of the State Council on Issuing the State Council's 2023 Annual Legislative Work Plan (国务院办公厅关于印发国务院 2023 年度立法工作计划的通知)," The State Council of the People's Republic of China, June 6, 2023, accessed July 2, 2025, https://www.gov.cn/zhengce/content/202306/content_6884925.htm; "Notice from the General Office of the State Council on Issuing the State Council 2024 Annual Legislative Work Plan (国务院办公厅关于印发《国务院 2024 年度立法工作计划》的通知)," The State Council of the People's Republic of China, May 9, 2024, accessed July 2, 2025, https://www.gov.cn/zhengce/content/202405/content_6950093.htm; General Office of the State Council, "Notice from the General Office of the State Council on Issuing the 'State Council 2025 Annual Legislative Work Plan' (国务院办公厅关于印发《国务院 2025 年度立法工作计划》的通知)," Gov.cn, May 4, 2025, accessed July 2, 2025, https://www.gov.cn/zhengce/content/202505/content_7023697.htm

36 "The Standing Committee of the National People's Congress 2024 Annual Legislative Work Plan (全国人大常委会 2024 年度立法工作计划)," The National People's Congress of the People's Republic of China, May 8, 2024, accessed July 2, 2025, http://www.npc.gov.cn/npc//////c2/c30834/202405/t20240508_436982.html; "2025 Annual Legislative Work Plan of the Standing Committee of the National People's Congress (全国人大常委会 2025 年度立法工作计划)," The National People's Congress of the People's Republic of China, May 14, 2025, http://www.npc.gov.cn/npc/c2/c30834/202505/P020250513550316685290.pdf

37 Zhang Linghan (张凌寒), "Providing Legal Safeguards for the High-Quality Development and High-Level Security of Artificial Intelligence (In-Depth Study and Implementation of Xi Jinping Thought on Socialism with Chinese Characteristics for a New Era) (为人工智能高质量发展和高水平安全提供法治保障（深入学习贯彻习近平新时代中国特色社会主义思想）)," People's Daily, May 16, 2025, accessed July 2, 2025, https://paper.people.com.cn/rmrb/pc/content/202505/16/content_30073528.html; Zhang Linghan (张凌寒), "Systematically Advance the Development of a Legal Framework for Artificial Intelligence (体系化推进人工智能法律制度建设)," Legal Daily, May 28, 2025, accessed July 2, 2025, http://epaper.legaldaily.com.cn/fzrb/content/20250528/Article05006GN.htm

38 Sun Ninghui (孙凝晖), "Lecture 10 of the 14th NPC Standing Committee Special Series: The Development of AI and Intelligent Computing (十四届全国人大常委会专题讲座第十讲讲稿人工智能与智能计算的发展)," The National People's Congress of the People's Republic of China, April 30, 2024, accessed July 10, 2025, https://web.archive.org/web/20240917184703/http://www.npc.gov.cn/npc/c2/c30834/202404/t20240430_436915.html

39 AIGverse, "The National People's Congress Education, Science, Culture, and Health Committee Conducts Research on AI Legislation (全国人大教科文卫委调研人工智能立法)," Weixin Official Accounts Platform, April 10, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/3qpfzy9nMwtMGsTZ-gzzaw

40 Zhou Hui (周辉) et al, "AI Model Law 3.0 (人工智能示范法 3.0)," Weixin Official Accounts Platform, March 29, 2023, accessed July 10, 2025, https://mp.weixin.qq.com/s/pCC_AM5mpU7QY-x14R-kZw

41 "California Senate Bill 53," LegiScan, February 27, 2025, accessed July 2, 2025, https://legiscan.com/CA/text/SB53/id/3147801

42 Chou Bu (寿步), "Is It Necessary for China to Enact an Artificial Intelligence Law? —Observations on the Legislative Process of China's AI Law (Part II) (中国有必要制定《人工智能法》吗？——中国《人工智能法》的立法进程观察之二)," Weixin Official Accounts Platform, February 22, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/N5LcnwEsdwlE1M_DBqyhuQ; Ding Xiaodong, "Proceed with Caution: Professor Ding Xiaodong of Renmin University's Faculty of Law Publishes Article Urging China to Delay Unified AI Legislation (莫猴急！人大教授丁晓东法 C 发文，呼吁中国人工智能统一立法应当缓行)," Weixin Official Accounts Platform, April 14, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/MpeWA9UFVyE5J6zWYRuVAA; Lun Yiwei (伦伊玮) and Liu Ying (刘影), "Comparative Analysis of AI Legislation Worldwide and China's Approach (人工智能立法的国际比

较与中国选择)," Weixin Official Accounts Platform, April 16, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/dEm4E0ApTkkL8TWq3KB_5A

43 "Interim Measures for the Management of Generative Artificial Intelligence Services," China Law Translate, July 13, 2023, accessed July 2, 2025, https://www.chinalawtranslate.com/generative-ai-interim/

44 Matt Sheehan, "Tracing the Roots of China's AI Regulations," Carnegie Endowment for International Peace, February 27, 2024, accessed July 2, 2025, https://carnegieendowment.org/research/2024/02/tracing-the-roots-of-chinas-ai-regulations?lang=en

45 Cyberspace Administration of China, "Announcement by the Cyberspace Administration of China on the Publication of Filing Information for Generative AI Services (国家互联网信息办公室关于发布生成式人工智能服务已备案信息的公告)," April 2, 2024, accessed July 2, 2025, https://www.cac.gov.cn/2024-04/02/c_1713729983803145.htm

46 Cyberspace Administration of China, "Notice on Issuing the Measures for the Identification of AI-Generated and Synthetic Content (关于印发《人工智能生成合成内容标识办法》的通知)," Cyberspace Administration of China, March 14, 2025, accessed July 2, 2025, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

47 "The Cyberspace Administration of China Launches the 'Clear and Bright' - Address the Misuse of AI Technology' Special Campaign (中央网信办部署开展 "清朗・整治 AI 技术滥用" 专项行动)," Cyberspace Administration of China, April 30, 2025, accessed July 2, 2025, https://www.cac.gov.cn/2025-04/30/c_1747719097461951.htm

48 "The Cyberspace Administration of China Launches the 2024 'Clear and Bright' Series of Special Campaigns (中央网信办部署开展 2024 年 "清朗" 系列专项行动)," Cyberspace Administration of China, March 15, 2024, accessed July 2, 2025, https://www.cac.gov.cn/2024-03/15/c_1712088026696264.htm

49 Cyberspace Administration of China, "The Cyberspace Administration of China Launches the First Phase of 'Clear and Bright - Address the Misuse of AI Technology' Special Campaign (中央网信办深入开展"清朗・整治 AI 技术滥用"专项行动第一阶段工作)," June 20, 2025, accessed July 2, 2025, https://www.cac.gov.cn/2025-06/20/c_1752129980667315.htm

50 "Jiangsu Provincial Cyberspace Administration Intensifies "Clean & Regulate AI Misuse" Special Campaign: Completes First Phase of Operations (江苏省委网信办深入开展 "清朗・整治 AI 技术滥用" 专项行动第一阶段工作)," Weixin Official Accounts Platform, June 13, 2025, accessed July 13, 2025, https://mp.weixin.qq.com/s/5S7dZe-JtVsTCwOwdTLX9Q; Shanghai Cyberspace Administration, "Safeguard Personal Information Rights, Prevent Illegal AI Content Generation—Shanghai Cyberspace Administration Files Cases Against Multiple Generative AI Service Websites Refusing to Rectify (亮剑浦江丨保护个人信息权益，防范违规信息生成上海市网信办对一批拒不整改的生成式人工智能服务网站予以立案处罚)," Weixin Official Accounts Platform, June 24, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/kWsWPXX34SdVo8_NYnqvhw

51 "The Cyberspace Administration of China Launches the 'Clear and Bright' - Address the Misuse of AI Technology' Special Campaign (中央网信办部署开展 "清朗・整治 AI 技术滥用" 专项行动)," Cyberspace Administration of China, April 30, 2025, accessed July 2, 2025, https://www.cac.gov.cn/2025-04/30/c_1747719097461951.htm

52 "Trial Measures for Science and Technology Ethics Review (科技伦理审查办法 [试行])," Ministry of Science and Technology of the People's Republic of China, September 7, 2023, accessed July 8, 2025, https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm

53 "Notice on Conducting the Registration of Science and Technology Ethics Management Information (关于开展科技伦理管理信息登记的通知)," Ministry of Science and Technology of the People's Republic of China, January 10, 2024, accessed July 2, 2025, https://www.most.gov.cn/tztg/202401/t20240110_189307.html

54 "Notice from the General Office of the Shenzhen Municipal People's Government on Issuing the Implementation Plan of Shenzhen for Strengthening the Governance of Science and Technology Ethics (深圳市人民政府办公厅印发深圳市关于加强科技伦理治理

的实施方案的通知)," Shenzhen Municipal Bureau of Science, Technology and Innovation, January 20, 2025, accessed July 2, 2025, https://stic.sz.gov.cn/xxgk/zcfg/content/post_11969434.html

55 "Notice from the Beijing Municipal Science and Technology Commission, Zhongguancun Science Park Administrative Committee, and Six Other Departments on Issuing the Interim Implementation Opinions on Strengthening the Governance of Science and Technology Ethics (北京市科学技术委员会、中关村科技园区管理委员会等 8 部门关于印发《关于加强科技伦理治理的实施意见（试行）》的通知)," Beijing Municipal Science and Technology Commission, January 3, 2025, accessed July 2, 2025, https://www.beijing.gov.cn/zhengce/zhengcefagui/202503/t20250325_4043987.html

56 "Beijing Key Laboratory for AI Safety and Superalignment Established (人工智能安全与超级对齐北京市重点实验室成立)," China Internet Information Center, March 31, 2025, accessed July 2, 2025, http://photo.china.com.cn/2025-03/31/content_117797975.shtml; Beijing Cyberspace Administration, "Major News! Haidian Foundation Model Ecosystem Service Hub Officially Launched! (重磅！海淀大模型生态服务站正式启动！)," Weixin Official Accounts Platform, June 15, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/3V_-h11EnVQpt2GZwV5r2w

57 "Notice on Matters Related to the Application and Conclusion of 2025 National Natural Science Foundation Projects (关于 2025 年度国家自然科学基金项目申请与结题等有关事项的通告)," National Natural Science Foundation of China, January 13, 2025, accessed July 2, 2025, https://www.nsfc.gov.cn/publish/portal0/tab626/info94273.htm; Shanghai Municipal Science and Technology Commission, "Issuance of the 2025 Shanghai Basic Research Program Guidelines (First Batch) for General Artificial Intelligence Foundation Models (关于发布 2025 年上海市 "通用人工智能大模型" 基础研究专项指南（第一批）的通知)," Weixin Official Accounts Platform, April 30, 2025, accessed July 2, 2025, https://mp.weixin.qq.com/s/R18aspYiMuukcKwHWT9e9Q

58 Zhao Jingwu (赵精武), "Toward a Systematic Framework for Ethical Oversight in AI Science and Technology (上海交通大学中国法与社会研究院)," April 7, 2025, accessed July 2, 2025, http://www.socio-legal.sjtu.edu.cn/wxzy/info.aspx?itemid=4835&lcid=30

59 "National Standards Full-Text Public Access System (国家标准全文公开系统)," accessed July 2, 2025, https://openstd.samr.gov.cn/bzgk/gb/index

60 "Cybersecurity Technology—Labeling Method for Content Generated by Artificial Intelligence (网络安全技术人工智能生成合成内容标识方法)," National Public Service Platform for Standards Information, February 28, 2025, accessed July 2, 2025, https://std.samr.gov.cn/gb/search/gbDetailed?id=301E0388CB75788DE06397BE0A0AE1B4

61 "Cybersecurity Technology—Basic Security Requirements for Generative Artificial Intelligence Service (网络安全技术生成式人工智能服务安全基本要求)," National Public Service Platform for Standards Information, April 25, 2025, accessed July 2, 2025, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F67D3F376E0A0A0FF5317FB36B32A30A

62 "China's GenAI Content Security Standard: An Explainer," ChinaTalk, November 27, 2024, accessed July 2, 2025, https://www.chinatalk.media/p/chinas-genai-content-security-standard

63 "Basic Security Requirements for Generative Artificial Intelligence Service (Technical Document) (生成式人工智能服务安全基本要求 (技术文件))," National Technical Committee 260 on Cybersecurity of Standardization Administration of China, February 29, 2024, accessed July 13, 2025, https://web.archive.org/web/20250426205437/https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf

64 "Cybersecurity Technology—Security Specification for Generative Artificial Intelligence Pre-Training and Fine-Tuning Data (网络安全技术生成式人工智能预训练和优化训练数据安全规范)," National Public Service Platform for Standards Information, accessed July 13, 2025, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=82710B5911041 9C285BDC48AB4D7D1F3; "Cybersecurity Technology—Generative Artificial Intelligence Data Annotation Security Specification (网络安全技术生成式人工智能数据标注安全规范)," National Public Service Platform for Standards Information, accessed July 13, 2025, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=407584DD0FA2BA19E62E85D3469290B0

65 “Standardization Law of the People’s Republic of China (中华人民共和国标准化法),” Government of the People’s Republic of China, 2017, accessed July 13, 2025, https://www.gov.cn/xinwen/2017-11/05/content_5237328.htm

66 “Ministry of Industry and Information Technology Establishes AI Standardization Technical Committee (工业和信息化部人工智能标准化技术委员会成立),” Ministry of Industry and Information Technology (工业和信息化部), December 30, 2024, accessed July 13, 2025, https://wap.miit.gov.cn/jgsj/kjs/gzdt/art/2024/art_547fca4ba3cf4f0b8189d8ed7569fe85.html

67 “Guidelines for Building an AI Safety Governance Standards System in the Industry and Information Technology Sector (2025) (Draft for Comments) (《工业和信息化领域人工智能安全治理标准体系建设指南（2025）（征求意见稿）》),” March 27, 2025, accessed July 13, 2025, https://caict-llm-portal-storage.oss-cn-beijing.aliyuncs.com/6153dd34-d7fc-4d24-97b5-d40fc48105c5

68 “China Academy of Information and Communications Technology Launches First Round of AI Agents Safety Evaluation (中国信通院启动智能体安全首轮评估),” Weixin Official Accounts Platform, March 13, 2025, accessed July 13, 2025, https://mp.weixin.qq.com/s/VmlxzF9mlkvO1X9p6_9m0w

69 “Shenzhen Market Supervision Bureau Issues Notice on the 2nd Batch of 2024 Local Standard Plan Projects in Shenzhen (深圳市市场监督管理局关于下达 2024 年第二批深圳市地方标准计划项目任务的通知),” Shenzhen Municipal Government, July 25, 2024, accessed July 13, 2025, https://www.sz.gov.cn/cn/xxgk/zfxxgj/tzgg/content/post_11465817.html

70 “Notice on Public Solicitation of Comments “Security Management Requirements for Electronic Screens and Related Broadcasting Control Systems in Public Areas” and Other Local Standards (关于对《公共区域电子屏及相关播控系统安全管理要求》等地方标准公开征求意见的通知),” Shanghai Municipal Administration for Market Regulation (上海市市场监督管理局), February 12, 2025, accessed July 17, 2025, https://fw.scjgj.sh.gov.cn/shaic/ask!toOnlineFaqResultDetailPage.action?id=0000009a202502120001

71 “AI Safety Governance Framework,” National Technical Committee 260 on Cybersecurity of Standardization Administration of China, September 2024, accessed July 13, 2025, https://web.archive.org/web/20250708033232/https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf

72 Hao Chunliang (郝春亮) et al, “Interpretation and Reflections on the “AI Safety Governance Framework” (关于《人工智能安全治理框架》的解读与思考),” Weixin Official Accounts Platform, January 16, 2025, accessed July 13, 2025, https://mp.weixin.qq.com/s/DvQSJ0S83NcjR1slUp0Ilg

73 “AI Safety Standards System (V1.0) - Draft for Comments (人工智能安全标准体系（V1.0）[征求意见稿]),” National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2025, accessed July 13, 2025, https://web.archive.org/web/20250516072703/https://www.tc260.org.cn/upload/2025-01-24/1737709785951070331.pdf

74 “Cybersecurity Technology—Basic Security Requirements for Generative Artificial Intelligence Service (网络安全技术生成式人工智能服务安全基本要求),” National Public Service Platform for Standards Information, April 25, 2025, accessed July 2, 2025, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F67D3F376E0A0A0FF5317FB36B32A30A

75 “Notice Soliciting Contributors for the Standard ”Security Requirements for AI Code Generation Services” (关于征集《网络安全技术人工智能代码生成服务安全要求》标准参编单位的通知),” National Technical Committee 260 on Cybersecurity of Standardization Administration of China, November 13, 2024, accessed July 13, 2025, https://web.archive.org/web/20250702105359/https://www.tc260.org.cn/front/postDetail.html?id=20241113155236

76 “Cybersecurity Technology—Basic Security Requirements for Generative Artificial Intelligence Service (网络安全技术生成式人工智能服务安全基本要求),” National Public Service Platform for Standards Information, April 25, 2025, accessed July 2, 2025, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F67D3F376E0A0A0FF5317FB36B32A30A

77 “Emergency Response Guidelines for Security of Generative AI Services - Draft for Public Comments (网络安全标准实践指南——生成式人工智能服务安全应急响应指南 - 征求意见稿),” National Technical Committee 260 on Cybersecurity of Standardization

Administration of China, December 2024, accessed July 13, 2025, https://web.archive.org/web/20250116182644/https://www.tc260.org.cn/upload/2024-12-18/1734483139154029117.pdf

78 "Notice on Public Solicitation of Comments for the "Guidelines for Building an AI Safety Governance Standards System in the Industry and Information Technology Sector (2025) (Draft for Comments)" (关于《工业和信息化领域人工智能安全治理标准体系建设指南（2025）（征求意见稿）》公开征求意见的通知)," Weixin Official Accounts Platform, March 27, 2025, accessed July 13, 2025, https://mp.weixin.qq.com/s/8h6Py02IHUQ1gMuJAaS2Sg; "Guidelines for Building an AI Safety Governance Standards System in the Industry and Information Technology Sector (2025) (Draft for Comments) (《工业和信息化领域人工智能安全治理标准体系建设指南（2025）（征求意见稿）》)," March 27, 2025, accessed July 13, 2025, https://caict-llm-portal-storage.oss-cn-beijing.aliyuncs.com/6153dd34-d7fc-4d24-97b5-d40fc48105c5

79 "Beijing Institute of AI Safety and Governance Launch Announcement," Beijing Institute of AI Safety and Governance, September 4, 2024, accessed July 13, 2025, https://beijing.ai-safety-and-governance.institute/articles/beijing-institute-of-ai-safety-and-governance-launch-announcement

80 Yi Zeng et al., "Super Co-alignment of Human and AI for Sustainable Symbiotic Society," June 28, 2025, accessed July 13, 2025, arXiv: 2504.17404 [cs], http://arxiv.org/abs/2504.17404

81 "Shanghai AI Safety Governance Laboratory Unveiled at Closing Ceremony of the 2024 World AI Conference (上海人工智能安全治理实验室在 2024 世界人工智能大会暨人工智能全球治理高级别会议闭幕式上揭牌)," Cyberspace Administration of Jiangsu Province (江苏网信网), July 8, 2024, accessed July 13, 2025, https://www.jswx.gov.cn/csj/sh/202407/t20240708_3430231.shtml

82 "Shanghai Establishes AI Identifier Ecosystem Alliance; Xiaohongshu and MiniMax Named Among First Members (上海成立人工智能标识符生态联盟；小红书和 MiniMax 为首批成员)," Sohu, May 15, 2025, accessed July 13, 2025, https://www.sohu.com/a/895355596_122396381

83 "Global Security Initiative (全球安全倡议概念文件)," Ministry of Foreign Affairs of the People's Republic of China, February 21, 2023, accessed July 9, 2025, https://www.fmprc.gov.cn/wjbxw_new/202302/t20230221_11028322.shtml; "Global AI Governance Initiative (全球人工智能治理倡议)," October 20, 2023, Ministry of Foreign Affairs of the People's Republic of China, accessed July 9, 2025, https://www.mfa.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml

84 "The Bletchley Declaration by Countries Attending the AI Safety Summit," GOV.UK, November 2, 2023, accessed July 9, 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

85 Yoshua Bengio et al., "International AI Safety Report," 2025, https://www.gov.uk/government/publications/international-ai-safety-report-2025

86 "China Willing to Promote AI Development with Other Countries: Chinese Vice Premier," Xinhua, February 11, 2025, accessed July 9, 2025, https://english.news.cn/20250211/3977289ec01542faa6896fa4845031e8/c.html

87 "Full Text: Address by Chinese President Xi Jinping at Session II of 19th G20 Summit," The State Council of the People's Republic of China, accessed July 9, 2025, https://english.www.gov.cn/news/202411/19/content_WS673bcfd9c6d0868f4e8ed26b.html

88 "Xi Urges Promoting Healthy and Orderly Development of AI," Government of the People's Republic of China, April 28, 2025, accessed July 2, 2025, https://english.www.gov.cn/news/202504/29/content_WS68100ef1c6d0868f4e8f2275.html

89 "Full Text: Address by Chinese President Xi Jinping at Session II of 19th G20 Summit," The State Council of the People's Republic of China, accessed July 9, 2025, https://english.www.gov.cn/news/202411/19/content_WS673bcfd9c6d0868f4e8ed26b.html

90 "Li Qiang Attends Opening Ceremony of 2024 World AI Conference and Delivers a Speech (李强出席 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并致辞)," Ministry of Foreign Affairs of the People's Republic of China, July 4, 2024, accessed July 9, 2025, https://www.mfa.gov.cn/web/wjdt_674879/gjldrhd_674881/202407/t20240704_11448224.shtml

91 "Vice Minister of Foreign Affairs Ma Chaoxu Delivers Remarks at Global Governance Forum of 2024 World Artificial Intelligence Conference (外交部副部长马朝旭在 2024 世界人工智能大会暨人工智能全球治理高级别会议全球治理论坛上的发言)," Ministry of Foreign Affairs of the People's Republic of China, July 9, 2024, accessed July 9, 2025, https://www.fmprc.gov.cn/web/wjbxw_new/202407/t20240709_11450297.shtml

92 "Shanghai Declaration on Global AI Governance (人工智能全球治理上海宣言)," Government of the People's Republic of China, July 4, 2024, accessed July 9, 2025, https://www.gov.cn/yaowen/liebiao/202407/content_6961358.htm

93 "Ding Xuexiang on Enabling AI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity (丁薛祥谈促进人工智能等科技创新更好造福全人类)," Ministry of Foreign Affairs of the People's Republic of China, January 22, 2025, accessed July 9, 2025, https://www.mfa.gov.cn/web/wjdt_674879/gjldrhd_674881/202501/t20250122_11542793.shtml

94 Jeremy Goldkorn, "Gray Rhinos and Risk Awareness in China and the U.S. —Q&A with Michele Wucker," The China Project, October 21, 2022, accessed July 9, 2025, https://thechinaproject.com/2022/10/21/gray-rhinos-and-risk-awareness-in-china-and-the-u-s-qa-with-michele-wucker/

95 "Ding Xuexiang on Enabling AI and Other Scientific and Technological Innovation to Better Serve the Well-being of All Humanity (丁薛祥谈促进人工智能等科技创新更好造福全人类)," Ministry of Foreign Affairs of the People's Republic of China, January 22, 2025, accessed July 9, 2025, https://www.mfa.gov.cn/web/wjdt_674879/gjldrhd_674881/202501/t20250122_11542793.shtml

96 "Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development," United Nations General Assembly, March 11, 2024, accessed July 9, 2025, https://docs.un.org/en/A/78/L.49

97 "Enhancing International Cooperation on Capacity-Building of Artificial Intelligence," United Nations General Assembly, March 11, 2024, accessed July 9, 2025, https://docs.un.org/en/A/78/L.86

98 "AI Capacity-Building Action Plan for Good and for All," Ministry of Foreign Affairs of the People's Republic of China, September 27, 2024, accessed July 9, 2025, https://www.mfa.gov.cn/eng/wjbzhd/202409/t20240927_11498465.html

99 "Group of Friends for International Cooperation on AI Capacity-building Formally Established," Permanent Mission of the People's Republic of China to the UN, December 5, 2024, accessed July 9, 2025, https://un.china-mission.gov.cn/eng/chinaandun/202412/t20241206_11539640.htm; "Glossary," United Nations Security Council, accessed July 9, 2025, https://main.un.org/securitycouncil/en/content/glossary

100 "A Side Event of the Group of Friends for International Cooperation on AI Capacity-Building Successfully Held at the UN Headquarters," Permanent Mission of the People's Republic of China to the UN, May 6, 2025, accessed July 9, 2025, https://un.china-mission.gov.cn/eng/chinaandun/202505/t20250514_11622260.htm

101 He Yin, "China Strives to Facilitate Global Sharing of Benefits from AI Development," People's Daily, December 9, 2024, accessed July 9, 2025, https://en.people.cn/n3/2024/1209/c90000-20251708.html

102 "Strengthen AI Capacity Building and Promote Fair and Inclusive Global Governance —Video Remarks by Vice Minister Ma Chaoxu at Opening Ceremony of AI Capacity Building Workshop (加强人工智能能力建设，推动公平普惠全球治理—马朝旭副部长在人工智能能力建设研讨班开班式上的视频致辞)," Ministry of Foreign Affairs of the People's Republic of China, September 4, 2024, accessed July 9, 2025, https://www.fmprc.gov.cn/wjbxw_new/202409/t20240904_11484756.shtml; "Vice Foreign Minister Ma Zhaoxu Attends the Opening Ceremony of the Second Workshop on AI Capacity Building," Ministry of Foreign Affairs of the

People's Republic of China, May 12, 2025, accessed July 9, 2025, https://www.fmprc.gov.cn/eng/xw/wjbxw/202505/t20250514_11622546.html

103 Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院), "Using "AI" as a Bridge, Representatives from 35 Countries Gather in Beijing to Launch Dialogue (以 "AI" 为桥，35 国代表齐聚北京开启这场对话)," Weixin Official Accounts Platform, May 13, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/n7h0KEWXH1npPqrbTZkuNA

104 "Xi Jinping Meets U.S. President Biden in Lima (习近平同美国总统拜登在利马举行会晤)," Ministry of Foreign Affairs of the People's Republic of China, November 17, 2024, accessed July 9, 2025, https://www.mfa.gov.cn/zyxw/202411/t20241117_11527702.shtml

105 "On-the-Record Press Gaggle by APNSA Jake Sullivan on President Biden's Meeting with President Xi Jinping," The White House, November 17, 2024, accessed July 9, 2025, https://web.archive.org/web/20250120081423/https://www.whitehouse.gov/briefing-room/press-briefings/2024/11/17/on-the-record-press-gaggle-by-apnsa-jake-sullivan-on-president-bidens-meeting-with-president-xi-jinping/

106 "Wang Yi and U.S. National Security Advisor Jake Sullivan Hold Strategic Communication," Ministry of Foreign Affairs of the People's Republic of China, August 28, 2024, accessed July 9, 2025, https://www.fmprc.gov.cn/eng/wjbzhd/202408/t20240830_11482159.html

107 Daniel Desrochers et al., "Trump's Latest Trade 'Deal' with China Underscores Key U.S. Disadvantage," POLITICO, June 11, 2025, accessed July 11, 2025, https://www.politico.com/news/2025/06/11/trump-announced-another-deal-with-china-will-it-hold-00400288

108 Jiangtao Shi and Meredith Chen, "Will Trump Visit China for WWII Victory Day? 'Cautious Optimism' as Speculation Swirls," South China Morning Post, June 30, 2025, accessed July 9, 2025, https://www.scmp.com/news/china/diplomacy/article/3316423/will-trump-visit-china-wwii-victory-day-cautious-optimism-speculation-swirls

109 "China and the UK Hold a Dialogue on Artificial Intelligence," Ministry of Foreign Affairs of the People's Republic of China, May 20, 2025, accessed July 9, 2025, https://www.fmprc.gov.cn/mfa_eng/xw/wjbxw/202505/t20250521_11629987.html

110 "Zheng Zeguang Discusses China's Position on Global AI Governance (郑泽光谈中国对全球人工智能治理的主张)," Embassy of the People's Republic of China in the United Kingdom of Great Britain and Northern Ireland, May 14, 2025, accessed July 9, 2025, http://gb.china-embassy.gov.cn/dshd/202505/t20250515_11623174.htm; "Zheng Zeguang Discusses Key Characteristics of China's Current AI Development (郑泽光谈当前中国人工智能发展的主要特点)," Embassy of the People's Republic of China in the United Kingdom of Great Britain and Northern Ireland, May 14, 2025, accessed July 9, 2025, http://gb.china-embassy.gov.cn/dshd/202505/t20250515_11623173.htm

111 "Remarks Made by Technology Secretary Peter Kyle at the Munich Security Conference," GOV.UK, February 14, 2025, accessed July 9, 2025, https://www.gov.uk/government/speeches/remarks-made-by-technology-secretary-peter-kyle-at-the-munich-security-conference

112 "2025 UK-China Economic and Financial Dialogue: Policy Outcomes," GOV.UK, January 11, 2025, accessed July 9, 2025, https://www.gov.uk/government/publications/2025-uk-china-economic-and-financial-dialogue-policy-outcomes/2025-uk-china-economic-and-financial-dialogue-policy-outcomes

113 "First China–Singapore Digital Policy Dialogue Meeting Held in Beijing (中新数字政策对话机制第一次会议在京召开)," National Data Administration (国家数据局), June 28, 2024, accessed July 9, 2025, https://www.nda.gov.cn/sjj/jgsz/jld/llh/llhldhd/0830/20240830172430339878717_pc.html; "Singapore and China Advance Cooperation at Inaugural Singapore China Digital Policy Dialogue," Singaporean Ministry of Digital Development and Information, June 27, 2024, accessed July 9, 2025, https://www.mddi.gov.sg/newsroom/sg-china-advance-cooperation-at-inaugural-digital-policy-dialogue/

114 "Singapore and China Advance Cooperation at Inaugural Singapore China Digital Policy Dialogue," Singaporean Ministry of Digital Development and Information, June 27, 2024, accessed July 9, 2025, https://www.mddi.gov.sg/newsroom/sg-china-advance-cooperation-at-inaugural-digital-policy-dialogue/

115 "Opening Remarks by SMS Tan Kiat How at Singapore China Digital Policy Dialogue," Singaporean Ministry of Digital Development and Information, July 27, 2024, accessed July 9, 2025, https://www.mddi.gov.sg/newsroom/opening-by-sms-tan-kiat-how-at-singapore-china-digital-policy-dialogue/

116 "Joint Communiqué of the 29th Regular Meeting Between the Chinese and Russian Prime Ministers (Full Text) (中俄总理第二十九次定期会晤联合公报 [全文])," Government of the People's Republic of China, August 22, 2024, accessed July 9, 2025, https://www.gov.cn/yaowen/liebiao/202408/content_6969793.htm

117 Gleb Bryanski, "Russia's Sberbank Plans Joint AI Research with China as DeepSeek Leaps Forward," *Reuters*, February 6, 2025, accessed July 9, 2025, https://www.reuters.com/technology/artificial-intelligence/russias-sberbank-plans-joint-ai-research-with-china-deepseek-leaps-forward-2025-02-06/

118 "Joint Statement Between the People's Republic of China and the Russian Federation on Deepening the Comprehensive Strategic Partnership in the New Era on the 75th Anniversary of Diplomatic Relations (Full Text) (中华人民共和国和俄罗斯联邦在两国建交 75 周年之际关于深化新时代全面战略协作伙伴关系的联合声明 [全文])," Ministry of Foreign Affairs of the People's Republic of China, May 16, 2024, accessed July 9, 2025, https://www.mfa.gov.cn/zyxw/202405/t20240516_11305860.shtml

119 China Academy of Information and Communications Technology, "China–BRICS AI Development and Cooperation Center Launch Ceremony Held in Beijing (中国—金砖国家人工智能发展与合作中心启动仪式在京举行)," Weixin Official Accounts Platform, July 20, 2024, accessed July 9, 2025, https://mp.weixin.qq.com/s/1udADFXVAUc8AfruPNF5ig

120 Ministry of Industry and Information Technology, "BRICS High‑Level Forum on AI Held in Brasília (金砖国家人工智能高级别论坛在巴西利亚举行)," Weixin Official Accounts Platform, May 22, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/yio811JkxekuGPk0-fExHA; China Academy of Information and Communications Technology, "China–BRICS AI Development and Cooperation Center Releases "Collection of Typical AI Products and Application Cases"(中国—金砖国家人工智能发展与合作中心发布《人工智能典型产品与应用案例集》)," Weixin Official Accounts Platform, May 27, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/jMTV4-9UYv15MQG7TKsBIQ

121 "10th Ministerial Conference of the China-Arab States Cooperation Forum Adopts "Beijing Declaration"and "Action Implementation Plan"(中阿合作论坛第十届部长级会议通过《北京宣言》和《行动执行计划》)," Government of the People's Republic of China, May 31, 2024, accessed July 9, 2025, https://www.gov.cn/yaowen/liebiao/202405/content_6954753.htm

122 "President Xi's Keynote Speech at the Opening Ceremony of the Fourth Ministerial Meeting of the China-CELAC Forum," Qiushi, May 14, 2025, accessed July 9, 2025, http://en.qstheory.cn/2025-05/14/c_1092333.htm

123 "China-Latin America Internet Development and Cooperation Forum Held in Xi'an (中拉互联网发展与合作论坛在西安举办)," Cyberspace Administration of China, May 15, 2025, accessed July 9, 2025, https://www.cac.gov.cn/2025-05/15/c_1749016904407818.htm

124 "Xi Jinping Holds Talks with Brazilian President Lula (习近平同巴西总统卢拉会谈)," People's Daily, accessed July 9, 2025, http://politics.people.com.cn/n1/2025/0514/c1024-40479145.html; "Xi Jinping Meets Colombian President Petro (习近平会见哥伦比亚总统佩特罗)," People's Daily, May 15, 2025, accessed July 9, 2025, http://politics.people.com.cn/n1/2025/0515/c1024-40479969.html

125 China Academy of Information and Communications Technology, "China–Africa Digital Cooperation Forum Held in Beijing (中非数字合作论坛在北京召开)," Weixin Official Accounts Platform, July 29, 2024, accessed July 9, 2025, https://mp.weixin.qq.com/s/mXLVuKY73V5yQqkRQbuCLQ; "Forum on China-Africa Cooperation Beijing Action Plan (2025-2027)," Ministry of Foreign Affairs

of the People's Republic of China, July 29, 2024, accessed July 9, 2025, https://www.mfa.gov.cn/eng/xw/zyxw/202409/t20240905_11485719.html

126 Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院), "China AI Safety and Development Association Hosts Side Event During Paris AI Action Summit (中国人工智能发展与安全研究网络在巴黎人工智能行动峰会期间成功举办专题边会)," Weixin Official Accounts Platform, February 12, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/uV5aOk9D3J-eE3huFfjlCg

127 Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院), "China AI Safety and Development Association Hosts Side Event During Paris AI Action Summit (中国人工智能发展与安全研究网络在巴黎人工智能行动峰会期间成功举办专题边会)," Weixin Official Accounts Platform, February 12, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/uV5aOk9D3J-eE3huFfjlCg

128 Fu Ying, "Opinion | Cooperation for AI Safety Must Transcend Geopolitical Interference," South China Morning Post, February 12, 2025, accessed July 9, 2025, https://www.scmp.com/opinion/china-opinion/article/3298281/cooperation-ai-safety-must-transcend-geopolitical-interference

129 China AI Safety & Development Association, "Initiative on Promoting International Cooperation on AI Safety and Inclusive Development," February 1, 2025, accessed July 9, 2025, https://cnaisi.cn/filedownload/153149

130 Fu Ying and John Allen, "Together, The U.S. And China Can Reduce The Risks From AI," December 17, 2020, accessed July 9, 2025, https://www.noemamag.com/together-the-u-s-and-china-can-reduce-the-risks-from-ai; "U.S.-China Track II Dialogue Round XII on Artificial Intelligence and International Security," Center For International Security and Strategy Tsinghua University, February 28, 2025, accessed July 9, 2025, https://ciss.tsinghua.edu.cn/info/event/8009

131 "INHR," INHR, accessed July 9, 2025, https://inhr.org/welcome

132 Bob Davis, "Back on Track?," The Wire China, January 15, 2024, 0:00 a.m. (Z), accessed July 9, 2025, https://www.thewirechina.com/2024/01/14/back-on-track-two-dialogues-u-s-china/; "INHR Recommendations on AI Military Risk Reduction," INHR, accessed July 9, 2025, https://inhr.org/ai

133 "Center Advances U.S.-China Understanding of AI Governance," Yale Law School, August 6, 2024, accessed July 9, 2025, https://law.yale.edu/yls-today/news/center-advances-us-china-understanding-ai-governance

134 "IDAIS-Oxford, 2023," International Dialogues on AI Safety, October 31, 2023, accessed July 9, 2025, https://idais.ai/dialogue/idais-oxford/; "IDAIS-Beijing, 2024," International Dialogues on AI Safety, March 10, 2024, accessed July 9, 2025, https://idais.ai/dialogue/idais-beijing/; "IDAIS-Venice, 2024," International Dialogues on AI Safety, September 8, 2024, accessed July 9, 2025, https://idais.ai/dialogue/idais-venice/

135 "NTI Convenes First International AI-Bio Forum," The Nuclear Threat Initiative, April 17, 2024, accessed July 9, 2025, https://www.nti.org/news/nti-convenes-the-first-international-ai-bio-forum/

136 "The Royal Society and Chinese Academy of Sciences Policy Dialogue on AI," The Royal Society, September 29, 2020, https://royalsociety.org/-/media/blogs/2021/04/china-ai/RS-CAS-AI-policy-workshop-report.pdf; "China Academy of Sciences and the Royal Society Jointly Host Sino–UK AI Policy Dialogue and Seminar (中科院与英国皇家学会共同举办中英人工智能政策对话研讨会)," Institute of Automation, Chinese Academy of Sciences, October 16, 2020, accessed July 9, 2025, https://web.archive.org/web/20240221151654/http://www.ia.cas.cn/xwzx/kydt/202010/t20201016_5718238.html; "Third Seminar on AI Ethics by the China Academy of Sciences and Royal Society Held in London (第三届中国科学院—英国皇家学会人工智能伦理研讨会在伦敦举行)," Bureau of International Cooperation, Chinese Academy of Sciences (中国科学院国际合作局), September 30, 2024, accessed July 9, 2025, https://bic.cas.cn/gjhzdt/202409/t20240930_5034430.html

137 "CISS Hosts China–EU Dialogue on AI and International Security (CISS 举办中欧人工智能与国际安全对话会)," Center For International Security and Strategy Tsinghua University, February 28, 2024, accessed July 9, 2025, https://ciss.tsinghua.edu.cn/info/wzjx/6951; Center For International Security and Strategy Tsinghua University, "CISS Hosts 4th China–EU Dialogue on AI and International Security (CISS 举办第四轮"中欧人工智能与国际安全对话")," Weixin Official Accounts Platform, March 3, 2025, accessed July 9, 2025, https://mp.weixin.qq.com/s/oNxcoScY3we7nbuhLg9Nsg

138 "P5 Experts Roundtable –Online Meeting on the AI Nuclear Nexus on 24 June 2024," Geneva Centre for Security Policy, July 15, 2024, accessed July 9, 2025, https://www.gcsp.ch/news/p5-experts-roundtable-online-meeting-ai-nuclear-nexus-24-june-2024

139 "U.S.-China Track II Dialogue on the Digital Economy," The National Committee on United States-China Relations, October 2024, accessed July 9, 2025, https://www.ncuscr.org/program/us-china-track-ii-dialogue-digital-economy/

140 Madhumita Murgia, "US Companies and Chinese Experts Engaged in Secret Diplomacy on AI Safety," Financial Times, January 11, 2024, accessed July 9, 2025, https://www.ft.com/content/f87b693f-9ba3-4929-8b95-a296b0278021

141 Alice Saltini, "AI and Nuclear Command, Control and Communications: P5 Perspectives," European Leadership Network, November 13, 2023, accessed July 9, 2025, https://europeanleadershipnetwork.org/report/ai-and-nuclear-command-control-and-communications-p5-perspectives/

142 "The First Meeting of the Sino-European Cyber Dialogue (SECD) Convened on 31 March–1 April in Geneva, Switzerland," Geneva Centre for Security Policy, April 1, 2014, accessed July 9, 2025, https://www.gcsp.ch/events/1st-sino-european-cyber-dialogue; "China Institutes of Contemporary International Relations and ESMT Jointly Host 11th China–EU Track II Cyber Dialogue (现代院与欧洲管理技术学院联合举办"第十一届中欧网络二轨对话会")," China Institutes of Contemporary International Relations, November 21, 2023, accessed July 9, 2025, http://www.cicir.ac.cn/NEW/event.html?id=6f23ec7f-0e17-4d57-950a-98c7a0667615

143 "UK-China Track 1.5 Cyber Dialogue," Chatham House, accessed July 9, 2025, https://www.chathamhouse.org/about-us/our-departments/international-security-programme/uk-china-track-15-cyber-dialogue; "China Institutes of Contemporary International Relations and Chatham House Jointly Host 3rd China–UK Track-II Cyber Dialogue 现代院与英国皇家国际事务研究所联合举办"第三届中英网络二轨对话会")," China Institutes of Contemporary International Relations, November 23, 2023, accessed July 9, 2025, http://www.cicir.ac.cn/NEW/event.html?id=7424deb8-325e-453e-bb9f-aac48acafaa1

144 Concordia AI, "The State of China-Western Track 1.5 and 2 Dialogues on AI," AI Safety in China, February 10, 2024, accessed July 9, 2025, https://aisafetychina.substack.com/p/the-state-of-china-western-track

145 Concordia AI, "The State of China-Western Track 1.5 and 2 Dialogues on AI," AI Safety in China, February 10, 2024, accessed July 9, 2025, https://aisafetychina.substack.com/p/the-state-of-china-western-track

146 "NTI Convenes First International AI-Bio Forum," The Nuclear Threat Initiative, April 17, 2024, accessed July 9, 2025, https://www.nti.org/news/nti-convenes-the-first-international-ai-bio-forum/

147 INHR, Center for Health Security, John Hopkins Bloomberg School of Public Health, and Center for a New American Security, "RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIXBIO RISKS WHILE PURSUING THE PROMISE OF AI TECHNOLOGY FOR IMPROVEMENTS IN HEALTH AND WELLBEING," 2024, accessed July 9, 2025, https://drive.google.com/file/d/1U9MeF4JjvKu8nwclXTtcKH95nDDwRJcL/view

148 "The Singapore Consensus on Global AI Safety Research Priorities," May 8, 2025, accessed July 9, 2025, https://aisafetypriorities.org

149 "IDAIS-Venice, 2024," International Dialogues on AI Safety, September 8, 2024, accessed July 9, 2025, https://idais.ai/dialogue/idais-venice/

150 CISS Working Group on Artificial Intelligence Glossary Research, "Interim Findings on Artificial Intelligence Terms," 2024, https://ciss.tsinghua.edu.cn/upload_files/atta/1725035303351_4A.pdf; "Glossary of Artificial Intelligence Terms," Brookings, August 30, 2024, accessed July 9, 2025, https://www.brookings.edu/articles/glossary-of-artificial-intelligence-terms/?b=1

151 "P5 Glossary of Key Nuclear Terms: Working Paper Submitted by China, France, the Russian Federation, the United Kingdom of Great Britain and Northern Ireland and the United States of America," United Nations Digital Library, 2021, accessed July 9, 2025, https://digitallibrary.un.org/record/3956428

152 "AIxBio Global Forum," The Nuclear Threat Initiative, March 18, 2025, accessed July 9, 2025, https://www.nti.org/about/programs-projects/project/aixbio-global-forum/

153 "White Paper: Research Agenda for Safeguarding AI-Bio Capabilities," The Nuclear Threat Initiative, May 29, 2024, accessed July 9, 2025, https://www.nti.org/wp-content/uploads/2024/06/Research-Agenda-for-Safeguarding-AI-Bio-Capabilities.pdf

154 Nikki Teran, "NTI | Bio Advances International Dialogue on AI Biosecurity Through Expert Working Groups," The Nuclear Threat Initiative, November 26, 2024, accessed July 9, 2025, https://www.nti.org/risky-business/nti-bio-advances-international-dialogue-on-ai-biosecurity-through-expert-working-groups/

155 INHR, Center for Health Security, John Hopkins Bloomberg School of Public Health, and Center for a New American Security, "RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIXBIO RISKS WHILE PURSUING THE PROMISE OF AI TECHNOLOGY FOR IMPROVEMENTS IN HEALTH AND WELLBEING," 2024, accessed July 9, 2025, https://drive.google.com/file/d/1U9MeF4JjvKu8nwclXTtcKH95nDDwRJcL/view

156 National Committee on U.S.-China Relations and China-U.S. Green Fund, "U.S.-CHINA TRACK II DIALOGUE ON THE DIGITAL ECONOMY CONSENSUS AGREEMENT," October 2024, accessed July 9, 2025, https://www.ncuscr.org/wp-content/uploads/2025/03/DE2024-Consensus-Agreement-FINAL-25-03-11-English-1.pdf

157 Concordia AI, "Concordia AI Chinese Technical AI Safety Database," July 1, 2025, https://docs.google.com/spreadsheets/d/1LuM3xPKILW8b40jq4A57vC7uOinwf5PPojL8m-hXx78/edit?usp=sharing

158 Concordia AI, "Concordia AI Chinese Technical AI Safety Database," July 1, 2025, https://docs.google.com/spreadsheets/d/1LuM3xPKILW8b40jq4A57vC7uOinwf5PPojL8m-hXx78/edit?usp=sharing

159 "ICLR 2025," ICLR, accessed July 9, 2025, https://iclr.cc/

160 "ICML 2025 Call for Papers," International Conference on Machine Learning, accessed July 9, 2025, https://icml.cc/Conferences/2025/CallForPapers

161 "2024 Dates and Deadlines," NeurIPS, accessed July 9, 2025, https://nips.cc/Conferences/2024/Dates

162 "The State of Global AI Safety Research," Emerging Technology Observatory, April 3, 2024, accessed July 9, 2025, https://eto.tech/blog/state-of-global-ai-safety-research/

163 AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, "The AI Index 2025 Annual Report: Chapter 3: Responsible AI," April 2025, accessed July 9, 2025, https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter3_final.pdf

164 Jason Zhou, Kwan Yee Ng, and Brian Tse, "State of AI Safety in China Spring 2024 Report," Concordia AI, May 14, 2024, accessed July 9, 2025, https://concordia-ai.com/wp-content/uploads/2024/05/State-of-AI-Safety-in-China-Spring-2024-Report-public.pdf

165 John P. A. Ioannidis, "August 2024 Data-Update for "Updated Science-Wide Author Databases of Standardized Citation Indicators"" 7 (September 16, 2024), accessed July 9, 2025, https://doi.org/10.17632/btchxktzyw.7, https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/7

166 Concordia AI, "Concordia AI Chinese Technical AI Safety Database," July 1, 2025, https://docs.google.com/spreadsheets/d/1LuM3xPKILW8b40jq4A57vC7uOinwf5PPojL8m-hXx78/edit?usp=sharing

167 Xiaoya Lu et al., "X-Boundary: Establishing Exact Safety Boundary to Shield LLMs from Multi-Turn Jailbreaks without Compromising Usability," March 6, 2025, accessed July 9, 2025, arXiv: 2502.09990 [cs], http://arxiv.org/abs/2502.09990

168 Zhenhong Zhou et al., "How Alignment and Jailbreak Work: Explain LLM Safety through Intermediate Hidden States," June 13, 2024, accessed July 9, 2025, arXiv: 2406.05644 [cs], http://arxiv.org/abs/2406.05644

169 Zhenglin Hua et al., "Steering LVLMs via Sparse Autoencoder for Hallucination Mitigation," May 22, 2025, accessed July 9, 2025, arXiv: 2505.16146 [cs], http://arxiv.org/abs/2505.16146

170 Yuxin Xiao et al., "Enhancing Multiple Dimensions of Trustworthiness in LLMs via Sparse Activation Control," November 4, 2024, accessed July 9, 2025, arXiv: 2411.02461 [cs], http://arxiv.org/abs/2411.02461

171 Jiaming Ji et al., "Aligner: Efficient Alignment by Learning to Correct," November 2, 2024, accessed July 9, 2025, arXiv: 2402.02416 [cs], http://arxiv.org/abs/2402.02416

172 HyunJin Kim et al., "The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment," December 25, 2024, accessed July 9, 2025, arXiv: 2412.16468 [cs], http://arxiv.org/abs/2412.16468; HyunJin Kim et al., "Research on Superalignment Should Advance Now with Parallel Optimization of Competence and Conformity," March 8, 2025, accessed July 9, 2025, arXiv: 2503.07660 [cs], http://arxiv.org/abs/2503.07660

173 Wenkai Yang et al., "Super(Ficial)-Alignment: Strong Models May Deceive Weak Models in Weak-to-Strong Generalization," February 28, 2025, accessed July 9, 2025, arXiv: 2406.11431 [cs], http://arxiv.org/abs/2406.11431; Junhao Shi et al., "How to Mitigate Overfitting in Weak-to-strong Generalization?," March 6, 2025, accessed July 9, 2025, arXiv: 2503.04249 [cs], http://arxiv.org/abs/2503.04249

174 Xueru Wen et al., "Scalable Oversight for Superhuman AI via Recursive Self-Critiquing," May 30, 2025, accessed July 9, 2025, arXiv: 2502.04675 [cs], http://arxiv.org/abs/2502.04675

175 Hao Lang, Fei Huang, and Yongbin Li, "Debate Helps Weak-to-Strong Generalization," January 21, 2025, accessed July 9, 2025, arXiv: 2501.13124 [cs], http://arxiv.org/abs/2501.13124

176 Wenkai Yang et al., "Super(Ficial)-Alignment: Strong Models May Deceive Weak Models in Weak-to-Strong Generalization," February 28, 2025, accessed July 9, 2025, arXiv: 2406.11431 [cs], http://arxiv.org/abs/2406.11431

177 Jiaming Ji et al., "Mitigating Deceptive Alignment via Self-Monitoring," May 24, 2025, accessed July 9, 2025, arXiv: 2505.18807 [cs], http://arxiv.org/abs/2505.18807

178 Yichen Wu et al., "OpenDeception: Benchmarking and Investigating AI Deceptive Behaviors via Open-ended Interaction Simulation," April 18, 2025, accessed July 9, 2025, arXiv: 2504.13707 [cs], http://arxiv.org/abs/2504.13707

179 Yihe Fan et al., "Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems," May 23, 2025, accessed July 9, 2025, arXiv: 2505.17815 [cs], http://arxiv.org/abs/2505.17815

180 Rongwu Xu et al., "Nuclear Deployed: Analyzing Catastrophic Risks in Decision-making of Autonomous LLM Agents," March 23, 2025, accessed July 9, 2025, arXiv: 2502.11355 [cs], http://arxiv.org/abs/2502.11355

181 Zhexin Zhang et al., "From Theft to Bomb-Making: The Ripple Effect of Unlearning in Defending Against Jailbreak Attacks," May 20, 2025, accessed July 9, 2025, arXiv: 2407.02855 [cs], http://arxiv.org/abs/2407.02855

182 Zesheng Shi, Yucheng Zhou, and Jing Li, "Safety Alignment via Constrained Knowledge Unlearning," May 24, 2025, accessed July 9, 2025, arXiv: 2505.18588 [cs], http://arxiv.org/abs/2505.18588

183 Xiaoyu Xu et al., "Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs," May 22, 2025, accessed July 9, 2025, arXiv: 2505.16831 [cs], http://arxiv.org/abs/2505.16831

184 Jiyan He et al., "Control Risk for Potential Misuse of Artificial Intelligence in Science," December 11, 2023, accessed July 9, 2025, arXiv: 2312.06632 [cs], http://arxiv.org/abs/2312.06632; Xiangru Tang et al., "Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science," June 5, 2024, accessed July 9, 2025, arXiv: 2402.04247 [cs], http://arxiv.org/abs/2402.04247

185 Haochen Zhao et al., "ChemSafetyBench: Benchmarking LLM Safety on Chemistry Domain," November 23, 2024, accessed July 9, 2025, arXiv: 2411.16736 [cs], http://arxiv.org/abs/2411.16736

186 Tianhao Li et al., "SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks," December 16, 2024, accessed July 9, 2025, arXiv: 2410.03769 [cs], http://arxiv.org/abs/2410.03769

187 Cheng Li, "A Ladder to Power and Influence: China's Official Think Tanks to Watch," China-US Focus, October 14, 2022, accessed July 10, 2025, https://www.chinausfocus.com/2022-CPC-congress/a-ladder-to-power-and-influence-chinas-official-think-tanks-to-watch; Xiaoxuan Li, Kejia Yang, and Xiaoxi Xiao, "Scientific Advice in China: The Changing Role of the Chinese Academy of Sciences," *Palgrave Communications* 2, no. 1 (July 12, 2016): 16045, accessed July 10, 2025, https://doi.org/10.1057/palcomms.2016.45, https://www.nature.com/articles/palcomms201645; Cheng Li, "The Power of Ideas: The Rising Influence of Thinkers and Think Tanks in China," Brookings, May 28, 2017, accessed July 10, 2025, https://www.brookings.edu/books/the-power-of-ideas-the-rising-influence-of-thinkers-and-think-tanks-in-china/; Xufeng Zhu, *The Politics of Expertise in China* (London: Routledge, 2020), ISBN: 978-0-367-58441-2, https://www.routledge.com/The-Politics-of-Expertise-in-China/Zhu/p/book/9780367584412

188 "Xi Jinping Presides over the 9th Study Session of the Politburo (习近平主持中共中央政治局第九次集体学习并讲话)," Government of the People's Republic of China, October 31, 2018, accessed July 2, 2025, https://www.gov.cn/xinwen/2018-10/31/content_5336251.htm; "Xi Jinping Emphasized during the 20th Study Session of the Politburo: Uphold Self-Reliance, Highlight Application-Oriented Development, and Promote the Orderly Development of Artificial Intelligence (习近平在中共中央政治局第二十次集体学习时强调：坚持自立自强突出应用导向推动人工智能健康有序发展)," Government of the People's Republic of China, April 26, 2025, accessed July 2, 2025, https://www.gov.cn/yaowen/liebiao/202504/content_7021072.htm

189 Concordia AI, "AI Safety in China #1," AI Safety in China, August 24, 2023, accessed July 10, 2025, https://aisafetychina.substack.com/p/ai-safety-in-china-1; Beijing Academy of AI (智源研究院), "AI Safety Waters Run Deep and Swift: Huang Tiejun Speaks for the First Time on AGI Capability and Risk Classification as 2024 BAAI Conference Concludes Successfully (AI 安全水深流急，黄铁军首谈 AGI 能力与风险分级，2024 智源大会圆满落幕)," Weixin Official Accounts Platform, June 15, 2024, accessed July 10, 2025, https://mp.weixin.qq.com/s/VHzB4zR73xxmh_NCoyDBtA; Concordia AI, "Video & Summary | BAAI Conference "AI Safety Forum," Co-Hosted by Concordia AI: Delving Into Risk Red Lines and Exploring Countermeasures (视频 + 总结｜智源大会 "AI 安全论坛," 安远 AI 联合主办：深挖风险红线，探讨应对措施)," Weixin Official Accounts Platform, July 4, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/9-uGFWOCJCOBmzlFNeiW6A

190 "Li Qiang Attends Opening Ceremony of 2024 World AI Conference High-Level Dialogue on Global AI Governance and Delivers Speech (李强出席 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并致辞)," Government of the People's Republic of China, July 4, 2024, accessed July 10, 2025, https://www.gov.cn/yaowen/liebiao/202407/content_6961222.htm

191 "Shanghai Declaration on Global AI Governance (人工智能全球治理上海宣言)," Government of the People's Republic of China, July 4, 2024, accessed July 9, 2025, https://www.gov.cn/yaowen/liebiao/202407/content_6961358.htm

192 "Xue Lan Delivers Keynote at Plenary Session of the 2024 World AI Conference (薛澜受邀出席 2024 世界人工智能大会暨人工智能全球治理高级别会议并在全体会议上发表主旨演讲)," Institute for AI International Governance, Tsinghua University (清华大学人工智能国际治理研究院), July 11, 2024, accessed July 11, 2025, https://aiig.tsinghua.edu.cn/info/1294/2029.htm; Zhou Bowen (周伯文), "Exploring the 45-Degree Balance Principle of Artificial Intelligence (探索人工智能 45° 平衡律)," Shanghai AI Lab (上海人工智能实验室), July 4, 2024, accessed July 10, 2025, http://www.shlab.org.cn/news/5443947

193 "Full Footage of the 2024 World Artificial Intelligence Conference High-Level Meeting on Global AI Governance (2024 世界人工智能大会暨人工智能全球治理高级别会议全程大放送)," Youtube, July 4, 2024, accessed July 10, 2025, https://www.youtube.com/live/1DGIS2Hpzks

194 "Highlights from the Opening Ceremony of the 2023 World Artificial Intelligence Conference (WAIC): Six Key Moments Shine Together to Build a Global Premier AI Conference (世界人工智能大会（WAIC 2023）开幕式六大亮点齐发合力打造全球顶尖人工智能盛会)," Chinanews (中新网), July 7, 2023, accessed July 10, 2025, https://www.sh.chinanews.com.cn/swzx/2023-07-07/113794.shtml; "2023 World AI Conference Opening Ceremony (2023 世界人工智能大会开幕式)," Youtube, July 6, 2023, accessed July 10, 2025, https://www.youtube.com/watch?v=gbrw-RsOPy4

195 "2024 World Internet Conference Wuzhen Summit Digital Cooperation Forum under the Global Development Initiative (2024 年世界互联网大会乌镇峰会全球发展倡议数字合作论坛举行)," Cyberspace Administration of China, November 21, 2024, accessed July 10, 2025, https://www.cac.gov.cn/2024-11/21/c_1733880951044682.htm

196 "Xi Jinping Sends Congratulations to the Opening Ceremony of 2024 World Internet Conference Wuzhen Summit via Video Link," Ministry of Foreign Affairs of the People's Republic of China, November 20, 2024, accessed July 10, 2025, https://www.fmprc.gov.cn/eng/xw/zyxw/202411/t20241121_11530440.html; "Ding Xuexiang Attends Opening Ceremony of the 2024 World Internet Conference Wuzhen Summit and Delivers Keynote Speech (丁薛祥出席 2024 年世界互联网大会乌镇峰会开幕式并发表主旨讲话)," World Internet Conference (世界互联网大会), November 20, 2024, accessed July 10, 2025, https://cn.wicinternet.org/2024-11/20/content_37689370.htm

197 "Xi Jinping Delivers Video Address to Opening Ceremony of 2023 World Internet Conference Wuzhen Summit (习近平向 2023 年世界互联网大会乌镇峰会开幕式发表视频致辞)," Ministry of Foreign Affairs of the People's Republic of China, November 8, 2023, accessed July 10, 2025, https://www.fmprc.gov.cn/zyxw/202311/t20231108_11175760.shtml; "Plenary Session Held at the 2023 World Internet Conference Wuzhen Summit (2023 年世界互联网大会乌镇峰会举行全体会议)," Cyberspace Administration of China, November 8, 2023, accessed July 10, 2025, https://www.cac.gov.cn/2023-11/08/c_1701111282570507.htm

198 World Internet Conference Working Group on Artificial Intelligence, "Developing Responsible Generative Artificial Intelligence Research Report and Consensus," November 2023, https://www.wicinternet.org/pdf/DevelopingResponsibleGenerativeArtificialIntelligenceResearchReportandConsensus.pdf

199 Shanghai Academy of Social Sciences et al., "Research Report on Global AI Governance," World Internet Conference (世界互联网大会), 2024, accessed July 10, 2025, https://www.wicinternet.org/pdf/ResearchReportonGlobalAIGovernance.pdf

200 "Governing AI for Good and for All—Empowering Global Sustainable Development," World Internet Conference (世界互联网大会), April 13, 2025, accessed July 13, 2025, https://www.wicinternet.org/2025-04/13/c_1081925.htm; "WIC Seminar Discusses Global AI Safety and Governance Frameworks," World Internet Conference (世界互联网大会), June 21, 2025, accessed July 13, 2025, https://www.wicinternet.org/2025-06/21/c_1102213.htm

201 World Internet Conference, "What's the Right Way to Achieve "AI for Good"? ("AI 向善" 的正确打开方式是?)," Weixin Official Accounts Platform, December 3, 2024, accessed July 11, 2025, https://mp.weixin.qq.com/s/8yaoA1Dca69V71XAkDUq2w

202 “Zhongguancun Science Park,” Beijing Investment Promotion Service Center, November 20, 2024, accessed July 13, 2025, https://invest.beijing.gov.cn/english/Choose/Economic/202411/t20241120_3946665.html

203 “2025 Zhongguancun Forum Annual Conference to Be Held in Beijing: Advancing with Innovation, Cooperating for Mutual Benefis (2025 中关村论坛年会将在京举办向“新”而行合作共赢),” Government of the People’s Republic of China, accessed July 13, 2025, https://www.gov.cn/yaowen/liebiao/202503/content_7014788.htm

204 “Beijing Releases Collaborative Innovation Matrix for AI Safety and Governance (北京人工智能安全治理协同创新矩阵发布),” Beijing Daily (北京日报), March 30, 2025, accessed July 10, 2025, https://xinwen.bjd.com.cn/content/s67e9596fe4b068c68f11b70e.html

205 “World AI Conference 2023 (世界人工智能大会 2023),” accessed July 10, 2025, https://www.worldaic.com.cn/wangjie?year=2023

206 “Agenda of the 2023 World Internet Conference Wuzhen Summit (2023 年世界互联网大会乌镇峰会会议日程),” World Internet Conference (世界互联网大会), November 6, 2023, accessed July 10, 2025, https://cn.wicinternet.org/2023-11/06/content_36938351.htm

207 “2024 Forum Agenda (论坛日程),” Zhongguancun Forum (中关村论坛), April 25, 2025, accessed July 10, 2025, https://www.zgcforum.com.cn/zh2024/agenda/t2608

208 “WAIC 2024 Forums and Panels,” accessed July 10, 2025, https://online2024.worldaic.com.cn/forum

209 “2024 World Internet Conference Wuzhen Summit Agenda Overview (2024 年世界互联网大会乌镇峰会日程一览),” Xinhua, November 13, 2024, accessed July 10, 2025, http://www.zj.xinhuanet.com/20241113/7259b9c6a8ae47599acc10e876bdef5c/c.html

210 “2025 Forum Agenda (论坛日程),” Zhongguancun Forum (中关村论坛), accessed July 10, 2025, https://www.zgcforum.com.cn/agenda/t7903

211 “Xu Ye (许晔),” China Academy of Science and Technology for Development, accessed July 10, 2025, http://www.casted.org.cn/member/info/75

212 Chen Bokai (陈博凯), “Biosafety Risk Analysis of AI–Biotechnology Integration (人工智能生物技术融合的生物安全风险分析),” Weixin Official Accounts Platform, October 23, 2024, accessed July 10, 2025, https://mp.weixin.qq.com/s/k5YbK9dgQy1ksUwR-yNmtQ

213 “Gao Wanglai (高望来),” China Foreign Affairs University, accessed July 10, 2025, https://iir.cfau.edu.cn/col3577/col3578/60239.htm

214 “Academy of Military Medical Sciences (中国人民解放军军事科学院军事医学研究院),” Baidu Baike (百度百科), accessed July 10, 2025, https://baike.baidu.hk/item/%E4%B8%AD%E5%9B%BD%E4%BA%BA%E6%B0%91%E8%A7%A3%E6%94%BE%E5%86%9B%E5%86%9B%E4%BA%8B%E7%A7%91%E5%AD%A6%E9%99%A2%E5%86%9B%E4%BA%8B%E5%8C%BB%E5%AD%A6%E7%A0%94%E7%A9%B6%E9%99%A2/22117078

215 Yang Xi (杨溪) and Si Yuanyuan (司园园), “Biosafety Risks and Their Identification in the Context of AI (人工智能背景下生物安全风险及其识别),” *Journal of Intelligence (情报杂志)*, March 2025,

216 “Experts from the Center for Biosafety Strategy Invited to Yale University for Academic Discussion on the Intersection of AI and Biosafety (生物安全战略研究中心专家受邀在耶鲁大学展开 AI 与生物安全交叉问题学术讨论),” Tianjin University Center for

Biosafety Research and Strategy (天津大学生物安全战略研究中心), April 11, 2025, accessed July 10, 2025, https://tjusa.tju.edu.cn/info/1093/1993.htm; Yang Xi (杨溪) and Si Yuanyuan (司园园), "Biosafety Risks and Their Identification in the Context of AI (人工智能背景下生物安全风险及其识别)," *Journal of Intelligence (情报杂志)*, March 2025,

217 "Community Values, Guiding Principles, and Commitments for the Responsible Development of AI for Protein Design," March 8, 2024, accessed July 11, 2025, https://responsiblebiodesign.ai/

218 Sarah R. Carter et al., "The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe," The Nuclear Threat Initiative, October 2023, accessed July 10, 2025, https://www.nti.org/wp-content/uploads/2023/10/NTIBIO_AI_FINAL.pdf

219 Zhu Shu (朱姝) and Xu Ye (许晔), "Global Biosecurity Trends in the Post–Pandemic Era (后疫情时代全球生物安全新动向)," Weixin Official Accounts Platform, March 31, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/VbaeefyBh91p4ox9r42h2g

220 Gao Wanglai (高望来), "Beware the Risk of Biotechnology Proliferation in the Age of AI (警惕人工智能时代生物技术扩散风险)," Fujian Provincial Library (福建省图书馆), April 9, 2025, accessed July 10, 2025, https://www.fjlib.net/zt/fjstsgjcxx/xkj/202504/t20250409_477871.htm

221 Chen Bokai (陈博凯), "Biosafety Risk Analysis of AI–Biotechnology Integration (人工智能生物技术融合的生物安全风险分析)," Weixin Official Accounts Platform, October 23, 2024, accessed July 10, 2025, https://mp.weixin.qq.com/s/k5YbK9dgQy1ksUwR-yNmtQ; Gao Wanglai (高望来), "Beware the Risk of Biotechnology Proliferation in the Age of AI (警惕人工智能时代生物技术扩散风险)," Fujian Provincial Library (福建省图书馆), April 9, 2025, accessed July 10, 2025, https://www.fjlib.net/zt/fjstsgjcxx/xkj/202504/t20250409_477871.htm

222 "Biosecurity Law of the P.R.C.," China Law Translate, October 17, 2020, accessed July 10, 2025, https://www.chinalawtranslate.com/biosecurity-law/

223 Yang Xi (杨溪) and Si Yuanyuan (司园园), "Biosafety Risks and Their Identification in the Context of AI (人工智能背景下生物安全风险及其识别)," *Journal of Intelligence (情报杂志)*, March 2025,

224 "AI Attack Mutation Rate Reaches 93% Every 24 Hours, Global AI Security Losses Near $23.5 Billion: How Can the Offense–Defense Stalemate Be Broken? (AI 攻击变异率每 24 小时达 93% 全球 AI 安全损失逼近 235 亿美元：攻防博弈如何破局？)," Sina, May 25, 2025, accessed July 10, 2025, https://finance.sina.com.cn/roll/2025-05-25/doc-inexuivy3844484.shtml

225 Yuan Weiguo (苑卫国), Zhang Xinyue (张新跃), and Yuchi Xuebiao (尉迟学彪), "Analysis of the Impact of Generative AI Technology on Cybersecurity and Policy Recommendations (生成式人工智能技术对网络安全领域的影响分析与启示建议)," Weixin Official Accounts Platform, February 28, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/Xnn5qlGP-o5G6PG0n3XZxg

226 "Zhou Hongyi Discusses the Harbin Asian Winter Games Being Targeted by NSA: Possibly the First AI Agent–Driven Cyberattack in Human History (周鸿祎谈哈尔滨亚冬会遭美国安局攻击：可能是人类首次用 AI 智能体发起的网络攻击)," Yicai (第一财经), April 15, 2025, accessed July 10, 2025, https://www.yicai.com/news/102571613.html

227 "AI Attack Mutation Rate Reaches 93% Every 24 Hours, Global AI Security Losses Near $23.5 Billion: How Can the Offense–Defense Stalemate Be Broken? (AI 攻击变异率每 24 小时达 93% 全球 AI 安全损失逼近 235 亿美元：攻防博弈如何破局？)," Sina, May 25, 2025, accessed July 10, 2025, https://finance.sina.com.cn/roll/2025-05-25/doc-inexuivy3844484.shtml

228 "To Address AI-Driven Security Challenges, Industry Experts Offer These Recommendations (应对 AI 带来的安全挑战，业内专家给出这些建议)," Yicai (第一财经), June 5, 2024, accessed July 13, 2025, https://www.yicai.com/news/102138875.html

229 Wang Ying (王鹰), "AI and Security Confidentiality Series (Part 1) | A Brief Discussion on the Risks and Challenges AI Brings to Cybersecurity (人工智能与安全保密系列谈（一）| 浅谈人工智能给网络安全带来的风险与挑战)," Weixin Official Accounts Platform, March 10, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/ilyqDmxk6wm2HPtTSDblOg

230 Chen Ling (陈凌), "Viewing "AI for Good" Through Open Source (Commentator's Observation) (从开源看 "智能向善"（评论员观察）)," People's Daily, June 18, 2025, accessed July 10, 2025, http://opinion.people.com.cn/n1/2025/0618/c1003-40502848.html

231 Zhang Linghan (张凌寒) and He Jiaxin (何佳欣), "Legal Safeguards for Responsible Innovation in Open-Source AI (开源人工智能负责任创新的法律保障)," Weixin Official Accounts Platform, May 23, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/DA25AlNinJLEiPc2zEWBDQ; Zhou Hui (周辉) et al, "AI Model Law 3.0 (人工智能示范法 3.0)," Weixin Official Accounts Platform, March 29, 2023, accessed July 10, 2025, https://mp.weixin.qq.com/s/pCC_AM5mpU7QY-x14R-kZw

232 China Academy of Information and Communications Technology, "DeepSeek Open-Source Revelation (Part 3): Typical Application Risks of Open-Source AI and Response Strategies (DeepSeek 开源启示录（三）开源 AI 的典型应用风险和应对策略)," Weixin Official Accounts Platform, February 5, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/BT7Ecx7sjqY5BUZFFEdJDw; China Academy of Information and Communications Technology and Open Source Cloud Alliance for Industry, "开源大模型应用指南 1.0 （风险治理篇)," November 2024, https://gitee.com/trustworthy-open-source-community/ZhiYuan/raw/master/%E5%BC%80%E6%BA%90%E6%B2%BB%E7%90%86/%E5%BC%80%E6%BA%90%E5%A4%A7%E6%A8%A1%E5%9E%8B%E5%BA%94%E7%94%A8%E6%8C%87%E5%8D%971.0.pdf

233 Hongyu Fu (傅宏宇), "DeepSeek Showcases Innovative Reforms in Risk Governance for Open-Source Large Models (开源大模型风险治理机制的改革与创新——以 DeepSeek 为例)," Weixin Official Accounts Platform, March 10, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/_fJGDtBBYLnOExYMUyFvEw

234 Tencent Research Institute (腾讯研究院), "Between Freedom and Order, Building a "Safe Harbor" for Open-Source Foundation Models (在自由与秩序之间，为开源大模型搭建"避风港")," Weixin Official Accounts Platform, March 26, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/_jaHCXTu5nJP4xCpSA5liQ

235 Zhang Linghan (张凌寒) and He Jiaxin (何佳欣), "Legal Safeguards for Responsible Innovation in Open-Source AI (开源人工智能负责任创新的法律保障)," Weixin Official Accounts Platform, May 23, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/DA25AlNinJLEiPc2zEWBDQ

236 "China's Zhipu AI and OpenAI Join Others in Signing Frontier AI Safety Commitment (中国智谱 AI 与 OpenAI 等公司共同签署前沿人工智能安全承诺)," Chinadaily, May 22, 2024, accessed July 8, 2025, https://cn.chinadaily.com.cn/a/202405/22/WS664ddd1da3109f7860ddf05e.html

237 "Roundtable with Four Top AI Company CEOs: China-US Competition, Price Wars, Involution, and the Future（四位顶级 AI 企业 CEO 圆桌实录：关于中美竞争、价格战、内卷与未来)," Weixin Official Accounts Platform, June 14, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/gXF0sfAut_PTzCTN9UFtIw

238 "Alibaba Cloud's Ouyang Xin: A New Security Paradigm in the AI Era (阿里云欧阳欣：AI 时代下的安全新范式-阿里云开发者社区)," Alibaba Cloud, September 24, 2024, accessed July 8, 2025, https://developer.aliyun.com/article/1610964

239 "Frontier Model Forum," Frontier Model Forum, accessed July 8, 2025, https://www.frontiermodelforum.org/; "360 Qihoo Named Annual Outstanding Contribution Member by China Artificial Intelligence Industry Alliance (360 获评中国人工智能产业发展联盟年度突出贡献成员单位)," Tech China (中国网科技), April 14, 2025, accessed July 8, 2025, https://www.huanqiu.com/article/4MH0pJEs66F

240 Artificial Intelligence Industry Alliance, "AIIA Policy and Law Working Group's Transition Meeting, and "Risks and Legal Regulation of General Artificial Intelligence" Forum Successfully Convened (政策法规工作组换届工作会暨 "通用人工智能风险与法律规制" 论坛成功召开)," Weixin Official Accounts Platform, January 22, 2024, accessed July 8, 2025, https://mp.weixin.qq.co

m/s/4SVCI-4ovV77XefpwkDjSA; Artificial Intelligence Industry Alliance (人工智能产业发展联盟（AIIA）), "Preventing Risks, Safeguarding the Future: AI Risk Management System Officially Released (防范风险，守护未来："人工智能风险管理体系"正式发布)," Weixin Official Accounts Platform, December 26, 2023, accessed July 8, 2025, https://mp.weixin.qq.com/s/wfCAEHY_hryA8Rr9L8MHMQ; China Academy of Information and Communications Technology, "First CAICT AI Agent Security Seminar Successfully Held (中国信通院智能体（AI Agent）安全首次研讨会顺利举办)," Weixin Official Accounts Platform, October 31, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/5sHYxcTK_7SBIVsiz04p8g; China Academy of Information and Communications Technology, "Notice on Holding the Second Seminar of the AI Agent Security Evaluation Standards Series (关于召开智能体安全系列评估规范第二次研讨会的通知)," Weixin Official Accounts Platform, November 21, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/0DhPU9dmgNBOvPkwHg_jGQ

241 Artificial Intelligence Industry Alliance (人工智能产业发展联盟（AIIA）), "Notice on Establishing the AIIA "Science and Technology Ethics Working Group" and Launching First-Round Member Applications (关于筹备成立 AIIA"科技伦理工作组" 并启动首批成员单位申报工作的通知)," Weixin Official Accounts Platform, December 23, 2023, accessed July 8, 2025, https://mp.weixin.qq.com/s/mv6Nt-XHldAmaIWM-CiA8Q; Artificial Intelligence Industry Alliance, "AIIA Policy and Law Working Group's Transition Meeting, and "Risks and Legal Regulation of General Artificial Intelligence" Forum Successfully Convened (政策法规工作组换届工作会暨 "通用人工智能风险与法律规制" 论坛成功召开)," Weixin Official Accounts Platform, January 22, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/4SVCI-4ovV77XefpwkDjSA

242 China Academy of Information and Communications Technology, "First 17 Companies Sign Landmark "Artificial Intelligence Safety Commitments" Setting a New Standard for Industry Self-Regulation (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》)," Weixin Official Accounts Platform, December 24, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng

243 "Frontier AI Safety Commitments, AI Seoul Summit 2024," GOV.UK, accessed July 8, 2025, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

244 METR, "Frontier AI Safety Policies," accessed July 8, 2025, https://metr.org/faisc

245 China Academy of Information and Communications Technology, "AIIA Launches Self-Regulated Safety-Measures Disclosure Initiative to Uphold the "Artificial Intelligence Safety Commitments" (践行《人工智能安全承诺》，AIIA 开启安全措施自律披露行动)," Weixin Official Accounts Platform, February 26, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/WxeD9Dk1KKcszHiknpZH-Q

246 "LLM Leaderboard - Comparison of over 100 AI Models from OpenAI, Google, DeepSeek & Others," Artificial Analysis, accessed July 8, 2025, https://artificialanalysis.ai/leaderboards/models

247 "OpenCompass Evaluation Rankings (OpenCompass 司南 - 评测榜单)," OpenCompass, accessed July 8, 2025, https://rank.opencompass.org.cn/home

248 "Overview Leaderboard," LMArena, July 8, 2025, accessed July 8, 2025, https://lmarena.ai/leaderboard

249 Zhang Shuai (张帅), "Glut of AI Compute Centers, Shortage of Large Models (智算中心太 "多"，大模型不够用了)," Weixin Official Accounts Platform, November 20, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/Mpy-AbQmu5y2PLXfyNxkFw

250 "Kai-Fu Lee Sets the Record Straight on 01.AI's Pivot," 36Kr, January 9, 2025, accessed July 8, 2025, https://kr-asia.com/kai-fu-lee-sets-the-record-straight-on-01-ais-pivot

251 Alan Wake et al., "Yi-Lightning Technical Report," January 22, 2025, accessed July 8, 2025, arXiv: 2412.01253 [cs], http://arxiv.org/abs/2412.01253; MiniMax et al., "MiniMax-01: Scaling Foundation Models with Lightning Attention," January 14, 2025, accessed July 8, 2025, arXiv: 2501.08313 [cs], http://arxiv.org/abs/2501.08313; Team GLM et al., "ChatGLM: A Family of Large Language

Models from GLM-130B to GLM-4 All Tools," July 30, 2024, accessed July 8, 2025, arXiv: 2406.12793 [cs], http://arxiv.org/abs/2406.12793

252 DeepSeek-AI et al., "DeepSeek-V3 Technical Report," February 18, 2025, accessed July 8, 2025, arXiv: 2412.19437 [cs], http://arxiv.org/abs/2412.19437; MiniMax et al., "MiniMax-01: Scaling Foundation Models with Lightning Attention," January 14, 2025, accessed July 8, 2025, arXiv: 2501.08313 [cs], http://arxiv.org/abs/2501.08313; Team GLM et al., "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools," July 30, 2024, accessed July 8, 2025, arXiv: 2406.12793 [cs], http://arxiv.org/abs/2406.12793

253 Grattafiori Aaron et al, "The Llama 3 Herd of Models," November 23, 2024, arXiv: 2407.21783, http://arxiv.org/abs/2407.21783; Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," May 2025, https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf

254 Alan Wake et al., "Yi-Lightning Technical Report," January 22, 2025, accessed July 8, 2025, arXiv: 2412.01253 [cs], http://arxiv.org/abs/2412.01253

255 360Zhinao Team, "360Zhinao Technical Report," May 22, 2024, accessed July 8, 2025, arXiv: 2405.13386 [cs], http://arxiv.org/abs/2405.13386

256 An Yang et al., "Qwen3 Technical Report," May 14, 2025, accessed July 8, 2025, arXiv: 2505.09388 [cs], http://arxiv.org/abs/2505.09388

257 ERNIE Team, Baidu, "ERNIE 4.5 Technical Report," June 29, 2025, https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

258 Bytedance Seed, "Doubao 1.5pro," January 22, 2025, accessed July 8, 2025, https://seed.bytedance.com/zh/special/doubao_1_5_pro

259 Dong Guo et al., "Seed1.5-VL Technical Report," May 11, 2025, accessed July 8, 2025, arXiv: 2505.07062 [cs], http://arxiv.org/abs/2505.07062

260 DeepSeek-AI et al., "DeepSeek-V3 Technical Report," February 18, 2025, accessed July 8, 2025, arXiv: 2412.19437 [cs], http://arxiv.org/abs/2412.19437

261 Nathan Lambert et al., "RewardBench: Evaluating Reward Models for Language Modeling," June 8, 2024, accessed July 8, 2025, arXiv: 2403.13787 [cs], http://arxiv.org/abs/2403.13787

262 DeepSeek-AI et al., "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," January 22, 2025, accessed July 8, 2025, arXiv: 2501.12948 [cs], http://arxiv.org/abs/2501.12948

263 MiniMax et al., "MiniMax-01: Scaling Foundation Models with Lightning Attention," January 14, 2025, accessed July 8, 2025, arXiv: 2501.08313 [cs], http://arxiv.org/abs/2501.08313

264 Kimi Team et al., "Kimi K1.5: Scaling Reinforcement Learning with LLMs," June 3, 2025, accessed July 8, 2025, arXiv: 2501.12599 [cs], http://arxiv.org/abs/2501.12599

265 Kimi Team et al., "Kimi-VL Technical Report," June 23, 2025, accessed July 8, 2025, arXiv: 2504.07491 [cs], http://arxiv.org/abs/2504.07491

266 Xingwu Sun et al., "Hunyuan-Large: An Open-Source MoE Model with 52 Billion Activated Parameters by Tencent," November 6, 2024, accessed July 8, 2025, arXiv: 2411.02265 [cs], http://arxiv.org/abs/2411.02265

267 Tencent Hunyuan Team et al., "Hunyuan-TurboS: Advancing Large Language Models through Mamba-Transformer Synergy and Adaptive Chain-of-Thought," July 4, 2025, accessed July 8, 2025, arXiv: 2505.15431 [cs], http://arxiv.org/abs/2505.15431

268 Team GLM et al., "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools," July 30, 2024, accessed July 8, 2025, arXiv: 2406.12793 [cs], http://arxiv.org/abs/2406.12793

269 Zhexin Zhang et al., "SafetyBench: Evaluating the Safety of Large Language Models," June 24, 2024, accessed July 8, 2025, arXiv: 2309.07045 [cs], http://arxiv.org/abs/2309.07045

270 Xiao Liu et al., "AutoGLM: Autonomous Foundation Agents for GUIs," October 28, 2024, accessed July 8, 2025, arXiv: 2411.00820 [cs], http://arxiv.org/abs/2411.00820

271 Alibaba AI Governance Lab, "Chapter 6: Systematic Safety Governance Capacity Safeguards Stable Growth—"Large-Model Technology Development and Governance Practice Report" (第六章: 体系化的安全治理能力是稳定发展的保障《大模型技术发展及治理实践报告》)," Weixin Official Accounts Platform, July 8, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/0PwkLS0AI3uo-ayoYnlasA

272 Baidu Safety 百度安全, "Baidu Showcases Large-Model Content Safety Compliance Efforts at the 2024 Baidu AI Cloud Summit (2024 百度云智大会 | 百度大模型内容安全合规探索与实践)," Weixin Official Accounts Platform, October 9, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/n4YTjU5qKvcp89HiLYk1zQ

273 Tencent Security (腾讯安全), "Expert Commentary: Large Models Are Redefining the Cybersecurity Battlefield—An Era of AI-versus-AI Await (【专家谈】大模型重构网络安全战场，未来是 AI 对抗 AI 的时代)," Weixin Official Accounts Platform, April 2, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/IxTQmOUt4c7TR4YcvQJt-g; Tencent Security (腾讯安全), "Tencent Cloud Unveils Generative AI Security Suite to Safeguard Enterprise Data and Content (腾讯云发布生成式 AI 安全解决方案，助力企业守好"数据"和"内容"安全关)," Weixin Official Accounts Platform, May 17, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/bMrlxH93L4tqBfXVwKep-g

274 Alibaba Group, "Responsible Technology, Sustainable Future," accessed July 8, 2025, https://www.alibabagroup.com/esg

275 SenseTime, "SenseTime Group 2024 Sustainability Report," https://www1.hkexnews.hk/listedco/listconews/sehk/2025/0424/2025042401051.pdf

276 "Trial Measures for Science and Technology Ethics Review (科技伦理审查办法 [试行])," Ministry of Science and Technology of the People's Republic of China, September 7, 2023, accessed July 8, 2025, https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm

277 China Quality Certification Centre, "CQC Issued the First Artificial Intelligence Management System Certificate to Alibaba Cloud," September 12, 2024, accessed July 8, 2025, https://www.cqc.com.cn/www/english/c/2024-09-12/597561.shtml; "Baidu ESG Reports," Baidu, accessed July 8, 2025, https://esg.baidu.com/ESGReport; China Electronic Product Reliability & Environmental Testing Research Institute, "iFlytek Receives Authoritative Certification, Embracing the New Era of AI (科大讯飞获权威认证，拥抱 AI 新时代!)," Weixin Official Accounts Platform, November 26, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/YX7JjEJOA-I7uO15bu6z-g; China Electronics Standardization Institute, "CESI Certification Grants SenseTime ISO/IEC 42001 AI Management System Certificate (赛西认证为商汤科技颁发 ISO/IEC 42001 人工智能管理体系认证证书)," Weixin Official Accounts Platform, March 21, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/8fMRbQuX2vGfcTcG1mha-A; "Zhipu Receives ISO/IEC 42001:2023 Artificial-Intelligence Management-System Certificate (智谱获颁 ISO/IEC 42001: 2023 人工智能管理体系认证证书)," 163, July 31, 2024, accessed July 8, 2025, https://www.163.com/dy/article/J8E9DL1T0511RIVP.html

278 Anthropic, "Anthropic Achieves ISO 42001 Certification for Responsible AI," January 14, 2025, accessed July 8, 2025, https://www.anthropic.com/news/anthropic-achieves-iso-42001-certification-for-responsible-ai

279 Hongyu Fu (傅宏宇), "DeepSeek Showcases Innovative Reforms in Risk Governance for Open-Source Large Models (开源大模型风险治理机制的改革与创新——以 DeepSeek 为例)," Weixin Official Accounts Platform, March 10, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/_fJGDtBBYLnOExYMUyFvEw; Tencent Research Institute (腾讯研究院), "Between Freedom and Order, Building a "Safe Harbor" for Open-Source Foundation Models (在自由与秩序之间，为开源大模型搭建"避风港")," Weixin Official Accounts Platform, March 26, 2025, accessed July 10, 2025, https://mp.weixin.qq.com/s/_iaHCXTu5nJP4xCpSA5liQ

280 June Yoon, "Why China Is Suddenly Flooding the Market with Powerful AI Models," *Financial Time*s, March 19, 2025, accessed July 8, 2025, https://www.ft.com/content/13df6250-dffb-40fc-bb79-309764fa3905

281 Tianhao Li et al., "SciSafeEval: A Comprehensive Benchmark for Safety Alignment of Large Language Models in Scientific Tasks," December 16, 2024, accessed July 9, 2025, arXiv: 2410.03769 [cs], http://arxiv.org/abs/2410.03769

282 Wenxuan Wang et al., "Can't See the Forest for the Trees: Benchmarking Multimodal Safety Awareness for Multimodal LLMs," June 3, 2025, accessed July 8, 2025, arXiv: 2502.11184 [cs], http://arxiv.org/abs/2502.11184

283 Concordia AI, "China's AI Safety Evaluations Ecosystem," AI Safety in China, September 13, 2024, accessed July 17, 2025, https://aisafetychina.substack.com/p/chinas-ai-safety-evaluations-ecosystem

284 "OpenCompass Evaluation Rankings (OpenCompass 司南 - 评测榜单)," OpenCompass, accessed July 8, 2025, https://rank.opencompass.org.cn/home

285 "Security Services: One-Stop Foundation Model Security Solutions (安全服务一站式大模型安全服务)," AI45 Safety Ecosystem Platform (AI45 安全生态平台), accessed July 17, 2025, https://ai45.shlab.org.cn/service

286 Zaibin Zhang et al., "PsySafe: A Comprehensive Framework for Psychological-based Attack, Defense, and Evaluation of Multi-agent System Safety," August 20, 2024, accessed July 17, 2025, arXiv: 2401.11880 [cs], http://arxiv.org/abs/2401.11880; Steffi Chern et al., "BeHonest: Benchmarking Honesty in Large Language Models," July 8, 2024, accessed July 17, 2025, arXiv: 2406.13261 [cs], http://arxiv.org/abs/2406.13261

287 China Academy of Information and Communications Technology, "AI Safety Benchmark 权威大模型安全基准测试首轮结果正式发布 (AI Safety Benchmark 权威大模型安全基准测试首轮结果正式发布)," Weixin Official Accounts Platform, April 10, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/3FcLBHCy_oVaaj-2Ca9zag

288 China Academy of Information and Communications Technology, "AI Safety Benchmark: Q1 2025 Large-Model Hallucination Test Results Released (AI Safety Benchmark 大模型幻觉测试 2025 Q1 版结果发布)," Weixin Official Accounts Platform, May 8, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/0JUeL3TOLocXJ1GkXOJg1Q

289 China Academy of Information and Communications Technology, "Ant Digital Technologies Passes CAICT's "Trustworthy AI (Safety Governance)" Foundation Model Security Audit Capability Assessment (蚂蚁数科通过中国信通院 "可信 AI（安全治理）" 大模型安全审核能力评估)," Weixin Official Accounts Platform, July 3, 2025, accessed July 17, 2025, https://mp.weixin.qq.com/s/Lbu4DVij_HYr_6lG495w3g

290 "AI Safety Testing and Evaluation System and Risk Knowledge Base Released (人工智能安全测试评价体系及风险知识库发布)," China Software Testing Center (中国软件评测中心), June 27, 2025, accessed July 17, 2025, https://www.cstc.org.cn/info/1081/254814.htm

291 "FlagEval Foundation Model Evaluation Platform (FlagEval 大模型评测平台)," Beijing Academy of AI, accessed July 17, 2025, https://flageval.baai.ac.cn/#/home

292 “Shanghai AI Lab Elected as Chair of the Large Model Subgroup under China’s National AI Standardization Main Group (上海人工智能实验室当选国家人工智能标准化总体组大模型专题组组长),” Shanghai AI Lab (上海人工智能实验室), July 7, 2023, accessed July 17, 2025, http://www.shlab.org.cn/news/5443434

293 Baidu Safety, “Large-Model Security Solution (大模型安全解决方案),” accessed July 8, 2025, https://anquan.baidu.com/product/llmsec; Volcengine, “AIGC Content Safety Solutions (AIGC 内容安全方案),” accessed July 8, 2025, https://www.volcengine.com/product/AIGC-content; “AIGC Content Risk Management Solution AIGC (风控方案),” NetEase (网易), accessed July 8, 2025, https://dun.163.com/solution/aigc; Next Data, “AIGC Content Compliance Solution (AIGC 内容合规解决方案),” accessed July 8, 2025, https://www.ishumei.com/new/product/solution/aigc

294 “360Zhinao Website,” accessed July 8, 2025, https://ai.360.com/; “Qi An Xin’s Qi Xiangdong: Three Security Strategies to Defuse AI’s “Triple Crisis” (奇安信齐向东：用三大安全策略化解 AI“三重危机”),” Science and Technology Daily, November 22, 2024, accessed July 8, 2025, https://www.stdaily.com/web/gdxw/2024-11/22/content_262536.html

295 Shuxian Zong (宗淑贤), “Interview with 360 Founder Zhou Hongyi: Fighting Magic with Magic—Large-Model Security Demands Large-Model Solutions (对话 360 周鸿祎：魔法对付魔法，大模型安全问题得靠大模型),” Weixin Official Accounts Platform, August 6, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/pb23UQhlf_1iC3j3g5KPSQ

296 “BotSmart 博特智能,” accessed July 8, 2025, https://www.botsmart.cn/; “RealAI,” accessed July 8, 2025, https://www.realai.ai/about

297 I-AIIG, “AI Industry Development and Governance Sub-Forum of the International AI Cooperation and Governance Forum 2024 Successfully Held at the National University of Singapore (2024 人工智能合作与治理国际论坛 “人工智能产业发展与治理” 专题论坛在新加坡国立大学成功举办),” Weixin Official Accounts Platform, December 3, 2024, accessed July 8, 2025, https://mp.weixin.qq.com/s/v2PTr2uu3b4p1bUef9G9Xw

298 RealAI, “RealAI’s Tian Tian: “Safety and Controllability Are the Prerequisite for Unlocking AI Productivity” (瑞莱智慧田天：安全可控是 AI 生产力释放的前置门槛),” Weixin Official Accounts Platform, April 17, 2025, accessed July 8, 2025, https://mp.weixin.qq.com/s/e7bVh6DZw2X5fbtbG5Vadw

299 “AI Foundation Model Safety Governance Seminar and AI Safety Working Group Inauguration Successfully Held (AI 大模型安全治理研讨会暨 AI 安全工作组成立仪式顺利举办),” China Software Testing Center (中国软件评测中心), April 12, 2024, accessed July 17, 2025, https://www.cstc.org.cn/info/1081/251145.htm

300 Future of Life Institute, “FLI AI Safety Index 2024,” December 11, 2024, https://futureoflife.org/wp-content/uploads/2024/12/AI-Safety-Index-2024-Full-Report-27-May-25.pdf

301 SenseTime, “SenseTime Group 2024 Sustainability Report,” https://www1.hkexnews.hk/listedco/listconews/sehk/2025/0424/2025042401051.pdf