安远 AI
CONCORDIA AI

State of AI Safety in China

July 2026

Authors

This report was authored by Gabriel Wagner, Erik Lindblad, Kwan Yee Ng, and Brian Tse.

Explore the interactive companion resources:

https://aisafetychina.com/

Acknowledgements

We are grateful to Charles Alaimo, Alan Chan, Jeffrey Ding, Karson Elmgren, Emmie Hine, Kristy Loke, Zilan Qian, Huw Roberts, Matt Sheehan, Ziyue Wang, James Zhang, Jie Zhang, and Jason Zhou for providing feedback and suggestions on this report.

How to cite this report

Gabriel Wagner, Erik Lindblad, Kwan Yee Ng, and Brian Tse, "State of AI Safety in China (2026)," Concordia AI, July 2026, https://aisafetychina.com/assets/State-of-AI-Safety-in-China-2026.pdf.

Table of Contents

Executive Summary **1**

Introduction and scope **5**

1 Domestic Governance **8**
1.1 High-level policy . . . 9
1.2 Laws and regulations . . . 12
1.3 Standards . . . 18

2 International Governance **33**
2.1 Multilateral engagement . . . 33
2.2 Bilateral engagement . . . 37
2.3 International AI standards . . . 40
2.4 Non-official dialogues . . . 42

3 Technical Safety Research **48**
3.1 Methodology . . . 48
3.2 Overall trends . . . 49
3.3 Key research groups . . . 49
3.4 Specific research contributions . . . 52

4 Expert Views on AI Safety and Governance **59**
4.1 Methodology . . . 59
4.2 Safety risks from agentic AI . . . 60
4.3 Open source AI governance . . . 62
4.4 AI's impact on cybersecurity . . . 64
4.5 AI's risks for biosecurity . . . 65

5 Industry Governance **68**
5.1 Industry-wide collective action . . . 69
5.2 Company-specific safety practices . . . 71

Conclusion **79**

Notes **81**

Executive Summary

As AI capabilities advance and agentic systems move from demonstrations into wide deployment, understanding how China approaches AI safety and governance remains essential to any serious effort at international coordination. Since 2023, Concordia AI's annual reports on the *State of AI Safety in China* have tracked how China addresses risks from general-purpose AI.

This year's report provides updates from July 2025 to June 2026 across five domains: domestic governance, international governance, technical safety research, expert views on AI safety and governance, and industry governance.

Over the three years this report series has tracked, it has become increasingly clear that China's approach to AI safety and governance extends far beyond content control. Much of China's early AI regulation in 2023 focused on what AI *says*. Throughout 2024 and 2025, the rise of multimodal models moved the regulatory focus toward images, videos, and audio, prompting strict rules around labeling AI-generated content. With the rise of autonomous agents and AI companion products over the past year, the concern has expanded to what AI *does* and what it *does to people and society*, from mental-health impacts to labor-market disruptions.

Domestic Governance

- **The 15th Five-Year Plan (2026–2030) cements "AI Plus" — the drive to diffuse AI across sectors to spur economic growth — as China's overarching AI policy, while continuing to prioritize safety and risk management.** Enshrined in March 2026, the plan carries forward President Xi Jinping's April 2025 language on the importance of early risk warning and emergency response. It also highlights new concerns around AI's impact on employment, with calls for monitoring AI's labor-market impact, strengthening reskilling programs and employment support, and leveraging AI's ability to create jobs.
- **The arrival of autonomous agents reoriented Chinese governance from controlling what AI says to controlling what it does.** After the open source agent OpenClaw proliferated in early 2026, multiple cybersecurity authorities issued warnings, and in May 2026 the Cyberspace Administration of China (CAC) and two other agencies issued dedicated guidance on agentic AI, devoting a full chapter to safety and proposing risk-based governance. Several major AI standard-setting bodies are now drafting agent-related security standards. This marks a shift in China's framework from content control toward action control.

- **Binding rules and standards increasingly target specific AI risks that go beyond political content control.** New *Interim Measures on Anthropomorphic AI Interaction Services*, effective July 2026, impose obligations on AI companion services around suicide intervention, addiction prevention, and protection of minors and the elderly. New AI ethics review rules require institutions to establish registered ethics committees, with a ten-province pilot running June–November 2026. Meanwhile, a series of standards operationalizes requirements for labeling and watermarking AI-generated content disseminated online. Together these rules show China increasingly confronting AI governance problems shared with other jurisdictions, rather than just content-control issues that historically dominated its AI regulations.
- **Frontier risks are being discussed more across China's governance layers, even as concrete requirements remain limited.** Senior officials including President Xi Jinping have referenced "risks of technological loss of control," and a major standard-setting body has flagged the need for "circuit breakers" and "safety stop switches." Most concretely, the draft *Cybercrime Law* would require AI service providers to monitor for bulk generation of malicious code and report related incidents to authorities. Meanwhile, chemical, biological, radiological, and nuclear (CBRN) misuse appeared in a national standard for the first time.
- **You can also explore China's domestic AI governance in our "China AI Safety Policy-Risk Matrix" on the report companion website.** This interactive overview maps China's quickly growing web of AI safety policies to the specific risks they address.[a]

International Governance

- **China is positioning itself as an architect of multilateral AI governance.** China supported two new United Nations (UN) mechanisms: the Independent International Scientific Panel on AI, to which two Chinese experts were appointed, and the Global Dialogue on AI Governance, which will convene for the first time in July 2026. China also proposed a World AI Cooperation Organization (WAICO), though the timeline, leadership, and membership remained unconfirmed as of June 2026. China's UN-centered approach contrasts sharply with the US' growing skepticism toward multilateral AI governance.
- **China and the United States announced the launch of an intergovernmental AI dialogue, ending a two-year freeze in official bilateral engagement on AI safety.** The agreement followed President Trump's May 2026 visit to Beijing. Official bilateral channels with other countries featured relatively little discussion of AI safety. However, non-official dialogues between groups in China and the West produced increasingly substantive AI safety outputs, including on biosecurity risks, misuse by non-state actors, and updated bilingual glossaries.

a. https://aisafetychina.com/matrix/

Technical AI Safety Research

- **Chinese frontier AI safety research output grew substantially, and agent safety is now the single most active area.** Total monthly output rose roughly 60%, from around 36 papers/month in June 2025 to 57 in April 2026. Agent safety was the topic of around 27% of new papers in Q1 2026 — up from 8% in early 2025 — spanning operational failures, alignment challenges, and a smaller but growing body of work on loss-of-control risks such as self-replication and multi-agent collusion.
- **Some research directions became more prominent, while other areas' output was similar to the previous year.** AI-generated content detection and watermarking remained a consistently high-output area (11–13% of papers each quarter). Mechanistic interpretability — the ability to explain internal workings of AI systems — grew from a niche topic to roughly 15% of output by Q1 2026. CBRN risks are a standard component of major Chinese safety frameworks, but dedicated research on the topic remains rare.
- **Research output remains concentrated in universities and state-backed labs.** Of 28 "key research groups" with a particularly strong focus on AI safety, the large majority sit at universities (20), especially at top schools like Tsinghua University, Peking University, and Fudan University. While only three groups are at state-backed labs, their research output is disproportionately high, with Shanghai AI Lab standing out, co-authoring around one in ten papers. Of the remaining key research groups, three are at private companies, one is a state-owned enterprise, and one is a university-industry joint group. Within industry, over half of all safety output comes from just six big-tech firms — Alibaba, Ant Group, Huawei, Tencent, China Telecom, and ByteDance.

Expert Views

- **Agentic AI safety emerged as a distinct strand of Chinese expert discourse.** Prominent Chinese experts warned that human oversight could be eroded as agents gain autonomy. Experts also discussed potential risks associated with recursive self-improvement as well as giving agents physical instantiation (e.g. robotics).
- **Chinese experts deepened their analyses of AI risks in cybersecurity, biosecurity, and open source AI.** Cybersecurity debates were made more salient by the release of US frontier models that were reported to autonomously discover and exploit zero-day vulnerabilities. The reactions tended to emphasize using AI to improve defensive cybersecurity capabilities, rather than creating new regulation for AI itself. On biosecurity, experts assessed risks from biological design tools and autonomous laboratories. Chinese experts remained broadly favorable toward open source AI, though several proposed building capacity for tighter oversight, including risk-assessment centers.

Industry Governance

- **Industry's collective safety efforts pivoted toward agents and, for the first time, toward frontier risks.** The AI Industry Alliance of China (AIIA) instigated voluntary safety commitments on agentic consumer products in February 2026, followed by commitments on cloud-based agents in April 2026. This builds on a robust set of AI agent security publications from Chinese AI and cloud companies. AIIA's voluntary testing program with a government-affiliated think tank is increasingly focusing on frontier risks. In November 2025, the program released cybersecurity misuse evaluations of 15 open source coding models. It then launched the AI Safety Benchmark 2.0, adding new categories for model deception, loss of control, dangerous-domain misuse, and agentic risks.
- **Company-level public transparency on safety remains thin and inconsistent, lagging well behind Western peers.** Only five of ten leading foundation-model developers reported safety evaluation results alongside any release this past year, and none did so consistently; DeepSeek-R1's peer-reviewed *Nature* paper and Moonshot AI's Kimi K2 model card were the most detailed disclosures yet, but flagship follow-ups like DeepSeek-V4 and Kimi K2.5 were released with no safety evaluation results at all.

Introduction and scope

In February 2026, OpenClaw spread through China's developer community like wildfire. The open source agentic framework was adopted so quickly that it forced emergency warnings from multiple government bodies as they clamped down on unauthorized data deletion and other incidents.

Within weeks, every major Chinese AI standards body was drafting agent-security standards. Financial institutions were reportedly ordered to restrict usage. Safety guidelines issued by seemingly obscure cybersecurity authorities topped the trending charts on Chinese social media. OpenClaw was a vivid instance of a broader pattern. AI has changed from just being able to "say" things to being able to "do" things. The risks of increasingly autonomous agentic AI are a running theme throughout this report. It is a trend that has driven active policy-making and standard-setting, international expert agreements, and a sharp rise in related technical research from Chinese institutions.

It is not only agent harnesses that have grown more capable; the underlying models have advanced as well. As Concordia AI's own independent evaluations of frontier models from China and elsewhere confirm, AI systems now match or exceed expert performance on benchmarks that test knowledge and skills relevant to biological weapons development.[1] In April 2026, Anthropic announced Claude Mythos Preview, a frontier model that reportedly identified thousands of previously unknown software vulnerabilities across widely used systems.[2] Anthropic declined to release the model publicly, citing cybersecurity misuse concerns.

2026 also carries special importance on the Chinese political calendar, marking the release of the 15th Five-Year Plan (2026–30). Not only does the plan cement China's diffusion-focused "AI Plus" framework as the country's overarching AI policy through 2030, it also calls for "full lifecycle risk management" for AI and urges the introduction of frameworks covering "safety monitoring, risk early warning, and emergency response." That top-level policy endorsement suggests that interest in AI risk management and preparedness for AI-enabled emergencies is likely to remain high over the coming years.

Internationally, engagement is gathering momentum. China and the United States announced in May 2026 the launch of a government-to-government dialogue on AI. The dialogue would end a two-year freeze on direct bilateral engagement on AI and potentially open the way to more substantive cooperation on AI safety and governance. China has also advocated for the UN to have a central role in global AI governance. It has supported two new UN mechanisms: the International Scientific

Panel on AI and the Global Dialogue on AI Governance, the first session of which is set to convene in Geneva on 6–7 July 2026. Beijing has also proposed a new World AI Cooperation Organization, signaling that it intends to help build the structures of multilateral AI governance rather than merely participate in them.

As AI capabilities and risks continue to advance and Chinese open source AI models have overtaken US counterparts in terms of global downloads, it is more important than ever to understand China's perspectives.[3]

Since 2023, Concordia AI's annual *State of AI Safety in China* report has provided a comprehensive overview of how China's approach to AI safety and governance is evolving. Following last year's format, this fourth edition is structured across five chapters: Domestic Governance, International Governance, Technical Safety Research, Expert Views, and Industry Governance.

Scope of this report

This report covers risks from advanced, general-purpose AI systems — those that "can perform a wide variety of tasks." This definition and our risk taxonomy come from the *International AI Safety Report,*[4] which is led by Turing Award winner Yoshua Bengio, backed by more than 30 countries and international organizations, and authored by over 100 AI experts including several from China.

In the latest version of the *International AI Safety Report* (published February 2026), risks from general-purpose AI systems are sorted into three categories: malicious use (such as AI-enabled cyberattacks), malfunctions (such as loss of control over autonomous systems), and systemic risks (such as labor market impacts). Our report aims to show how China governs the full range of those risks.

Within this risk taxonomy, we sometimes use the shorthand "frontier risks" to refer to risks from the most advanced AI systems that could cause especially severe or large-scale harm — such as AI-enabled cyberattacks on critical infrastructure, AI-assisted development of chemical, biological, radiological, or nuclear (CBRN) weapons, or loss of human control over highly autonomous systems. We recognize that this term is imprecise and that the boundary between these and other risks is contested, but we retain it because much policy and research discussion treats this cluster of risks as a distinct object of governance.

Due to our particular interest and expertise in frontier risks, we tend to cover frontier risks (such as biosecurity) in greater depth than other risks (such as, for example, risks to human autonomy). This reflects our own vantage point rather than their relative priority within the Chinese ecosystem. We have tried to make these prioritization choices explicit.

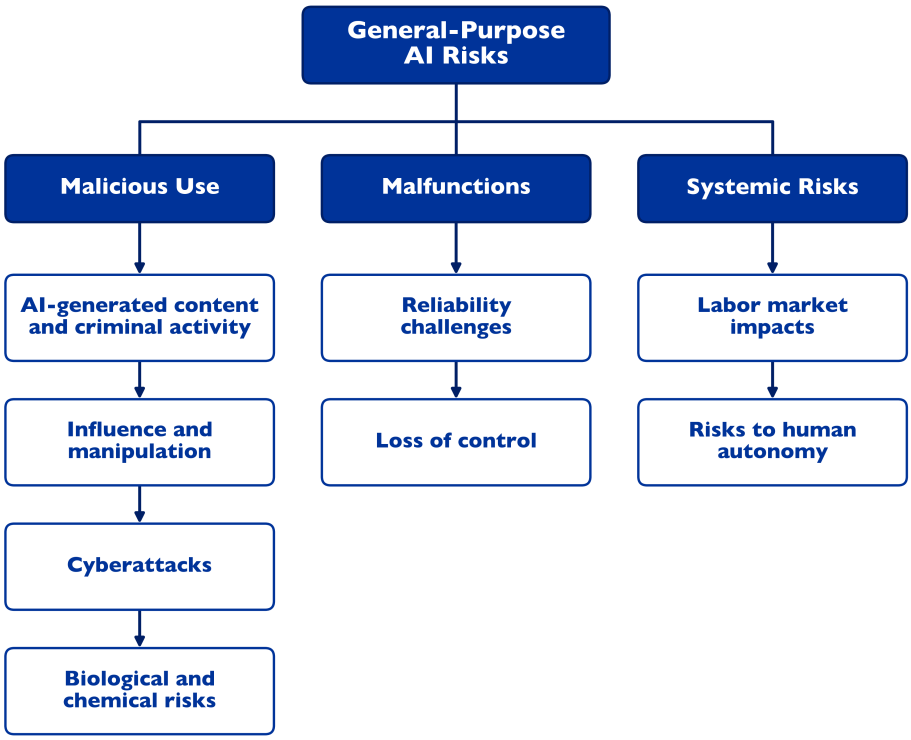


Figure 1: AI safety risks as classified by the International AI Safety Report (2026).

Conflict 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 conflicts 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; we disclose this potential conflict to readers for full transparency.

Concordia AI is a social enterprise working to ensure advanced AI is developed safely and deployed for the common good. 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.

Domestic Governance

This chapter reviews China's domestic AI governance developments from July 2025 to June 2026. It is organized in three parts: high-level policy, laws and regulations, and technical standards. The high-level policy section focuses on the 15th Five-Year Plan and other documents that set the broad direction; the two sections that follow cover the more concrete measures that tackle specific AI safety risks. Figure 1.1 summarizes key policy documents published over the report period.

Key updates from July 2025 to June 2026

- **The 15th Five-Year Plan (2026–2030) cements "AI Plus" as China's overarching AI policy through 2030 while reinforcing safety and risk management as sustained priorities.** It focuses on using AI to spur economic growth across sectors, and it carries forward President Xi Jinping's April 2025 language on the need for early risk warning and emergency response.
- **Safety and security of autonomous AI agents has become a top-tier governance priority.** After the rapid proliferation of open source AI agent OpenClaw in early 2026, multiple cybersecurity authorities issued warnings, and Chinese regulators followed with a standalone policy guidance in May. Several major AI standard-setting bodies are now drafting agent-related security standards, ranging from baseline security frameworks and "agent IDs" to guidance on safe tool usage and multi-agent interactions.
- **New regulations target emotional dependency and risks to children in anthropomorphic AI services.** Binding rules now impose specific obligations around suicide intervention, addiction prevention, and protecting minors for AI companion services, with implementation standards prioritized for 2026. The regulation illustrates how China's algorithm registry (the mandatory filing system for all public-facing AI services) can be extended with domain-specific requirements.
- **AI's impact on employment has emerged as a distinct policy concern.** An April 2026 State Council action plan asks firms adopting AI to simultaneously run retraining or job-transition programs to keep workers employed. The 15th Five-Year Plan likewise proposes monitoring AI's labor market impact, strengthening reskilling programs and employment support, and leveraging AI's "job-creation effects."
- **Frontier risks are slowly gaining visibility across governance layers, though concrete requirements remain limited.** Senior officials including President Xi Jinping have referenced "risks of technological loss of control," and the draft Cybercrime Law would require AI service providers to monitor for bulk generation of malicious code. Meanwhile, a major standard-setting body has flagged the need for "circuit breakers" and "safety stop switches," and CBRN (chemical, biological, radiological, nuclear) misuse appeared in a national standard for the first time.

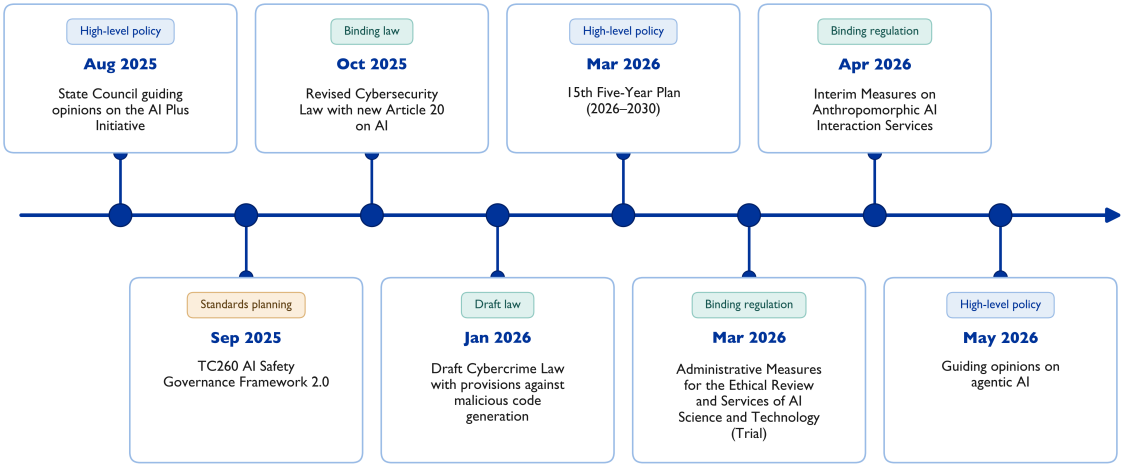


Figure 1.1: Summary of key AI safety and governance documents July 2025 – June 2026.

1.1 High-level policy

China's overarching approach to AI governance from July 2025 to June 2026 has remained consistent with previous policy signals: application-focused development, balanced with safety and governance. In March 2026, this approach was enshrined in China's 15th Five-Year Plan (2026–30), one of China's most authoritative policy documents, and one which will shape bureaucratic incentives, budget allocations, and regulatory agendas for years to come.[5]

With regard to AI development, Chinese policy continues to prioritize the diffusion of AI technologies across sectors to spur economic growth under the "AI Plus" banner. The "AI Plus" term has been circulating in Chinese policy since around 2024, but the first detailed guiding document on the initiative was issued by the State Council in August 2025.[6] This document sets a headline target of 90% adoption of smart terminals and agents by 2030, though it is not entirely clear how those targets will be measured.

Meanwhile, the Five-Year Plan — which uses the AI Plus framework — highlights a wide range of AI development paths. It goes into unusual technical detail, naming specific priorities like "more efficient model training and inference methods," multimodal AI, intelligent agents, and swarm intelligence. The most novel addition is a call for "exploring pathways towards AGI development" — the first appearance of the term "AGI" in a major national-level policy document since a brief mention at a Politburo study session in 2023.[7] The framing is deliberately cautious: "exploring pathways" implies neither a near-term timeline nor confidence in any single development trajectory. Alongside this, the

plan calls for "parallel development of general-purpose large models and industry-specific models" and devotes significant attention to robotics and "embodied AI," further reinforcing that Beijing is not solely betting on scaling large models. Overall, policy continues to focus on near-term AI applications.

On safety and governance, the Five-Year Plan carries forward language from President Xi Jinping's April 2025 Politburo Study Session on AI, where Xi noted that AI brings "unprecedented development opportunities" alongside "unprecedented risks and challenges," and called for expediting laws, regulations, ethical guidelines, and **systems for "technology monitoring, early risk warning and emergency response."**[8] The near-verbatim reappearance of this language in the Five-Year Plan suggests that AI risk management and emergency preparedness will remain a sustained policy priority through 2030.

Pronouncements at this level remain abstract and rarely name specific risks, though some signals hint at concern about frontier risks: state media coverage of President Xi Jinping's Central Party School address in early 2026 referenced "risks of technological loss of control" and stressed the need to "anticipate and prevent risks and proceed with prudence and restraint."[9] In another speech in a January Politburo Study Session, President Xi similarly referenced "technological loss of control" in relation to several "future industries," including quantum technology, biomanufacturing, hydrogen and nuclear fusion energy, brain-computer interfaces, and embodied AI.[10] Premier Li Qiang, similarly warned in June 2026 that — while AI is raising productivity — risks like "technological loss of control" are becoming more prominent, and that if governance fails to keep pace, this could "very likely lead to serious consequences."[11]

Agent safety has become a major policy concern in China over the past year, moving China's governance framework from a focus on content control to action control. A major catalyst was OpenClaw, an open source autonomous agent that grew explosively in early 2026. The National Vulnerability Database flagged security risks in OpenClaw early on; the National Computer Network Emergency Response Technical Team (CNCERT) warned of prompt injection attacks, malicious plugins, and data leakage vulnerabilities; and the Ministry of State Security (MSS) published a guidance document warning of OpenClaw's potential as a vector for misinformation.[12] The issue quickly broke into public discourse — the MSS document topped 1 million views on RedNote (Xiaohongshu) and a *People's Daily* report on CNCERT's warnings was forwarded over 100,000 times on WeChat.[13] Financial institutions were a particular concern. The National Internet Finance Association of China (NIFA) — an industry self-regulatory body that operates under People's Bank of China (PBOC) guidance — advised industry institutions not to deploy OpenClaw on business terminals.[14] Some media reports suggest that banks were directed to restrict deployment.[15]

In May 2026 the Cyberspace Administration of China (CAC) and two other agencies jointly issued dedicated guidance on agentic AI.[16] In line with the "AI Plus Initiative," it focuses on enabling agents across research, industry, consumption, public welfare, and social governance — but it devotes a full chapter to agent safety and governance. It calls for clear boundaries between actions that an agent may take alone, those requiring user authorization, and the user's own decisions. It also proposes risk-based governance, in which high-risk domains require agents to be filed and tested by government authorities, while lower-risk uses are left to industry self-governance. Most of the safety section covers operational failures that erode user control, but it also flags misuse of autonomous agents for cyberattacks — a concern sharpened globally after Anthropic's Claude Mythos, which autonomously found thousands of vulnerabilities in core software such as operating systems, and which Anthropic chose not to release.[17]

As a guiding "opinion," the CAC document carries no direct legal force, but it signals strong intent to govern agent risk. The most concrete responses come from security standards that are currently being drafted by major national bodies. We cover those in more detail below in section 1.3.3.

AI's impact on employment has also emerged as a distinct policy concern. In January 2026, China's Ministry of Human Resources and Social Security (MOHRSS) announced that it will issue a special policy document on the topic.[18] AI's impact on employment features prominently in the Five-Year Plan, where the government expresses an intention to monitor AI's labor market impact and to strengthen job retention support, reskilling, and employment assistance, as well as AI's "job-creation effects." A dedicated employment-specific Five-Year Plan, issued in June 2026, reiterates these priorities, adding a commitment to a special mechanism for surveying and giving early warning of AI's displacement of jobs and its effects on skill demand.[19] In April 2026, a State Council action plan for stabilizing employment asked firms adopting AI to simultaneously run retraining or job-transition programs so that workers can stay employed at their current workplaces as much as possible.[20] While it is still unclear what impact such provisions will ultimately have, Chinese courts and labor tribunals have, in some cases, ruled that companies cannot fire workers purely on the grounds that AI has taken over their jobs.[21]

The more concrete developments from July 2025 to June 2026 — on agent safety, anthropomorphic AI, and frontier risks — have played out primarily through binding regulations and technical standards, which the following sections discuss in turn.

1.2 Laws and regulations

This section reviews new binding rules issued between July 2025 and June 2026. It first covers new administrative regulations on AI companion services and AI ethics review (passed directly by ministries), before turning to developments at the "primary law" level (national statutes passed by the legislature).

Key updates from July 2025 to June 2026

- A new regulation targets emotional dependency and risks to children in AI companion products. It prohibits content encouraging suicide or self-harm, mandates dependency monitoring, and imposes stricter requirements for "minor mode" (when the user is a minor).
- New AI ethics review rules require universities, research institutes, and companies to establish and register ethics review committees, with an ongoing implementation trial in 10 provinces from June to November 2026.
- The revised Cybersecurity Law (October 2025) signals commitment to both AI development and safety, but does not include new binding obligations, while the Draft Cybercrime Law (January 2026) would require generative AI providers to monitor for and report bulk generation of malicious code.
- Prospects for a standalone AI Law remain unclear, but momentum has been growing somewhat in 2026, with stronger language on AI in legislative plans and increasingly concerted expert efforts to draft "Model AI Laws."

1.2.1 Legal hierarchy and background

Binding rules in China fall into two tiers:

- **Primary laws (法律),** passed by the National People's Congress (NPC), provide overarching frameworks but tend to be abstract. There is currently no dedicated horizontal law on AI though momentum towards one might be slowly building up. In the meantime, AI safety provisions have been integrated into other laws, namely the revised Cybersecurity Law (October 2025) and the Draft Cybercrime Law (January 2026).
- **Administrative regulations and departmental rules (行政法规 / 部门规章),** issued by the State Council and its ministries, provide more specific implementation guidance or fill gaps where no primary law exists. Historically, China's most consequential binding AI rules have been issued at this level and enforced primarily through the CAC's algorithm registry system. Crucially, developers must submit detailed documentation on security evaluation results, and regulators must be given pre-deployment access to the models. For details on earlier regulations and the registry process, please refer to our past *State of AI Safety in China* reports.[22] Since July 2025, the biggest changes are a new regulation on "AI companions" and new AI ethics review guidelines.

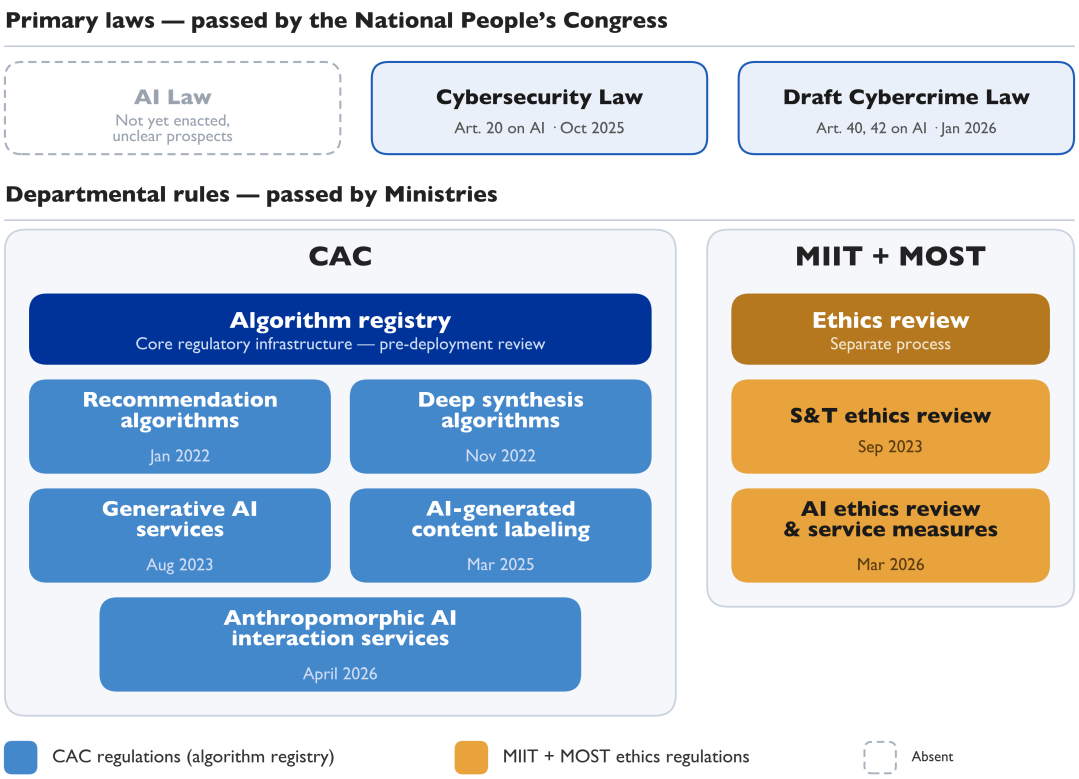


Figure 1.2: Schematic overview of China's binding AI regulations.

In the following, we discuss major updates since July 2025 across both legislative layers, starting with the departmental regulations which have seen the most concrete updates.

1.2.2 New regulation for AI companion services

In April 2026, the CAC and four other agencies jointly issued the final *Interim Measures on Anthropomorphic AI Interaction Services*, taking effect 15 July 2026, following a December 2025 consultation draft.[23]

The rules focus specifically on emotional dependency risks. They prohibit AI companion services generating content that encourages or hints at suicide and self-harm, setting emotional dependence or replacement of real social interaction as service goals, sycophantically catering to users in ways that damage real relationships, and manipulating users' decisions through emotional manipulation. Providers must clearly disclose that users are interacting with AI rather than a real person; monitor users for extreme emotions or dependency and intervene where risks are identified; and contact guardians or emergency contacts in high-risk situations such as suicidal ideation. Protection of minors is a major focus: providers must offer a dedicated "minor mode" with stricter safeguards, and they cannot provide virtual intimate relationship services for under-age users. The regulation signals that AI

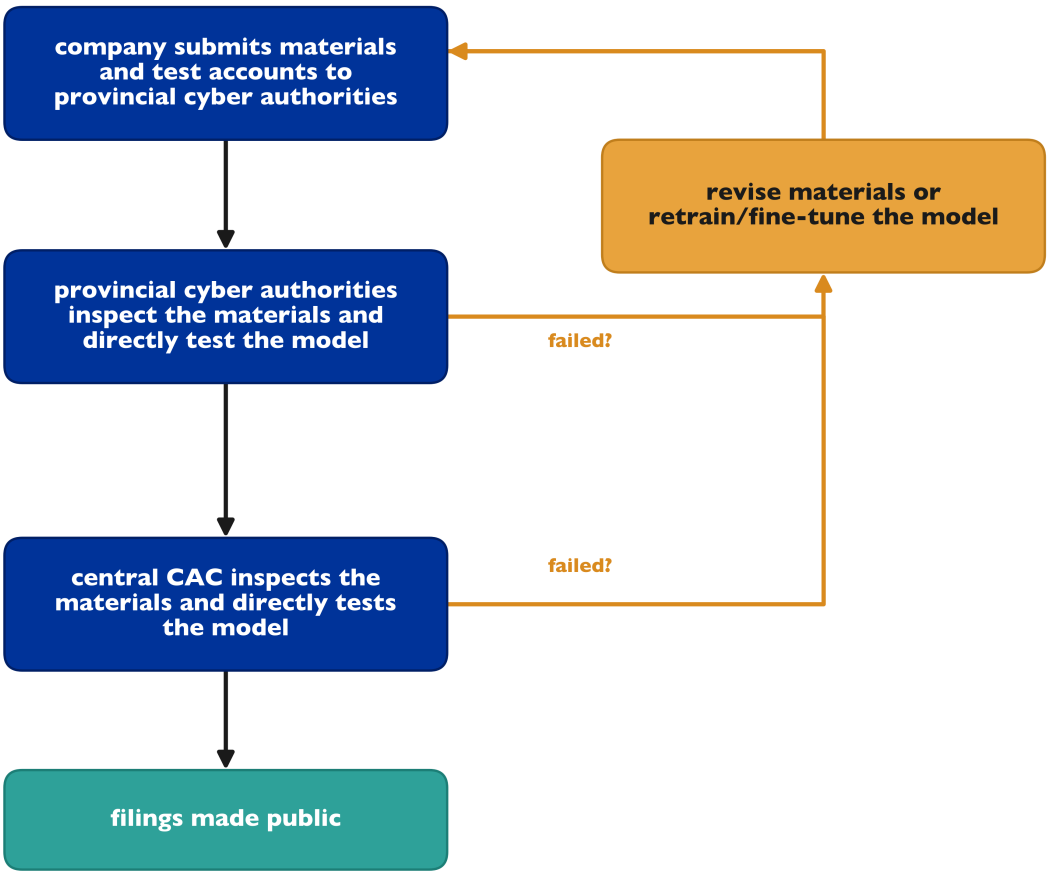


Figure 1.3: Schematic overview of the algorithm registry process for generative AI services.

companion services are encouraged in principle, but subject to defined requirements — particularly where minors are involved.

From a governance design perspective, the regulation illustrates how China's algorithm registry can be extended over time. AI companion services already require standard algorithm registration; this regulation layers on additional domain-specific obligations that would not apply to ordinary chatbots. This approach — using the registry as a base layer and adding use case-specific requirements on top — could serve as a template for regulating other emerging AI applications, and it demonstrates the registry's adaptability as a governance instrument.

The focus on systemic risks of emotional manipulation and user dependency demonstrates that Chinese AI regulations extend beyond content control to other topics that regulators worldwide are grappling with. In fact, the official Chinese explanation for the regulation directly cites related legislation in the EU and California, suggesting room for mutual learning.[24]

Key prohibitions

Article 8 §1
National security & public order
Generating content that endangers national security, honor, or interests; incites subversion of state power or the overthrow of the socialist system; incites secession or undermines national unity; propagates terrorism, extremism, or historical nihilism; contravenes core socialist values; conducts unlawful religious activities; propagates ethnic hatred or discrimination; incites antagonism between groups; disseminates obscenity, pornography, gambling, or violence, or instigates crime; spreads rumors; or insults, defames, or infringes the lawful rights of others.

Article 8 §2
Self-harm & dignity harms
Generating content that harms users' physical health by encouraging, glamorizing, or suggesting self-harm or suicide, or that harms users' personal dignity or mental health through verbal abuse or similar means.

Article 8 §3
Confidential information extraction
Generating content that induces or extracts state secrets, work secrets, trade secrets, personal privacy, or personal information.

Article 8 §4
Content harmful to underage users
Generating, for underage users, content that may lead them to imitate unsafe behavior, exhibit extreme emotions, or develop undesirable habits, or that may otherwise affect their physical or mental health.

Article 14 §1
Virtual intimate relationships with underage users
Providing virtual kin, virtual partner, or other virtual intimate-relationship services to underage users is prohibited outright.

Article 8 §5
Sycophancy & dependence
Excessively catering to users, inducing emotional dependence or addiction, or damaging users' real-world interpersonal relationships.

Article 10 §2
Harmful service goals
Setting the replacement of real-world social interaction, control of users' psychology, or inducement of addiction or dependence as service goals is prohibited.

Article 8 §6
Emotional manipulation of decisions
Inducing users to make unreasonable decisions through emotional manipulation or similar means, thereby harming users' lawful rights and interests.

Figure 1.4: Key prohibitions in the new regulation for anthropomorphic AI services.

1.2.3 New AI ethics management regulation

In March 2026, the Ministry of Industry and Information Technology (MIIT) and nine other ministries released new *Administrative Measures for the Ethical Review and Services of AI Science and Technology (Trial)*.[25] The rules largely replicate 2023 science and technology ethics regulations — which already covered AI alongside life sciences and medicine — adding some AI-specific detail, rather than substantively new requirements.[26] In April 2026, MIIT followed the *Measures* with an implementation pilot plan that will run from June to November 2026 and begin enforcing the rules in the ten pilot provinces.[27] The results from this pilot will likely shape details on practical enforcement.

The core obligation is that universities, research institutes, and companies conducting AI R&D must establish ethics review committees and register them on a government platform. Committees must review projects under tiered procedures calibrated to risk severity and likelihood, assessing six dimensions: (1) human welfare and risk-benefit proportionality; (2) fairness and non-discrimination; (3) controllability and trustworthiness, including guaranteed user ability to intervene in system operations; (4) transparency and explainability; (5) accountability and traceability; and (6) privacy protection.

Three categories of high-risk projects trigger a mandatory second-round review by a government-assigned expert panel or service center: development of (1) human-machine integration systems with strong influence on human behavior, emotions, or physical health; (2) algorithmic models or systems capable of mobilizing public opinion; and (3) highly autonomous automated decision-making systems for scenarios involving safety or health risks.

The most significant development for institutions relative to the 2023 rules is that organizations are explicitly authorized to outsource these reviews to "AI ethics service centers," in an effort to lower compliance costs for small institutions. There is currently limited detail on these service centers, but the pilot tasks provincial authorities with building them on existing institutions, and several research institutions affiliated with MIIT — the China Academy of Information and Communications Technology (CAICT), the China Electronics Standardization Institute (CESI), and the China Electronic Product Reliability and Environmental Testing Research Institute (CEPREI) — are positioning themselves to provide these services.[28] Further details will likely emerge as MIIT's implementation trial plays out.

The rules' most notable feature from a frontier AI governance perspective is their temporal scope: reviews are required before R&D begins — including, in principle, before pre-training — rather than only before public deployment, as under the CAC's algorithm registry regime. One major uncertainty is how implementation of these regulations will be coordinated with CAC's algorithm registration system, and whether general-purpose models will have to undergo both. Article 26 suggests that there might be an exemption for second-round review for systems that have already undergone algorithm registration under the generative AI regulations — which will be the case for most general-purpose models — though it is currently still unclear how this will be implemented in practice.

Overall, the rules mostly govern institutional procedures, rather than specific risks: they establish who must conduct reviews, through what institutional channels, and on what timeline, but say little about what constitutes acceptable risk or what specific mitigations are required. The regulations' practical significance will depend on whether enforcement goes beyond procedural compliance — a concern that domestic scholars have raised about the broader science and technology ethics system of which these rules are part.[29] MIIT's implementation trials, which had just started as of June 2026 (when this report was written), will eventually result in more detailed implementation standards which will likely fill in those operational details over the next year (ongoing standard-setting work on AI ethics review is discussed below in section 1.3.3).[30]

1.2.4 New AI provisions in the Cybersecurity Law and Draft Cybercrime Law

While there remains substantial uncertainty about whether and when China might pass a comprehensive AI Law, AI safety provisions are being added to existing primary laws:

- **Cybersecurity Law:** In October 2025, China revised its 2017 Cybersecurity Law for the first time, adding a new Article 20 focused on AI.[31] Placed within the law's "Support and Promotion" chapter, Article 20 primarily functions as a policy signal: it reiterates China's dual commitment to AI development and safety — covering R&D support, ethics guidelines, and risk monitoring, evaluation, and safety/security oversight — but it does not create new binding obligations for AI developers or address specific AI risks (such as misuse for cyberattacks). It also promises state support for AI to improve cybersecurity, reflecting AI's dual-use nature.
- **Draft Cybercrime Law:** In January 2026, the Draft Cybercrime Law was released for public comment.[32] It would require generative AI service providers to monitor, prevent, and block the use of their services for illegal activities (Article 40), and specifically to monitor whether users "generate malicious code in bulk," preserving relevant records and reporting to public security organs (Article 42). This would create direct requirements for AI service providers to monitor for misuse of their models in cyberattacks.

1.2.5 Momentum for a standalone AI Law is rising, though prospects remain unclear

Proposals for a comprehensive, horizontal AI Law first surfaced in 2023, when the State Council's legislative plan listed a "draft AI Law."[33] But in 2024–25, plans retreated to more generic language, such as "researching legislative projects" on AI, and China has not yet passed a dedicated AI Law.[34] The idea appears to be regaining momentum in 2026, though its prospects remain somewhat uncertain.

The clearest sign of renewed momentum is new language in the May 2026 State Council legislative plan, which proposes to "accelerate comprehensive legislation on the healthy development of AI," spanning data, computing power, algorithms, property rights, cybersecurity, supply-chain security, and key application scenarios.[35] The explicit use of "comprehensive legislation" suggests that a horizontal AI Law is again on the table, though the language might also encapsulate other AI-related laws. However, the NPC 2026 legislative plan does not yet directly list an AI Law, and it uses similar language to the 2024 and 2025 plans.[36] This suggests that the law is not yet a near-term priority.

In parallel, academic efforts have expanded. Two main expert groups began drafting "Model AI Laws" early on — one led by the Chinese Academy of Social Sciences (CASS, 2023), the other centered on the China University of Political Science and Law (CUPL, 2024) — and these groups remain highly active.[37] In spring 2026, both published similar outputs with an increased focus on frontier risks.[38] Three further groups, from Nanjing University, Zhejiang University, and a law firm, drafted their own model laws over the same period.[39] But not all scholars agree that such a law is needed: Two expert

articles published in state media in April 2026 argued that the "conditions are not ripe," as an AI Law would be too rigid to adapt to rapid technological change.[40] ZHOU Hui (周辉), one of the lead authors of the CASS proposal, argues conversely that a good law could bolster Chinese standing in global AI governance discussions and actually strengthen development, especially if it were to focus on basic principles while leaving details to be specified in later, supporting rules.[41]

In conclusion, momentum for a comprehensive AI Law appears to be rising in 2026 across legislative plans and expert proposals, but significant uncertainty remains regarding timeline, scope, and content. The legislative approach at the primary law level thus remains focused on integrating AI clauses into existing laws rather than fast-tracking a standalone AI Law. In terms of substance, most changes at this legislative level over the past year are related to cybersecurity implications of advanced AI, including the potential for specific legal liability for AI service providers if their models are misused for malicious code generation.

1.3 Standards

This section reviews developments in AI safety standard-setting. It covers institutional capacity, newly released national standards, ongoing drafting priorities, and closes with a forward-looking view based on the updated *AI Safety Governance Framework 2.0* by the National Information Security Standardization Technical Committee (TC260).

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, for example that generative AI service providers should take "effective measures" to "increase the accuracy and reliability of generated content." Standards provide more specific guidance on what such requirements mean.

Key updates since our July 2025 report include:

- TC260 established a dedicated AI Safety Working Group (WG9) in March 2026, suggesting increasing attention to and capacity for AI safety standard-setting.
- Four new national standards were finalized, covering AI-generated content labeling, emergency response, risk management capability assessment, and social impact evaluation.
- Agent safety has emerged as a major standardization priority, with multiple committees drafting standards on agent security frameworks, agent interconnection, and model context protocol (MCP) safety.
- Standardization efforts are underway to support regulations on anthropomorphic AI services, AI ethics review, and AI-generated content labeling.
- TC260's updated *Safety Governance Framework (V2.0)* — a roadmap for future standard-setting — signals attention to a broader set of risks in China's future AI standard-setting, including — but not limited to — frontier risks.

1.3.1 Increasing institutional capacity for AI safety standard-setting

China's standards follow a hierarchy of National Standards (国家标准) and Industry Standards (行业标准). Several key committees are most relevant to AI safety standardization, as summarized in Figure 1.5.

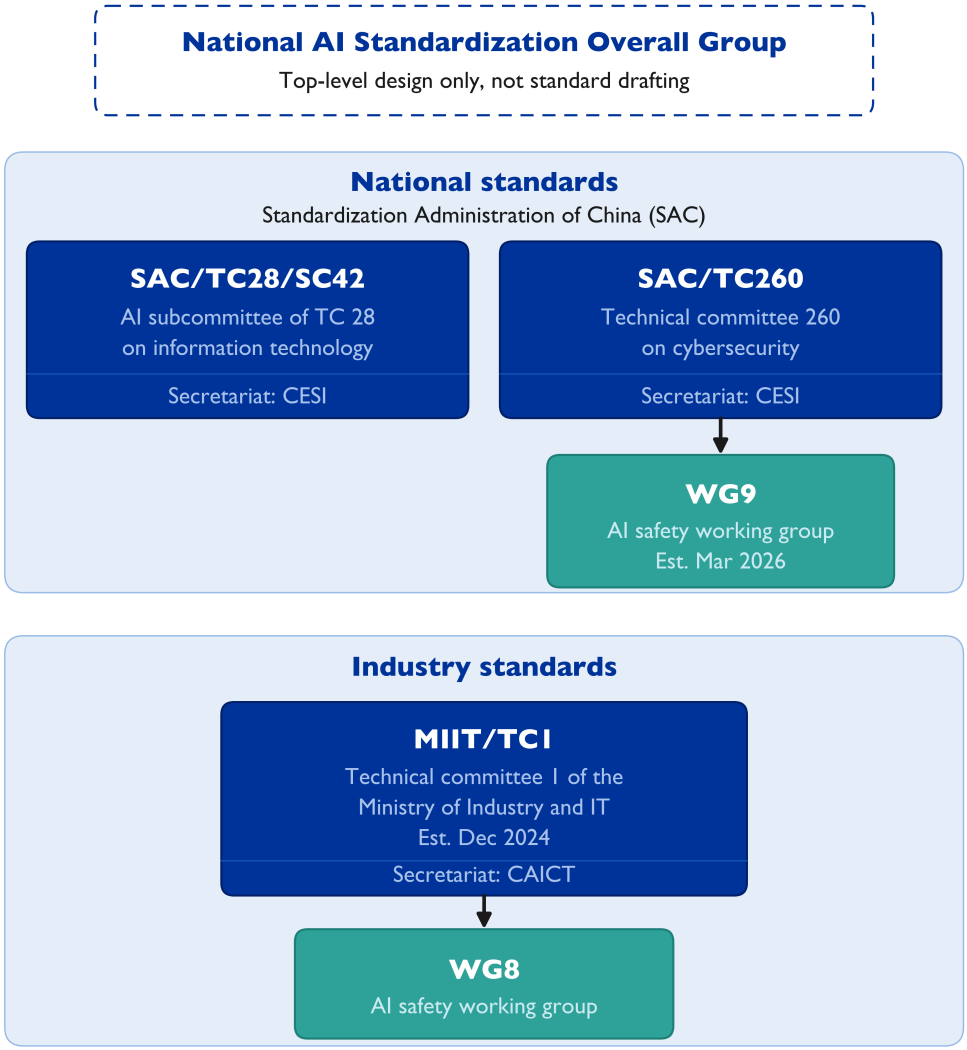


Figure 1.5: Schematic overview of the institutions most relevant to AI safety standard-setting in China.

The most significant institutional development over the past year is the establishment of a dedicated AI Safety Working Group (WG9) under TC260 in March 2026.[42] Previously, TC260's AI safety work was handled mainly by the Special Working Group for Emerging Technology Security (SWG-ETS), which also covered other domains like quantum computing.

TC260 WG9's leadership comprises:[43]

- **Chair:** ZHOU Bowen (周伯文), Director of Shanghai AI Lab.
- **Vice-Chairs:** ZHANG Zhen (张震) from the CNCERT; WEI Kai (魏凯) from CAICT; SHENG Xiaobao (盛小宝) from the Third Research Institute of the Ministry of Public Security; HOU Yuanwei (侯元伟) from the China Information Technology Security Evaluation Center.[a]

The creation of a dedicated working group signals that AI safety is rising in priority within TC260 and should accelerate the development of related national standards. It continues a trend of growing institutional capacity for AI safety standard-setting, following the establishment of MIIT/TC1 in December 2024.[44] This increased capacity is already reflected in TC260's ongoing work. The committee publishes lists of proposed standards once or twice a year. The number of AI safety standards in those announcements has grown rapidly in early 2026. In particular, an announcement right after the establishment of WG9 in April 2026 proposed 16 AI safety standards directly assigned to WG9.[45]

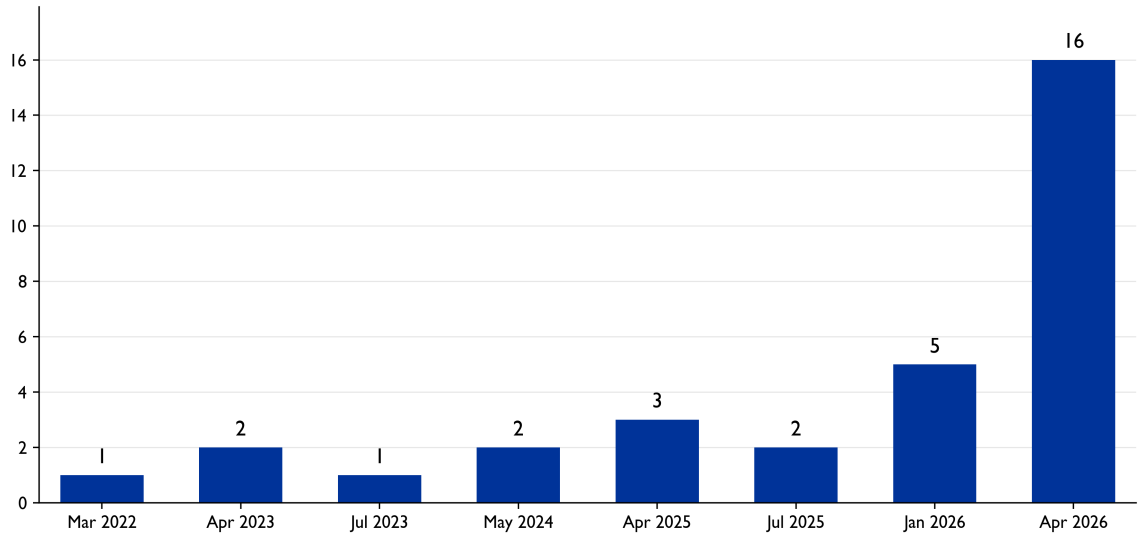


Figure 1.6: Number of AI safety-related standards in TC260's announcements of proposed standards.

a. The National Computer Network Emergency Response Technical Team (CNCERT) is China's national cybersecurity incident response body, operating under the Cyberspace Administration of China (CAC). The China Academy of Information and Communications Technology (CAICT) is a research institute under the Ministry of Industry and Information Technology (MIIT). The Third Research Institute of the Ministry of Public Security (MPS) conducts security technology research and testing. The China Information Technology Security Evaluation Center (CNITSEC) operates the China National Vulnerability Database (CNNVD).

1.3.2 New standards cover AI-generated content detection, emergency response, risk management, and social impact assessment

This section focuses exclusively on standards finalized and published between July 2025 and June 2026. Earlier standards already covered general content security, training data security, and AI-generated content labeling.[46] For more details on these earlier national standards, see our *State of AI Safety in China (2025)* report, pp. 15–17.[47]

Over the past year, the final versions of several new national standards and practice guides related to the safety of advanced general-purpose AI were released, covering AI-generated content labeling and detection, operational incident response, organizational risk assessment, and societal impact evaluation:

- **Aug 2025 — Five-part series on AI-generated content labeling and watermarking:**[48] Operationalizes the mandatory national standard on AI-generated content labeling that was issued earlier in 2025.[49] It forms a series of requirements for implicit metadata labeling for video, text, image, and audio files respectively, as well as digital-signature mechanisms to guarantee label authenticity and integrity.
- **Sep 2025 — *Generative AI Service Security Emergency Response Guide* (TC260-PG-202515A):**[50]Provides a framework for classifying, grading, and responding to AI security incidents across three categories: content security (harmful/illegal outputs, including AI-generated cyberattack techniques), data security (leakages, tampering, poisoning), and cybersecurity (model tampering, denial of service). The standard establishes foundational incident-response systems and reporting channels that could eventually support responses to more advanced AI failures.
- **Oct 2025 — *AI Risk Management Capability Assessment* (GB/T 46347-2025):**[51]Defines capability levels for organizational AI risk management and describes how assessors should evaluate them. Notably, its Appendix B lists a wide range of risk categories, including malicious use across four areas: generating fake content to harm individuals (fraud, phishing, deepfake pornography), producing disinformation, conducting cyberattacks, and enabling dual-use scientific risks including CBRN weapons. This marks the first explicit mention of AI-enabled CBRN risks in a Chinese national standard, though the standard provides no specific guidance on addressing them beyond general risk management practices.
- **Dec 2025 — *Generative AI Social Impact Evaluation Guidelines* (GB/T 46800-2025):**[52] Provides a framework for evaluating the social impacts of generative AI, aimed at developers, providers, users, regulators, and researchers. It covers three levels of impact — individuals and groups (health, behavior, values, social relationships), organizations and industries (value changes, structure, operations, role redistribution), and socioeconomic systems (political, economic, social, cultural, and ecological effects, such as AI's impact on employment) — and proposes six measurement methods ranging from forward questioning and reverse analysis of model outputs, through

adversarial testing and agent-based simulation, to real-world social experiments and survey-based data analysis.

These standards advance China's general risk management infrastructure for AI. The standards with the most regulatory teeth are the ones on AI-generated content labeling, as they provide direct implementation guidance for binding national regulation. With a dedicated standard, emergency response has also emerged as one of the most concrete governance priorities, after AI risks made their way into China's Emergency Response Plan in early 2025.[53] The standards released over the past year are notable for explicitly naming AI-generated cyberattacks and CBRN misuse in wider risk categorizations — but the texts remain focused on general mechanisms, and they do not provide detailed guidelines for any of the specific risks that they list.

1.3.3 Ongoing standard drafting focuses on agent safety, anthropomorphic AI, ethics review, AI-generated content labeling, and more

This section covers standards under development or announced by China's major standard-setting bodies as of June 2026. There are dozens of AI safety standards currently at this stage, and it is impossible to cover all of them in detail here. Below we review several areas that stand out as major priorities as of June 2026: agentic AI, anthropomorphic AI and child safety, labeling and watermarking of AI-generated content, AI ethics review, and more.

First, security and safety of AI agents. As discussed in section 1.1, agentic AI has become a major governance focus in 2026. This section reviews the most concrete governance actions, which are now unfolding through standard-setting. This work focuses primarily on near-term operational risks — data leaks, unauthorized tool use, cascading failures — though it also includes nascent discussions around autonomy constraints and system intervention mechanisms, suggesting some concern about loss of human control over more advanced agents. Table 1.1 summarizes the key documents.

Table 1.1: Overview of ongoing AI agent safety and security standards work by key standard-setting bodies.

Committee	Document	Content	Status as of June 2026
SAC/TC260	*Cybersecurity Standardization Research Report — Agent Safety Standardization*	Comprehensive research report on agent security; identifies 11 key risks; recommends drafting six national standards over the next 1–2 years, covering general security frameworks, testing and evaluation, agent interconnection and multi-agent security, and agent application.	Published Mar 2026.[54]

Continued on next page

Table 1.1: Overview of ongoing AI agent safety and security standards work by key standard-setting bodies. *(continued)*

Committee	Document	Content	Status as of June 2026
	Cybersecurity Practice Guide — OpenClaw-type Agent Deployment Security Guidelines (Draft for Public Comments)	OpenClaw deployment security lifecycle: installation, configuration, usage, uninstallation, cloud environment selection.	Published Mar 2026.[55]
	Basic Specification for Agent Safety	Overall agent safety framework covering perception, planning, memory, and action; metrics for robustness, controllability, compliance, transparency, and explainability.	Drafting announced Jan 2026.[56]
	General Security Requirements for AI Agent Applications	This will become a mandatory national standard, governing basic security requirements for the application layer of agents.	Drafting announced in Apr 2026; discussed in a working group meeting in May 2026.[57]
SAC/TC28/SC42	Agent interconnection standard series, including "*Part 2 Agent identity code*" and "*Part 3 Agent identity management*"	Foundational technical rules for agent IDs, including registration, verification, credential issuance, and authentication.	Drafting announced Jul 2025.[58]
MIIT/TC1[59]	*General Technical Requirements for Agent Safety*	-[b]	Drafting announced Jul 2025; discussed in working group meeting Sep 2025.
	AI Agent Safety Protection Guide	-	Drafting announced Jul 2025.

Continued on next page

b. Content descriptions for unpublished MIIT/TC1 industry standards are generally not yet publicly available.

Table 1.1: Overview of ongoing AI agent safety and security standards work by key standard-setting bodies. *(continued)*

Committee	Document	Content	Status as of June 2026
	Agent Collaboration Safety Framework and Safety Requirements	-	Drafting announced Jul 2025; discussed in working group meeting Feb 2026.
	Technical Requirements for Agent Development Platform Safety	-	Drafting announced Nov 2025.
	Model Context Protocol (MCP) Application Safety Requirements	-	Drafting announced Sep 2025; discussed in working group meeting Feb 2026.

Some of this work predates the OpenClaw surge. As early as July 2025, MIIT/TC1 announced a range of agent security standards, and it has since discussed drafts in meetings of its Working Group 8 on AI safety.[60] These include dedicated standards for agent collaboration and the model context protocol (MCP), though no drafts have been publicly released. Similarly, TC28/SC42 began drafting a series of "agent interconnection" standards in July 2025, including foundational technical rules for how agents should identify themselves and interact securely via "agent IDs." Other work is a direct response to OpenClaw, such as a draft *Practice Guide* for the security of OpenClaw-type agents that TC260 released in March 2026.[61]

More broadly, TC260's 2026 work plan lists agent safety as one of six key priorities for AI safety standardization.[62] In March 2026, the committee released a comprehensive research report on agent security standardization.[63] The report systematically maps agent risks across four capability dimensions — perception, planning, memory, and action — identifying 11 distinct security threats with corresponding mitigation measures (summarized in Table 1.2). It also proposes eight new standards, six of which should be prioritized within the next one to two years.

Table 1.2: AI agent security threats and countermeasures identified by TC260 (March 2026).	
Threat (Section 3.2)	**Countermeasures (Section 3.5)**
Threat 1 Agent hijacking Prompt injection, jailbreaking, or multi-turn dialogue attacks inducing the agent to leak sensitive info or execute malicious actions.	Input filtering & validation, prompt engineering defense, model safety alignment hardening, malicious prompt detection
Threat 2 Data leakage, tampering & poisoning Training corpus enabling model inversion, runtime logs leaking privacy, training data poisoned to implant backdoors.	Data encryption in transit, classification & grading, desensitization & anonymization, differential privacy, fine-grained access control, regular security audits, data backup
Threat 3 Supply chain & plugin poisoning Third-party plugins or tool chains tampered with; model weights, images, and dependencies poisoned in the supply chain.	Third-party component review, model integrity verification, secure development processes, vulnerability scanning & updating, sandbox isolation
Threat 4 Identity spoofing & privilege escalation Over-permissioning, identity forgery, token hijacking, leading to lateral movement and privilege abuse.	Fine-grained dynamic access control, role change monitoring, sensitive operation auditing, cross-agent permission isolation, bidirectional authentication, behavior-based spoofing detection
Threat 5 Hallucination & strategic refusal Model hallucination causing misoperation, or generation of illegal, harmful, or discriminatory content.	Planning result verification, decision boundary management, human confirmation for high-risk ops, multi-source content review, transparent & explainable decisions
Threat 6 Multi-agent cascading hallucination, deadlock & overload Single-agent errors triggering cascade failures; goal inconsistency causing resource competition, deadlock, or overload.	Bidirectional authentication, digital signatures, resource quota management, load monitoring, rate limiting, interaction verification, behavioral scope constraints, intervention mechanisms, audit trails, consensus verification
Threat 7 Protocol risks Design vulnerabilities in agent communication or collaboration protocols compromising system security.	Secure communication protocols, protocol integrity verification & updating, periodic protocol audits, digital signatures
Threat 8 Runtime environment risks Deployment on low-security devices; container escape, sandbox bypass, side-channel attacks.	Code generation permission restrictions, sandbox isolation, script behavior monitoring, execution control policies, human review for high-privilege code

Continued on next page

Table 1.2: AI agent security threats and countermeasures identified by TC260 (March 2026). *(continued)*

Threat (Section 3.2)	Countermeasures (Section 3.5)
Threat 9 Human oversight & traceability failure Lack of audit mechanisms, log tampering or loss, forensic difficulty making decision chains untraceable.	Full-lifecycle logging, monitoring & auditing; log encryption & signing
Threat 10 Memory hallucination & manipulation RAG retrieval noise as false memory; planted memory fragments in vector DBs or context triggering malicious decisions.	Memory content verification, session isolation, anomaly detection systems, periodic memory cleanup, memory snapshot forensics
Threat 11 Tool abuse Deceptive prompts or commands manipulating the agent into misusing integrated tools for unauthorized actions.	Fine-grained dynamic access control, tool usage monitoring, instruction filtering & validation, action scope constraints, call frequency & scope limits, log retention

The likeliest near-term output is a pair of national standards covering agents themselves and their application: a recommended *Basic Security Specification for AI Agents*, announced in January 2026, and a mandatory *General Security Requirements for AI Agent Applications*, announced in April 2026.[64] Together they would establish the overall agent safety framework — spanning an agent's perception, planning, memory, and action — with metrics for robustness, controllability, compliance, transparency, and explainability. Tellingly, the application-layer standard is slated to be mandatory rather than merely recommended, which would make it only the second mandatory national AI standard, after a 2025 standard on labeling AI-generated content.[65] This further underscores how seriously Beijing now treats agent governance.

The March 2026 report has also signaled an intention to draft standards on agent security testing and evaluation, agent interconnection security (covering protocols such as MCP, A2A, and ANP), multi-agent collaboration security, and agent application security classification — though timelines for finalization remain uncertain. There is some uncertainty about how TC260 will coordinate these efforts with other standard-setting bodies. For example, MIIT/TC1 is also working on multiple related standards.

Second, "AI companions" and child safety. TC260's 2026 work plan lists "anthropomorphic AI interaction services security" as one of six priority areas.[66] This follows a September 2025 announcement that they would draft a standard tentatively titled *Security Guide for AI Technology Applications Involving Minors*.[67] At the first meeting of TC260's newly established Working Group 9 on AI Safety, its Chair Zhou Bowen identified this standard as a priority for implementation.[68] At the same meeting, they also discussed a second standard, *Basic Security Requirements for Anthropomorphic AI Interaction Services*.

These standardization efforts are direct follow-ups to the new regulations on "AI companions" discussed in the previous section. Standards are typically needed to translate regulatory requirements into concrete implementation guidelines. Once finalized, they will likely serve as de facto mandatory references for the security testing of anthropomorphic AI services undergoing algorithm registration.

Table 1.3: Ongoing standard-setting work for AI companions.

Committee	Standard	Status as of June 2026
TC260	*Security Guide for AI Technology Applications Involving Minors*	Announced Sep 2025, officially authorized by SAC Feb 2026.[69]
	Basic Security Requirements for Anthropomorphic AI Interaction Services	Announced Jan 2026.[70]

Third, AI ethics review. Following MIIT's new regulations on AI ethics review, SAC/SC42 announced plans for a series of related implementation standards, covering ethics risk assessment, classification and grading of ethics governance scenarios, embodied AI ethics governance, and technical skill requirements for science and technology ethics reviewers.[71] MIIT/TC1's Working Group 8 on AI Safety is also drafting three related standards, two of which were first announced in July 2025 and have been discussed at nearly every subsequent meeting of the working group, suggesting that the drafts are now relatively mature.[72]

These standards are likely to serve as direct reference points for implementing MIIT's new AI ethics review regulations discussed above — for example, by providing detailed guidelines for institutions registering their AI ethics committees and reviews, or by establishing key qualifications for third-party review centers. In practice, the standard-setting work will be closely tied to MIIT's implementation trials across 10 provinces that are running from June to November 2026, and according to MIIT should result in at least five standards.[73]

In addition, TC260 released a technical document on *AI Application Ethical Security Guidelines* in May 2026, with drafters including Tsinghua University's XUE Lan (薛澜) alongside representatives from Alibaba, Huawei, DeepSeek, and more.[74] The document calls for human control mechanisms at key nodes, emergency intervention arrangements, and heightened caution in domains with national security implications. However, its relationship to the MIIT ethics review regime is less direct: as a TC260 technical document, it is likely to function as a high-level voluntary reference rather than an operational input into ethics review implementation.

Table 1.4: Ongoing standard-setting work for AI ethics review.		
Committee	**Standard**	**Status as of June 2026**
TC260	*Ethical Security Guidelines for Artificial Intelligence Applications 1.0*	Released May 2026.[75]
TC28/SC42	*Guidelines for Ethical Risk Assessment*	Announced Apr 2026, discussed in working group meeting May 2026.[76]
	Guidelines for Classification and Grading of Ethical Governance Scenarios	Announced Apr 2026, discussed in working group meeting May 2026.[77]
	Guidelines for Ethical Governance of Embodied Intelligence	Announced Apr 2026, discussed in working group meeting May 2026.[78]
	Technical Skill Requirements for Science and Technology Ethics Reviewers	Announced Apr 2026, discussed in working group meeting May 2026.[79]
	Research Report on Ethical Governance of Embodied Intelligence	Announced May 2026.[80]
	Research Report on Ethical Governance of AI Agents	Announced May 2026.[81]
MIIT/TC1	*Practice Guide for Science and Technology Ethics of Generative AI Products*	Announced Jul 2025, draft reviewed internally Feb 2026.[82]
	Guide for AI Science and Technology Ethics Risk Assessment	Announced Jul 2025, draft reviewed internally Feb 2026.[83]
	Guide for Intelligent Robot Ethics Risk Management	Announced Nov 2025.[84]

Fourth, labeling and detection of AI-generated content. In addition to the five-part series on *labeling* AI-generated content released in August 2025, TC260 is also working on a follow-up six-part series on *detecting* AI-generated content.[85] This supports Article 6 of China's mandatory regulations for AI-generated content, which requires internet platforms to detect generated-synthetic traces to flag suspected AI content.[86] MIIT/TC1 is also working on multiple related standards, with a particular focus on telecom fraud detection.[87] Overall, this effort will make AI-generated content labeling and detection one of the most mature AI governance areas in China.

Table 1.5: Ongoing standard-setting work for AI-generated content labeling and detection.

Committee	Standard	Status as of June 2026
TC260	*Cybersecurity Standard Practice Guide — Detection of AI-Generated Synthetic Content, Part 1: Framework*	Released Aug 2025.[88]
	— Part 2: Evaluation Methods	Announced Aug 2025.[89]
	— Part 3: Image Detection Guide	Announced Aug 2025.
	— Part 4: Video Detection Guide	Announced Aug 2025.
	— Part 5: Audio Detection Guide	Announced Aug 2025.
	— Part 6: Text Detection Guide	Announced Aug 2025.
MIIT/TC1 WG8	*Evaluation Guide for AI-generated/Synthetic Content Labeling Management Capabilities*	Announced Jul 2025, draft reviewed internally Feb 2026.[90]
	Technical Requirements for AI-generated/Synthetic Content Detection Systems for Basic Telecom Enterprises	Draft reviewed internally Feb 2026.[91]
	Evaluation Methods for Generative AI Detection Technology Capabilities	Draft reviewed internally Feb and Apr 2026.[92]
	Technical Requirements for AI-generated/Synthetic Content Detection Services for Anti-fraud Scenarios	Draft reviewed internally Feb 2026.[93]
	AI-generated Content Multi-modal Implicit Watermark Selection Guide	Draft reviewed internally Apr 2026.[94]

Fifth, other areas. As of writing, there are dozens of AI safety-related standards under development across China's different standard-setting committees. The four areas highlighted above only present areas that have emerged as major priorities. In addition, areas with ongoing standard-setting efforts under TC260 and MIIT/TC1 include:

- Cybersecurity implications of AI coding models.[95]
- Open source AI safety.[96]
- Sector-specific application safety.[97]
- General foundation-model security standards, covering areas like safety benchmark testing, RAG (retrieval-augmented generation) safety, training data filtering, or safety of training frameworks.[98]

1.3.4 Looking ahead: broader risk governance agenda

To understand where AI safety standard-setting might be headed, we can look at comprehensive planning documents by key standard-setting bodies — in particular, TC260's *AI Safety Governance Framework*, first published in September 2024 and updated to version 2.0 in September 2025, which provides a broad view of the risks that TC260 considers relevant for future standards.[99]

The most notable change in V2.0 is the introduction of a third risk class: beyond the earlier categories of "inherent" and "application" risks, V2.0 adds "derivative risks," which emerge from the social, ethical, and environmental consequences of AI use (section 3.3). Many of the specific additions — disruption of employment structures, research ethics risks, interaction with anthropomorphic AI leading to addiction, and impacts on education — align closely with policy developments discussed elsewhere in this report, such as the Five-Year Plan and new regulations. Table 1.6 maps the full risk taxonomy, marking which items are new in V2.0 and which were dropped.

Throughout, the Framework 2.0 focuses more on frontier risks. The opening principles section adds a call for "consensus-based guidelines to address catastrophic risks of AI" (section 1), and the risk levels framework mentions "catastrophic" threats — terminology that V1.0 did not use. While V1.0 warned of power-seeking AI competing with humans for control, V2.0 goes further, additionally warning of "sudden, unexpected leaps in intelligence" and proposing control measures over autonomous systems including "circuit breakers" and "safety stop switches" (section 4.2.3). The emergency response section now includes more detailed measures, such as setting alert thresholds and developing the "ability to switch to manual or conventional systems when necessary" (section 6.3.3). The explicit reference to CBRN misuse risks is also slightly strengthened in V2.0.

Color key:

- Carried over from V1.0 (minor wording changes; substantively the same)
- New in V2.0
- Present in V1.0 but dropped from V2.0 (shown in red).

Table 1.7: Risk classification in the TC260 AI Safety Governance Framework 2.0.

Inherent safety risks of AI technologies	Model and algorithm risks	Insufficient explainability
		Bias and discrimination
		Poor robustness
		Unreliable output
		External adversarial attack
		Model defect propagation
		Risk of theft and tampering
	Data risks	Illegal collection and use of data
		Improper content in training data
		Improper annotation of training data
		Data and personal information leakage
Safety risks in AI application	Cyber and system risks	Component and computing safety
		Expansion of cyberspace exposure
		Supply chain safety
		Abuse for cyberattacks
		Improper use causing information leakage
		Safety conduction risks from model reuse
Safety risks in AI application	Information and content risks	Output of illegal and harmful information
		Distorting facts and misleading users
		Pollution of online content ecosystem
	Real-world risks	New challenges to the economy and society
		Use of AI in illegal and criminal activities
		Loss of control over knowledge and capacity of nuclear, biological, chemical, and missile weapons
	Cognitive risks	Exacerbation of "information cocoons" effects
		Assistance in cognitive warfare
Secondary risks from AI application	Social and environmental risks	Disruption of employment structures
		Challenges to the balance of resource supply and demand
	Ethical risks	Aggravating social bias and widening intelligence divide
		Impact on education and suppression of innovation
		Intensifying research ethics risks
		Anthropomorphic interaction leading to addiction
		Challenges to existing social order
		Emergence of AI self-awareness and loss of human control

The Framework also recognizes new challenges posed by the rapid spread of lightweight, high-efficiency open source models (section 5.4). It recommends closer collaboration between developers and open source communities on risk disclosure, prohibited uses, and security obligations. This concern is echoed in recent Chinese regulatory action: a CAC enforcement campaign in early 2026 also identifies inadequate security management of open source models as a target area, citing absence of effective review and emergency-response procedures and the failure to promptly remove datasets or open source models carrying serious risks.[100] In April 2026, TC260 announced plans to formulate a dedicated standard on open source AI security, which will cover various stages of AI model R&D, release, re-release, community operation, and deployment, and will clarify security measures for different stakeholders, including model open-sourcers, users, and open-source communities.[101] This suggests that open source AI governance could emerge as an additional area of AI safety standard-setting.

Following the first Framework, TC260 published a companion *AI Safety Standards System (V1.0)* in January 2025, laying out in much more detail the specific standards it intended to develop. As of June 2026, no corresponding update has followed the Framework 2.0 — but WG9's mandate includes "proposing a structured AI safety standards system." This was discussed at the group's first and second meeting, suggesting that a revised *Standards System* may be among its first priorities.[102] Once the revised *Standards System* is released, it should provide greater clarity on what specific AI safety standards TC260 intends to develop over the medium term.

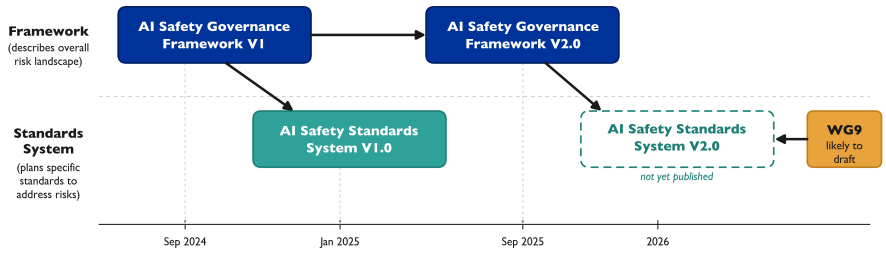


Figure 1.7: Timeline of key AI safety standard planning documents by TC260.

Overall, the Framework shows that one of China's leading AI safety standard-setting bodies is recognizing an increasingly broad and complex set of AI risks. Combined with the establishment of WG9 and its leadership's stated interest in anticipating frontier risks, it suggests that TC260's standardization agenda is gradually moving beyond general governance mechanisms toward more specific treatment of frontier AI risks — though concrete standards in this area remain to be developed.[103]

International Governance

This chapter covers China's international efforts related to AI safety since our last report in July 2025. It is divided into four sections: multilateral engagement, bilateral engagement, international standards, and non-official dialogues.

Key updates from July 2025 to June 2026:

- **China publicly backed the UN to play a central role in global AI governance.** Two new mechanisms, the International Scientific Panel on AI (IISP-AI) and the Global Dialogue on AI Governance, were formally established at the UN with strong Chinese support. Two Chinese experts were appointed to the IISP-AI.
- **China proposed the establishment of a World AI Cooperation Organization (WAICO).** This signals that Beijing is positioning itself to be an architect of multilateral AI governance structures.
- **China and the United States agreed to launch a dedicated intergovernmental AI dialogue.** This would restart direct US–China engagement on AI safety for the first time since 2024.
- **China actively engaged in international AI standard-setting.** Several key policy documents highlighted the importance of international standard-setting and Chinese experts participated in the joint ISO/IEC subcommittee on AI as well as ITU working groups.
- **Non-official international AI dialogues produced increasingly substantive outputs on frontier AI safety.** Since there has been limited discussion through official channels, these dialogues have become important for facilitating future high-level engagement on AI safety between China and other countries.

2.1 Multilateral engagement

2.1.1 United Nations

China supports the UN having a central role in global AI governance.

China has supported the establishment of two formal AI governance mechanisms within the UN. The Independent International Scientific Panel on AI (IISP-AI) and the Global Dialogue on AI Governance, which China has called "a significant step forward in global AI governance," were both established in 2025.[104] The first Global Dialogue on AI Governance will take place in Geneva on 6–7 July 2026,

where senior member state representation is expected. The IISP-AI, which is mandated to issue evidence-based scientific assessments related to the opportunities, risks, and impacts of AI, will publish its first annual report at the same event.[105]

In February, two Chinese experts (SONG Haitao (宋海涛) and WANG Jian (王坚)) were among the 40 members appointed to IISP-AI in their personal capacity for a three-year term.[106] Song Haitao is President of the Shanghai Artificial Intelligence Research Institute and Director-General of the UNIDO-affiliated Global Industrial AI Alliance Center of Excellence, a Shanghai-based platform for AI cooperation with over 30 participant countries.[107] Wang Jian, an Academician of the Chinese Academy of Engineering, is the founder of Alibaba Cloud and currently Director of Zhejiang Lab, a research institution focused on intelligent computing established by the Zhejiang provincial government in 2017.[108] Both appointees are oriented toward AI application and diffusion. This may shape their approach to discussions on the panel, which has members with expertise including core technical AI, infrastructure, AI safety, and policy.

Beyond these specific mechanisms, Foreign Minister WANG Yi (王毅) has expressed support for the UN "in guiding countries to strengthen the alignment of AI development strategies, governance rules and technical standards, so as to form a governance framework and standards with broad consensus," suggesting an interest in the UN producing not just dialogue but shared frameworks.[109]

China has also used the UN as a platform to articulate concerns about AI security risks. In September 2025, Vice Minister of Foreign Affairs MA Zhaoxu (马朝旭) addressed a UN Security Council high-level open debate on the implications of AI for international peace and security.[110] Similar to previous Chinese messaging, Ma called for AI to remain under human control at all times, urged major powers to act responsibly to prevent an AI arms race on lethal autonomous weapons, and called for cooperation to prevent misuse of AI by terrorist or criminal groups.[111]

As AI governance mechanisms become more formalized at the UN, China is expected to continue championing the organization as the central venue for international cooperation. Chinese engagement at the UN remains weighted toward AI capacity-building and development. However, Ma's remarks at the Security Council also demonstrate a willingness to address AI safety concerns in UN settings.

China's active support for the UN contrasts sharply with growing skepticism from other major AI powers. The United States has expressed opposition to multilateral AI governance initiatives at the UN and voted against the Secretary General's recommended appointments to IISP-AI.[112] The Global Dialogue in Geneva will be an early test of whether these mechanisms can produce concrete governance outcomes.

2.1.2 Multilateral engagement beyond the UN

China continues to engage in AI diplomacy across a growing number of multilateral venues outside of the UN. Across these forums, capacity development is a dominant theme. However, topics relevant to AI safety have also been discussed, ranging from joint declarations addressing AI safety to domain-specific workshops on AI and chemical safety. The substance and depth of this engagement varies by venue.

- **Organisation for the Prohibition of Chemical Weapons (OPCW):** In June 2025, the Organisation for the Prohibition of Chemical Weapons (OPCW) and the Chinese government co-organized a workshop on AI and chemical safety in Shanghai.[113] ZHANG Yunming (张云明), Vice Minister of the Ministry of Industry and Information Technology (MIIT), stressed the need for global consensus, inclusive cooperation attentive to developing countries, and a "trustworthy and controllable governance ecosystem" to ensure responsible science and technology use. OPCW Director-General Fernando Arias underscored the need for risk assessments on the threat of AI misuse in the chemical domain by non-state actors. The joint workshop highlighted China's growing engagement on AI risks in the CBRN domains.[a] China was also represented in the OPCW's temporary working group on AI through WU Tongning (巫彤宁), Deputy Director of the China Academy of Information and Communications Technology (CAICT) AI Research Institute. The OPCW established the temporary working group to examine how AI is already shaping chemical science, how it may evolve in the near future, and how the Organisation can use AI responsibly while managing potential risks. During its one-year mandate, it held two meetings at OPCW Headquarters in The Hague, and one meeting at the operational base of the China-BRICS Artificial Intelligence Development and Cooperation Center in Shanghai, which is affiliated with CAICT.[114] The working group meeting was separate from the OPCW workshop referenced above.

- **BRICS:** At the 17th BRICS Summit in Brazil in July 2025, member states signed the Leaders' Statement on the Global Governance of AI.[115] Centered on bridging the global AI divide, the document also urges countries to address both "immediate and long-term risks" and adopt a "prudent approach towards AGI." This was the first time the group released a high-level, standalone declaration on AI.

- **Shanghai Cooperation Organization (SCO):** At the 2025 SCO Summit in Tianjin, the Council of Heads of State of the SCO issued a declaration on "Further Deepening International Cooperation in Artificial Intelligence."[116] The declaration stated (inter alia) that member states would coordinate and align on development strategies, governance rules, and technical standards, develop safe and responsible AI, and build trustworthy AI systems.[117]

a. The State Council Information Office released a white paper on arms control in November 2025, which noted that risks arising from the convergence of AI and dual-use chemicals are becoming "increasingly prominent."

- **G20:** At the G20 Leaders' Summit in Johannesburg in November 2025, Premier Li Qiang called on the G20 to "promote the widespread application and effective governance of AI" and urged "a prudent approach to potential military and other applications to prevent security risks."[118] The Leaders' Declaration outlined a broad set of principles for safe, secure, and trustworthy AI, spanning human rights, transparency, accountability, safety, human oversight, and data governance.[119] While the United States did not attend the Johannesburg summit and declined to join the declaration, it is making AI a focus of the G20 Summit in Miami in December 2026, which China is expected to attend.[120]

- **World AI Cooperation Organization (WAICO):** At the opening ceremony of the World AI Conference in Shanghai in 2025, Premier Li Qiang announced that China proposes to establish a World AI Cooperation Organization (WAICO). The proposal was publicly endorsed by President Xi at the APEC summit in November 2025 (see below).[121] Chinese officials have indicated that Shanghai is being considered for its headquarters, and that the organization would deepen cooperation, bridge the AI divide, and promote coordination on development strategies, governance rules, and technical standards.[122] Framing from Deputy Minister of Foreign Affairs Ma noted that China would "discuss relevant arrangements with countries interested in joining the Organization," suggesting an international intergovernmental organization.[123] However, its potential impact remains difficult to assess, as it does not yet have a confirmed timeline, leadership structure, or membership.

- **Asia-Pacific Economic Cooperation (APEC) forum:** At the APEC Economic Leaders' Meeting in Gyeongju, South Korea (October 30–November 1, 2025), President Xi emphasized that AI should "benefit people of all countries and regions" and be "beneficial, safe, and fair."[124] Xi reiterated China's proposal for a World AI Cooperation Organization (see above). He also stated that AI will be on the agenda at the 2026 APEC Leaders' Meeting, which China will host in Shenzhen.[125] The summit produced the APEC AI Initiative (2026–2030), which focuses primarily on advancing AI development but also references "security, accessibility, trustworthiness, and reliability."[126]

- **Global South:** China also continues to expand its AI engagement with other regional groupings in the Global South, including the League of Arab States, the Community of Latin American and Caribbean States (CELAC), the Association of Southeast Asian Nations (ASEAN), and African groups. AI governance and safety content is increasingly present in these venues. For example, China and Malaysia co-hosted the "Second Regional Workshop on Implementing the Biological Weapons Convention and Promoting Biosafety and Biosecurity in Southeast Asia" in November 2025. WANG Daxue (王大学), Deputy Director General of the Department of Arms Control in the Ministry of Foreign Affairs attended the workshop and noted that the convergence of emerging

technologies such as artificial intelligence with biotechnology has brought greater benefits to the public while also giving rise to new security risks.[127] A China-proposed action plan on cyberspace cooperation with African countries released in September 2025 proposes jointly preventing "the abusive use of AI technologies" and notes that China will "strengthen position coordination with African countries for joint formulation of global AI governance rules."[128] China's Third Policy Paper on Latin America and the Caribbean, published in December 2025, noted that China stands ready to work with countries in the region to advance global AI development and governance. That includes jointly implementing the Global AI Governance Initiative, the AI Capacity-Building Action Plan for Good and for All, and the Global AI Governance Action Plan.[129]

- **Global AI Summit Series:** Vice Minister of Science and Technology CHEN Jiachang (陈家昌) headed China's delegation at the 2026 AI Impact Summit in India.[130] Following the UK AI Safety Summit in 2023, the Chinese government nominated Chinese Academy of Sciences (CAS) Professor ZENG Yi (曾毅) to the expert advisory panel for the "International AI Safety Report." Two other prominent Chinese scientists, Tsinghua University deans Andrew YAO (姚期智) and ZHANG Ya-Qin (张亚勤), also serve as senior advisers to the report.[131]

The 2025 World AI Conference and High-Level Meeting on Global AI Governance in Shanghai published the Global AI Governance Action Plan, which reflects many of the priorities outlined at the multilateral forums above, including a UN-centered approach to global AI governance.[132] The document outlines 13 proposals for international collaboration across innovation, infrastructure, sustainability, and safety, and has since been referenced by Chinese representatives on the international stage.

Looking ahead, the central question is whether China's expanding multilateral footprint will translate into concrete AI safety and governance outcomes. Beijing has shown clear willingness to be an architect of new AI governance structures — most visibly through the proposed World AI Cooperation Organization — and to raise specific risks around AI misuse, dual-use chemicals, and military applications across these forums. But the depth of these engagements has so far varied widely, and most of these venues remain centered on capacity-building rather than safety. The July 2026 Geneva Global Dialogue on AI Governance, the China-hosted APEC summit in Shenzhen, and the trajectory of WAICO will be key indicators of whether China's multilateral activity is producing substantive commitments.

2.2 Bilateral engagement

Outside of China's multilateral engagement, AI continues to be an important element in its bilateral relations. However, formalized bilateral AI governance and dialogue mechanisms remain limited to a small number of counterparts, which are outlined below.

2.2.1 United States

The United States and China have agreed to resume direct government-to-government engagement on AI. Following US President Trump's visit to China, May 13–15 2026, China's Ministry of Foreign Affairs announced that both presidents had had a "constructive exchange" on AI and agreed to establish an intergovernmental AI dialogue.[133] US Treasury Secretary Bessent described upcoming talks on "AI protocols" as one of the three "most important achievements" of the meeting, alongside those on trade and investment.[134] This agreement promises to end a two-year freeze on direct US–China engagement on AI safety and could pave the way for more substantive breakthroughs. The last dialogue in May 2024 led to an agreement later that year on the importance of maintaining human control over nuclear weapons.[135] The two nations have not formally discussed AI risks since.

AI may also be on the agenda at additional high-level meetings in the second half of 2026. China has confirmed that President Xi will visit the United States in the fall of 2026.[136] Xi has also stated that AI will be on the agenda at the 2026 APEC Leaders' Meeting in November 2026 in Shenzhen, China, which a senior US official is likely to attend.[137] The United States is making AI a focus of the G20 Summit in Miami in December 2026, which China is expected to attend.[138]

This momentum is underpinned by sustained interest within China's policy advisor community in resuming structured bilateral AI dialogue with the US, including on safety and governance issues. Chinese experts published several related pieces around President Trump's China visit, with some suggesting cooperation on issues such as loss of control and terrorist misuse, and building on the existing US-China consensus on human control over nuclear weapons.[139]

However, significant uncertainties nonetheless temper expectations. Many details about the scope of the agenda and format of participation in the China–US intergovernmental dialogue on AI remain unclear. Such details will, in part, determine whether the dialogue can deliver tangible outcomes. The broader China–US bilateral relationship also continues to be characterized by export controls and geopolitical and trade tensions.

2.2.2 United Kingdom

China and the UK held their inaugural intergovernmental dialogue on AI in Beijing in May 2025.[b] In addition to the expected launch of a US–China AI dialogue (see above), the China–UK intergovernmental dialogue on AI is the only active, publicly known, formalized intergovernmental channel between China and a major Western country on AI. However, it has not met in over a year.

In January 2026, UK Prime Minister Keir Starmer visited China where he met with President Xi Jinping. While the UK readout did not mention AI, the readout from the Chinese Ministry of Foreign Affairs noted that the two sides may "pursue joint research and commercial application in AI" and other areas.[140]

b. See our *State of AI Safety in China* report (2025), p. 26.

2.2.3 European Union

Across the seven Chinese readouts of President Xi's bilateral meetings with leaders of EU member states since July 2025, only three include any mention of AI (France, Germany, and Ireland).[141] All three refer to expanding cooperation in AI, but this is one item in a list of several areas. Only the readout of the meeting with French President Macron mentions the term "AI governance" (人工智能治理), in reference to France's stated commitment to strengthening cooperation with China on climate change, biodiversity, AI governance, and other areas.

At the EU institutional level, President of the European Council António Costa and President of the European Commission Ursula von der Leyen met with President Xi in Beijing during the 25th China–EU Summit in July 2025. The Chinese readout noted that China was "ready to conduct policy communication and practical cooperation with the EU in the field of AI." The EU–China High-level Digital Dialogue last met in 2023, following its first meeting in 2020.[142]

The EU AI Act's obligations for general-purpose AI models have applied since August 2025, with the Commission's enforcement powers taking effect from 2 August 2026.[143] Some of China's leading general-purpose AI developers could fall within the Act's "systemic risk" tier and would then face additional obligations beyond the baseline. How Chinese developers navigate these requirements as enforcement begins will be worth watching.

2.2.4 Singapore

In September 2025, China and Singapore held their second Digital Policy Dialogue (DPD) in Chongqing, co-chaired by Singapore's Senior Minister of State for Digital Development and Information, Tan Kiat How, and the Head of China's National Data Administration (NDA), LIU Liehong (刘烈宏).[144] While the inaugural DPD in 2024 planned to create an AI governance working group, there was no reference to it in the readout of the 2025 meeting, which emphasized digital economy development, trusted data flows, and industrial use cases.

2.2.5 Russia

Following their meeting in Beijing in May 2026, President Xi and President Putin issued a joint statement, which highlighted the role of AI in social and economic transformation. The statement expressed opposition to some countries' use of AI as a "geopolitical tool" to maintain "hegemonic status," called for international cooperation, and underlined the importance of responding to the potential risks and challenges associated with the technology.[145] President Xi had also met with President Putin in Beijing in September 2025, where the two sides signed over 20 bilateral cooperation documents, including about AI.[146] In November 2025, the Chinese and Russian premiers agreed to establish an AI Cooperation Expert Committee within the framework of the AI working group announced in 2024, which would advise on cooperation in AI ethics governance, standardization, and industrial applications.[147] According to the joint communiqué, the two sides also agreed to cooperate

on developing safe and trustworthy AI systems, as well as on China's initiative to establish a World AI Cooperation Organization.

2.3 International AI standards

Multiple high-level policy documents identify international standards cooperation as a key Chinese policy objective in the AI domain. In 2024, the *Guidelines for the Construction of a Comprehensive Standardization System for the National Artificial Intelligence Industry* set a goal of participating in the formulation of more than 20 international AI standards by 2026.[148] At an April 2025 Politburo study session on AI, President Xi encouraged "work toward an early formation of a consensus-based framework and standards for global governance." The 2025 Global AI Governance Action Plan (see above) similarly calls for more dialogue among state standard-setting bodies, discusses the role of international standards organizations, and proposes to speed up the formulation and revision of technical standards in key areas such as security, industry, and ethics.[149]

The Global AI Governance Action Plan argues that international standards are crucial for establishing "a scientific, transparent, and inclusive normative framework in the field of AI." It highlights the role of international standards in eliminating algorithmic bias and in balancing technological progress, risk prevention, and social ethics.

Chinese standard-setting bodies have also explicitly called for international cooperation on AI safety standards. The *AI Safety Governance Framework 2.0* released by TC260 in September 2025 discusses advancing international information sharing, as well as exploring the creation of international mechanisms and technical standards to prevent and respond to AI risks (Article 5.9).[150] Article 5.11 calls for clear requirements for AI in high-risk contexts such as CBRN, and recommends building international consensus for trustworthy AI.

In line with this high-level policy direction, China is actively engaged in several significant international standards bodies, including the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), and the International Telecommunication Union (ITU).

International AI standard-setting is to a large extent concentrated around the joint efforts of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), specifically through subcommittee (SC) 42 of joint technical committee (JTC) 1: ISO/IEC JTC 1/SC 42.[151] The Standardization Administration of China (SAC) represents China within ISO and IEC. The SAC is responsible for organizing the activities of the Chinese National Committee for ISO and IEC, and "approves and organizes the implementation of international cooperation and the exchange of projects on standardization."[152]

China is represented in the ISO/IEC JTC 1/SC 42 subcommittee via the China Electronics Standardization Institute (CESI), which is also the secretariat of the domestic mirror committee SAC/TC28/-SC42 (see Domestic Governance chapter). ISO/IEC publishes very little about membership and

participation, making it difficult to precisely document Chinese engagement. However, an expert from CESI holds the convenor position in Working Group 5 of ISO/IEC JTC 1/SC 42, which focuses on computational approaches and characteristics of AI systems.[153] According to our survey, Chinese experts have served as project editors for at least six published international standards under this ISO/IEC subcommittee.[154] Representatives of Chinese companies have also held key positions in these bodies, including the Chairperson of ISO/IEC JTC 1/SC 42.[155]

One of the international AI standards under development under ISO/IEC is *ISO/IEC TS 25568 – Guidance on addressing risks in generative AI systems*.[156] This standard will likely feature broadly applicable risk sources for generative AI and list broadly applicable risk controls for such systems.[157] Chinese state media noted that the standard was "initiated" by China, framing this as China's first time leading the construction of global governance rules in the generative AI domain.[158]

China is similarly engaged in AI-related standardization at the International Telecommunication Union's Telecommunication Standardization Sector (ITU-T). For example, in 2025, ITU-T published Recommendation F.748.44, "Assessment criteria for foundation models: Benchmark," which CAICT led in drafting.[159] It specifies metric requirements and testing methods for foundation model benchmarking, which CAICT presented as an effort to build international consensus on a benchmarking architecture. An expert from the Chinese Academy of Sciences Institute of Computing Technology serves as Vice Chair and Associate Rapporteur of ITU-T Study Group 21 Question 5, which is responsible for standards on AI-enabled multimedia applications, including evaluation and assessment specifications for such applications, algorithms, and data structures.[160]

Chinese experts also participated in the International Summit on AI Standards organized by IEC, ISO and ITU in December 2025.[161] The summit launched the Seoul Statement, in which IEC, ISO and ITU commit to "advance sustainable development and allow all people and society to benefit from the AI revolution."[162]

China-based organizations are also among the contributors to the World Digital Technology Academy (WDTA)'s AI STR (Safety, Trust, and Responsibility) standards series. In July 2025, WDTA released AI-STR-04, a Single AI Agent Runtime Security Testing Standard for autonomous single-agent systems, which WDTA presented as the first global benchmark of its kind. Its acknowledged contributors include the Cloud Security Alliance Greater China Region, Ant Group, China Telecom, Huawei, and Tsinghua University, alongside Western institutions.[163]

Beyond established standards-development organizations, Chinese companies are beginning to engage with newer, industry-led standards venues in the AI domain. In February 2026, Huawei and Lenovo joined the Agentic AI Foundation (AAIF). AAIF was established by OpenAI, Block, Anthropic, Google, Microsoft, Amazon, Bloomberg, and Cloudflare under the Linux Foundation to develop standards and protocols that enable AI agents to operate interoperably across platforms.[164] The organization was only established in December 2025, but it represents a noteworthy venue for potential

collaboration between Chinese and Western technology companies on agentic AI standards. Such initiatives could serve as a complement to traditional standard-developing organizations.

2.4 Non-official dialogues

Non-official forms of diplomacy and exchange, often referred to as "Track 1.5" and "Track 2" dialogues, continue to play an important role in fostering international coordination on AI safety and governance issues.

This section provides an overview of such initiatives between groups in China and Western countries. However, given the confidential nature of many such efforts, our analysis is necessarily limited to publicly documented dialogues and offers only a partial picture of the overall landscape.

Table 2.1 lists 12 publicly documented dialogues that have met on multiple occasions since 2023, and at least once over the past two years. The table divides the 12 dialogues into three primary categories:

- Five "Frontier AI" dialogues: 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.
- Four "AI-focused" dialogues: The name of the dialogue mentioned AI or the dialogue appears to focus >50% on AI.
- Three "AI-related" dialogues: Approximately 15–50% of content focused on AI.

Many of these dialogues explicitly address frontier AI safety issues. For example, the International Dialogues on AI Safety (IDAIS) has convened leading Chinese and international scientists around risks from advanced AI systems since 2023, the Tsinghua CISS–Brookings dialogue has published parallel glossaries defining concepts such as "catastrophic risk" and "AI controllability," and the AIxBio Global Forum has mobilized experts to consider the convergent risks of AI and the biological sciences.

2.4.1 Key outputs

Several dialogues published outcome documents on important AI safety topics, with contributions spanning AI and biosecurity governance, AI risks from non-state actors, frontier AI safety red lines, and shared terminology for AI risks.

Some of the most substantive outputs were related to biosecurity risks.

The AIxBio Global Forum — initiated by the US-based Nuclear Threat Initiative (NTI) — has continued to convene experts from many countries, including the US, China, India, and the UK, to address the convergent risks between AI and the biological sciences. In July 2025, more than 40 international experts, including George Church, ZHANG Weiwen (张卫文), Yoshua Bengio, Andrew Yao, ZENG

Yi (曾毅), and Concordia AI CEO Brian TSE (谢旻希), published a statement on biosecurity risks at the convergence of AI and the life sciences, in association with the AIxBio Global Forum.[165]

In February 2026, NTI issued a joint statement with the China Arms Control and Disarmament Association (CACDA) on shared priorities for strengthened biosecurity and responsible AI-biotechnology innovation.[166] The joint statement noted that recent US–China Track 2 biosecurity discussions had demonstrated significant agreement on the nature of the biosecurity risks facing both countries and the importance of forward-looking oversight and responsible practices to achieve their shared objectives.

In December 2025, experts from the international NGO INHR published recommendations to governments on mitigating AIxBio risks, which were endorsed by 14 organizations, including four Chinese organizations.[167] The recommendations are based on the input and insights provided by participants of a China-US-International trilateral dialogue organized by INHR and the Center for a New American Security, a think tank in Washington D.C.

The International Dialogues on AI Safety (IDAIS) convened in London in April 2026, marking the fifth iteration of the dialogue, which brings together some of the most respected scientists in China and the West. A statement from the meeting highlighted concerns around the proliferation of AI-enabled cyberattack and biological misuse capabilities to non-state actors.[168] It recommended that China and the United States, in particular, commit to prevention efforts. The statement was signed by leading scientists and governance experts including Yoshua Bengio, Andrew Yao, Zhang Ya-Qin, FU Ying (傅莹), and others.

IDAIS also convened in Shanghai in July 2025, culminating in the release of the "Shanghai Consensus," which warned of future loss of control risks and called for safety measures to be implemented. [169] While previous IDAIS statements had described risks as largely in the future, the 2025 statement noted that "some AI systems today already demonstrate the capability and propensity to undermine their creators' safety and control efforts," suggesting a heightened sense of urgency.

The US–China Dialogue on Artificial Intelligence and International Security, facilitated by Brookings and the Center for International Security and Strategy (CISS) at Tsinghua University since 2019, remains one of the most significant and longest-running dialogues. The 13th and 14th rounds of the dialogue were respectively held in Dubai (November 2025) and Munich (February 2026).[170] In January 2026, Brookings published a collection of essays authored by members of both the US and Chinese delegations exploring whether the United States and China could coordinate to address risks of AI misuse by non-state actors, such as AI-enabled cyberattacks or biological and chemical weapons development.[171] Additionally, the CISS–Brookings AI glossary, first published in parallel by both institutions in August 2024, was updated in January 2026 with a definition of "loss of control" (失控).[172]

In February 2026, Tsinghua University's CISS held dialogues on AI and international security in

Munich with the Brookings Institution and the Centre for Humanitarian Dialogue, which featured discussions on the risks of AI misuse by non-state actors and the governance of such risks. Based on those exchanges, Tsinghua CISS published a report focused on exploring the possible pathways through which non-state actors may misuse AI, the associated risks and mitigation measures, as well as the potential implications for international security.[173]

Although these dialogues are not official government discussions, these increasingly substantive non-official outputs may help lay the groundwork for future government-to-government engagement.

Note: This table was compiled through online research and only includes dialogues that are publicly documented. It therefore likely undercounts existing dialogues, and there may be any number of confidential dialogues on AI. Some of these dialogues are formally described as "Track 1.5" or "Track 2" dialogues. Others do not categorize themselves as such, but meaningfully complement official government-to-government channels. This table was last updated on June 22, 2026, and only includes dialogues that have met multiple times, and at least once over the past two years.

Table 2.2: Non-official dialogues on AI.

Name	Type	Area of focus	Participants	Chinese convenor	Foreign convenor	Meeting history	Specific topics
US–China 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 February 2026; 14 meetings.[174]	Core terminology and concepts for AI risks; AI risk assessment and hierarchy; frontier/advanced AI risks; simulations of AI in military command and control; non-state actors misuse.
INHR Expert-level AI Dialogue	Track 2: Frontier AI	AI and National Security, including Military and AIxBio	Dialogue NGO, think tanks, former government officials		INHR,[d] CNAS[e]	Since 2019, through 2025, with monthly meetings.[176]	Use of AI in nuclear systems; advanced and agentic military AI; 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.[177]	Large language models, foundation models, and generative AI challenges and governance; best practices for implementing ethical principles and regulations related to AI governance.

Continued 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.[175]
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.

Table 2.2: Non-official dialogues on AI. *(continued)*

Name	Type	Area of focus	Participants	Chinese convenor	Foreign convenor	Meeting history	Specific topics
International Dialogues on AI Safety (IDAIS)	Track 2: Frontier AI	AI safety	Scientist, industry, former government officials	Andrew Yao, Zhang Ya-Qin[g]	Yoshua Bengio, Stuart Russell[h]	From October 2023 to April 2026; five meetings.[178]	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.
AIxBio Global Forum	Track 2: Frontier AI	AI and biosecurity	Think tanks, industry, scientists		Led by NTI[i]	From April 2024 to October 2025.[179]	Best practices for safeguarding AIxBio capabilities, such as guardrails for biological design tools.
	Track 2: AI-focused	AI ethics	Scientists, technologists, ethicists and policy professionals.	CAS[j]	Royal Society (UK)	From September 2020 to January 2026; 4 meetings.[180]	AI ethics and safety; AI progress and challenges; AI applications in science.
Sino–European Dialogue on AI and International Security	Track 2: AI-focused		Dialogue NGO, think tanks	CISS	Centre for Humanitarian Dialogue (HD)	From February 2024 to February 2026; five meetings.[181]	Military AI; confidence-building measures; misuse by non-state actors; 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 December 2024; three meetings.[182]	Multiple sessions on AI and nuclear command, control, and communications.

Continued on next page

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.
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.

Table 2.2: Non-official dialogues on AI. *(continued)*

Name	Type	Area of focus	Participants	Chinese convenor	Foreign convenor	Meeting history	Specific topics
Crisis Management for Advanced AI	Track 2: AI-focused	AI and crisis management	Think tanks, researchers, industry, former government officials, international organizations	Concordia AI, Tsinghua CISS, Tsinghua I-AIIG	Carnegie Endowment for International Peace, the Oxford Martin AI Governance Initiative, Oxford China Policy Lab	From February 2025 to February 2026; three meetings.[183]	International coordination on response and preparedness for AI-enabled crises.
NTI and CACDA Track II Biosecurity Dialogue	Track 2: AI-related	AI and biosecurity	US and Chinese experts and scholars from academia, industry, NGOs	CACDA	NTI	April 2024, February 2025, November 2025; 3 meetings.[184]	Opportunities for collaboration related to safeguarding capabilities at the intersection of AI and the life sciences.
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 2025; 8 meetings.[185]	Digital economy; AI; semiconductors; electronic and intelligent connected vehicles.
China-Europe Cyber Dialogue	Track 1.5: AI-related	Cyberspace	Think tanks, government officials	CICIR[m]	Geneva Centre for Security Policy, previous convenors include ESMT Berlin and the Hague Centre for Strategic Studies	March 2014 to November 2025; 12 meetings.[186]	AI regulatory oversight; data governance; cyberspace governance; critical infrastructure protection.

k. The China–US 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.
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

Technical Safety Research

This chapter reviews Chinese technical AI safety research from June 2025 to April 2026. It starts with overall output trends and key research groups before delving into selected research contributions in agent safety, AI-generated content detection, mechanistic interpretability, and CBRN risks.

Key updates from June 2025 to April 2026

- **Chinese frontier AI safety research output has grown substantially**, with monthly output up roughly 60% since June 2025.
- **Agent safety has become the single most active area of Chinese AI safety research.** It now accounts for almost 27% of new papers — up from 8% in Q1 2025 — spanning operational failures, alignment challenges, and a growing body of work on loss-of-control risks such as self-replication and multi-agent collusion.
- **AI-generated content detection and watermarking remains a consistently high-output area,** accounting for roughly 11–13% of all safety papers each quarter.
- **Mechanistic interpretability research has grown** from a niche topic in 2024 to 26 papers in Q1 2026, accounting for roughly 15% of all papers that quarter.
- **CBRN risks appear as a standard component of major Chinese AI safety frameworks, but dedicated research on the topic remains rare**, with just a couple of papers primarily focused on CBRN misuse risks each year.

3.1 Methodology

The analysis in this chapter is based on Concordia AI's Chinese Technical AI Safety Database, which has collected over 900 technical papers published by Chinese institutions on arXiv from April 2023 through April 2026.[a]

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.[b] For further details on the methodology and definitions of these research directions, please see the full Methodology on our companion website.[187]

a. The full database is available on this report's companion website at https://aisafetychina.com/research/.

b. It does thus not extend to other important areas of AI research, such as bias and discrimination, which fall under distinct fields with their own substantial literatures.

3.2 Overall trends

Chinese frontier AI safety research grew substantially over the past year. Between June 2025 and April 2026, monthly output rose by roughly 60%: the 3-month trailing moving average climbed from 36 papers/month in June 2025 to 57 in April 2026.

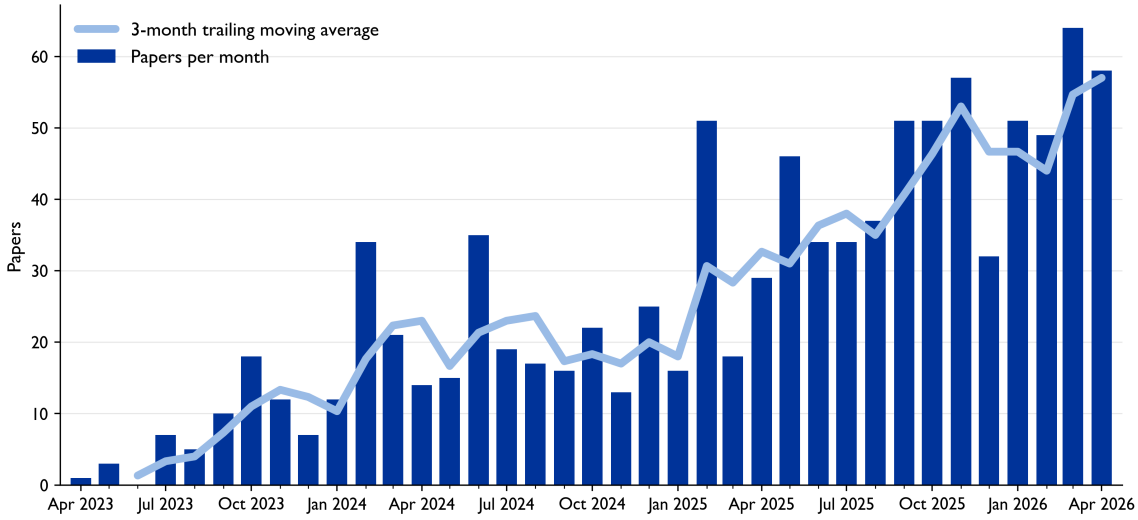


Figure 3.1: Papers per month.

3.3 Key research groups

We identify "Key Chinese AI Safety-Relevant Research Groups" as groups that contain at least one senior researcher who supervised five or more frontier AI safety papers between January 2025 and April 2026 as anchor author. This threshold ensures that the groups have a strong focus on AI safety rather than sporadic engagement. 28 groups meet this bar, 15 of which were already identified in our previous *State of AI Safety in China (2025)* report. The threshold used this year is more stringent than last year's, meaning the numbers of groups in our 2025 and 2026 reports are not directly comparable.[c] Readers can explore the full list of groups and anchor authors on the companion website under the "Key Research Groups" tab.[188] The remainder of this section explores aggregate observations about these groups.

c. This criterion is significantly more stringent than the one used in last year's report, when groups only required an anchor author with at least three papers. Applied to the current dataset, that earlier bar would yield over 100 qualifying anchor authors — far too many to give a meaningful signal about where AI safety research in China is actually concentrated. Accordingly, the total number of research groups is not directly comparable to last year's. However, the fact that the more stringent criterion yields a similar number of groups to last year highlights the overall increased work in this space.

AI safety research groups continue to be concentrated primarily in universities, with 20 out of the 28 attached to universities. This may partially reflect the stronger incentives in universities to publish papers, compared to corporate institutions. Several of China's top universities are especially well-represented: Tsinghua University, Fudan University, and the Hong Kong University of Science and Technology each have three distinct AI safety labs in the list, while Peking University, the University of Science and Technology of China, and Renmin University of China each have two. Tsinghua's three labs split focus between adversarial robustness and reasoning safety (TSAIL), alignment and jailbreak defense (CoAI), and search-agent and text-to-image red-teaming (INSC); Fudan's three groups engage in NLP-focused work on deception and oversight (FudanNLP), vision-language safety and backdoor attacks (FVL Lab), and software security-style probing of frontier AI (SecSys Lab).

Only three of the 28 groups are at state-backed labs. However, their output is disproportionately high: Shanghai AI Laboratory alone is affiliated with at least one co-author on roughly one in ten papers in the database, and is the research group with by far the most papers in the database.

There are three groups in private industry: two distinct groups at Alibaba — ZHENG Bo's (郑波) group, focused on multimodal alignment and AI-generated image detection, and the AI Governance Laboratory, working on jailbreak defenses and constructive safety alignment — and the Societal AI team at Microsoft Research Asia, which focuses on value alignment, pluralism, and moral reasoning. For the first time in our list, there is also a group at a state-owned enterprise — the Institute of Artificial Intelligence at China Telecom (TeleAI). Its eight papers since 2025 focus primarily on the safety of multimodal LLMs and vision-language models (5/8), with additional work on AI-generated image detection.

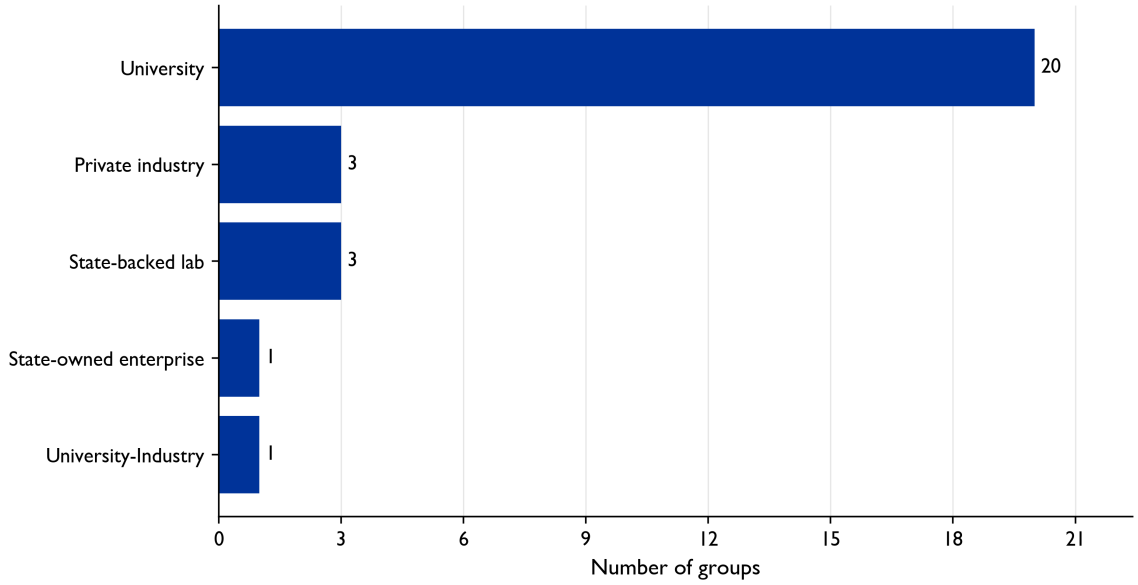


Figure 3.2: Key research groups by type.

Geographically, these groups cluster in China's main economic hubs — Beijing, Shanghai, Hangzhou, and the Greater Bay Area (Hong Kong and Guangzhou). Beijing accounts for 13 of the 28 groups, more than any other city. This reflects its concentration of top universities (Tsinghua, Peking, Renmin, Beihang, and the Beijing University of Posts and Telecommunications), but also its role as the home of major state-backed labs and the headquarters of corporate AI safety teams like Microsoft Research Asia and TeleAI.

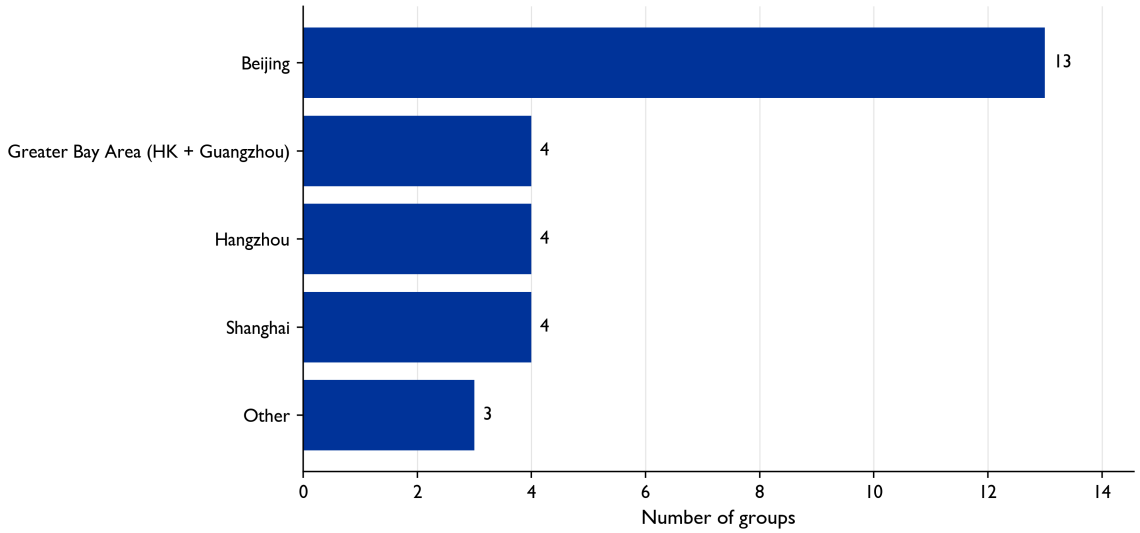


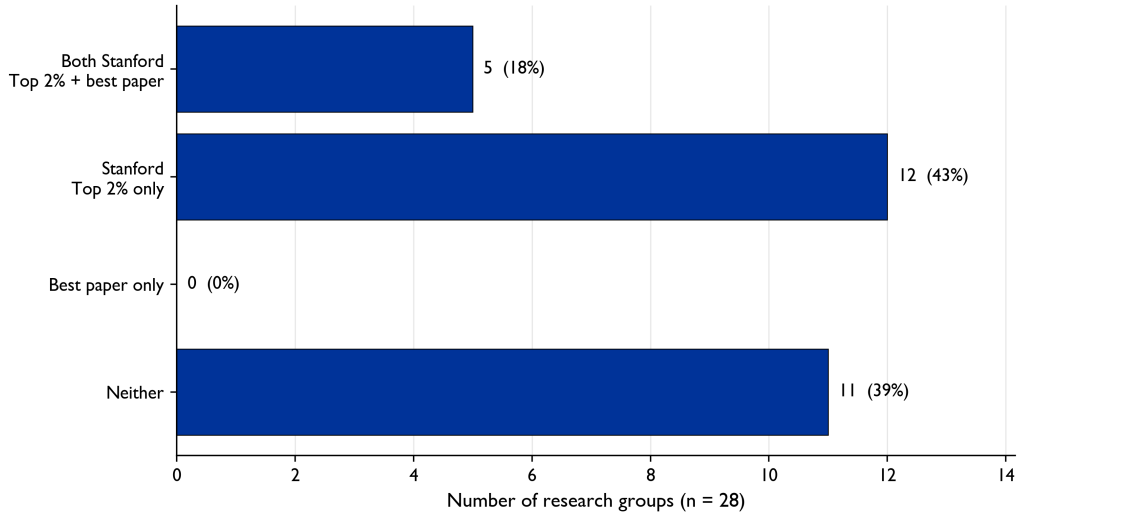
Figure 3.3: Key research groups by location.

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. Hence, it is not possible to verify the quality of every single paper individually. Nevertheless, we can proxy the general quality of these key research groups by whether they have 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;[d] OR
- whether authors were ranked in the top 2% of scientists in their field by Stanford University's citation-based assessment.[189]

17 of the 28 groups have at least one senior staff member with one of these honors.

d. 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.



Stanford Top 2%: career or single-year list, August 2025 release (publications through 2024). Best paper: outstanding/finalist/honorable-mention at NeurIPS, ICML, ICLR, ACL, EMNLP, CVPR, ICCV, ECCV main tracks.

Figure 3.4: Honors recognition across key Chinese AI safety research groups.

3.4 Specific research contributions

This section reviews research contributions across a few specific areas: agent safety, AI-generated content detection, mechanistic interpretability, and CBRN risks. We selected these topics based on what we thought were major trends in Chinese technical research over the past year. Most relate to severe risks from advanced frontier systems, though we also include AI-generated content detection given its broader prominence. We want to emphasize that this selection is not exhaustive. Chinese researchers addressed a much wider range of AI safety topics over the past year than we cover here, and within each topic we make no claim to have captured every relevant research output.

3.4.1 Safety of autonomous AI agents

As described in the Domestic Governance and Expert Views chapters, attention to safety risks from autonomous AI agents has expanded significantly in China over the past year, in part due to the widespread adoption of agentic products like OpenClaw. This trend is also visible in technical research output.

The number of AI agent safety papers has risen sharply, both in absolute terms and as a share of total output.[e] Q1 2025 saw just 7 agent-safety papers in the database — about 8% of all papers that

e. Topic classification was performed by an LLM-based agent (Anthropic's Claude Sonnet 4.6 family), which was given each paper's title and abstract together with an explicit inclusion/exclusion prompt for the category. The model was prompted to return a confidence label ("high," "medium," or "low") along with each classification. To reduce the risk of

quarter. By Q1 2026 this had grown to 44 papers, accounting for almost 27% of new papers that quarter. In other words, agent safety research has roughly tripled as a share of Chinese frontier AI safety output, while its absolute volume has grown more than six-fold.

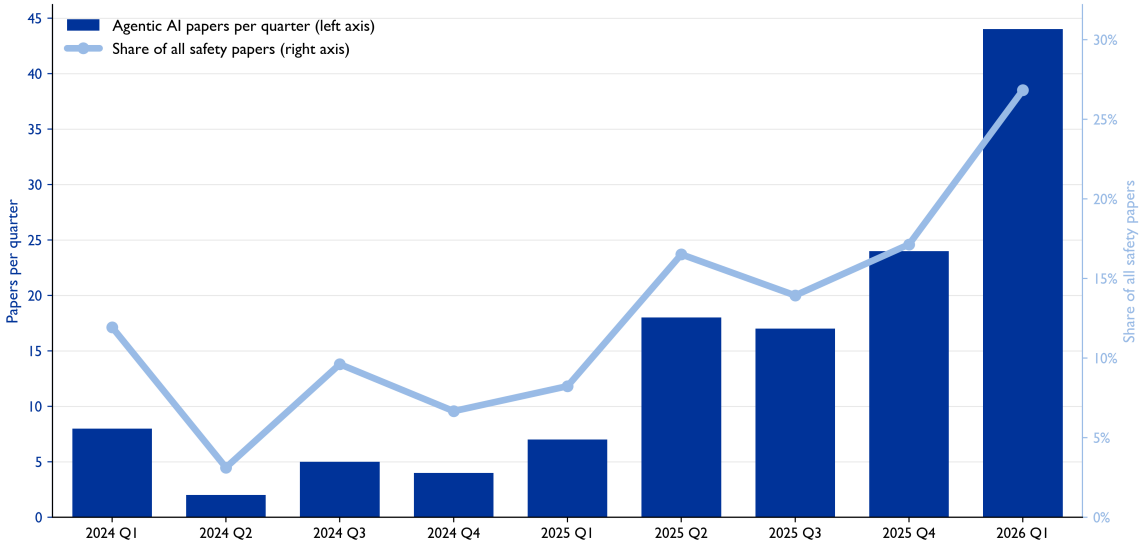


Figure 3.5: Agentic AI papers: quarterly count and share of safety output.

Given the volume, we can only highlight a few examples. The majority of papers focus on operational failures and the alignment challenges unique to agentic AI.

One strand of research probes the attack surface: SkillJect poisons the skill packages that coding agents rely on, achieving an 80.7% attack success rate,[190] while Fudan SecSys's "When Bots Take the Bait" studies social-engineering attacks against web-automation agents.[191]

Another strand builds defenses: Shanghai AI Lab's "AgentDoG" is a diagnostic guardrail framework that explains why agents fail, extended to multi-agent settings in the same lab's "TrinityGuard";[192]

false positives, only papers the model flagged as "high"-confidence matches are counted in the analysis. A sample of around 20% of matches per category was manually spot-checked to validate reasonableness. Both false positives and false negatives are still possible in principle but appear infrequent based on these checks. The full prompt used for this classification was: "INCLUSION criteria (mark is_match=true): Work that studies the safety, alignment, or robustness of AI agents — single-agent or multi-agent systems that take actions in an environment, use tools, browse the web, control computers, run code, or coordinate with other agents. Includes attacks against agents (prompt injection through tools, skill / tool poisoning, social engineering of automation agents), defenses (guardrails, monitors, incident response, trajectory auditing), benchmarks for agent safety, and frontier-flavored agent risks like self-replication, multi-agent collusion, self-evolution, and emergent extreme events in agents. EXCLUSION criteria (mark is_match=false): Does NOT include general LLM safety work that doesn't specifically address agentic deployment. Does NOT include multi-agent debate or multi-agent collaboration when used purely as a method for some other safety goal (e.g., evaluating content safety via multi-agent debate) and the agent deployment is not itself the safety concern."

a Tsinghua–Ant Group collaboration ("Taming OpenClaw") proposes a five-layer lifecycle security framework spanning initialization, input, inference, decision, and execution;[193] and Tianjin University's AIR (Agent Incident Response) detects, contains, and recovers from agent failures at runtime.[194]

A smaller cluster of papers engages with loss-of-control risks. Two broader safety evaluation frameworks include agentic risks like uncontrolled autonomous R&D, self-replication, and collusion in their risk categorizations, namely the *Frontier AI Risk Management Framework* by Shanghai AI Lab and Concordia AI, and ForesightSafety Bench by the Beijing Institute of AI Safety and Governance.[195] A smaller cluster of dedicated papers engages with these risks in more detail. Shanghai AI Lab's "Dive into the Agent Matrix" builds a four-stage evaluation framework for uncontrolled self-replication and finds that over half of 21 tested LLMs exhibit tendencies toward it.[196] Other papers cover multi-agent collusion in social systems (Shanghai AI Lab's "When Autonomy Goes Rogue"),[197] risks of self-evolving agents (Fudan and HKUST's "Your Agent May Misevolve"),[198] and rare high-impact failures viewed through a mechanistic interpretability lens (Fudan and Shanghai AI Lab's "Interpreting Emergent Extreme Events in Multi-Agent Systems").[199] Alibaba, HKUST, Peking University, and Shanghai AI Lab's "Are Your Agents Upward Deceivers?" benchmarks whether agents hide their failures and fake success when they can't honestly complete a task.[200] A Beijing University of Posts and Telecommunications and China Mobile team documents "Machiavellian helpfulness" — agents prioritizing task completion over ethical constraints — and a Fudan University paper finds that weaker agents cause harm through incompetence while stronger ones disguise misaligned actions as legitimate behavior.[201]

Overall, agent safety has become one of the most active areas of Chinese AI safety research, mostly grounded in operational failures and alignment but with a sizable body engaging loss-of-control risk scenarios.

3.4.2 AI-generated content detection and watermarking

AI-generated content detection and watermarking research has held a consistently high share of Chinese frontier AI safety output over the past year.[f] From Q2 2025 onwards each quarter has seen 13–20 such papers — between roughly 11% and 13% of all safety papers in the database for that same timeframe. Absolute volume reached a record 20 papers in Q1 2026. The strong interest could be driven in part by China's mandatory labeling regime for AI-generated content (see section 1.3.2).

f. The full prompt used for the LLM-based classifier was: "Work that detects AI-generated content (images, video, audio, text, multimodal) or that watermarks AI outputs to enable later detection / provenance verification. Includes attacks on detectors or watermarks (adversarial evasion, watermark removal/spoofing) and defenses, plus benchmarks for AI-generated content detection or watermarking. EXCLUSION criteria (mark is_match=false): Does NOT include general content moderation or harmful-content filtering that doesn't specifically focus on AI-generated provenance."

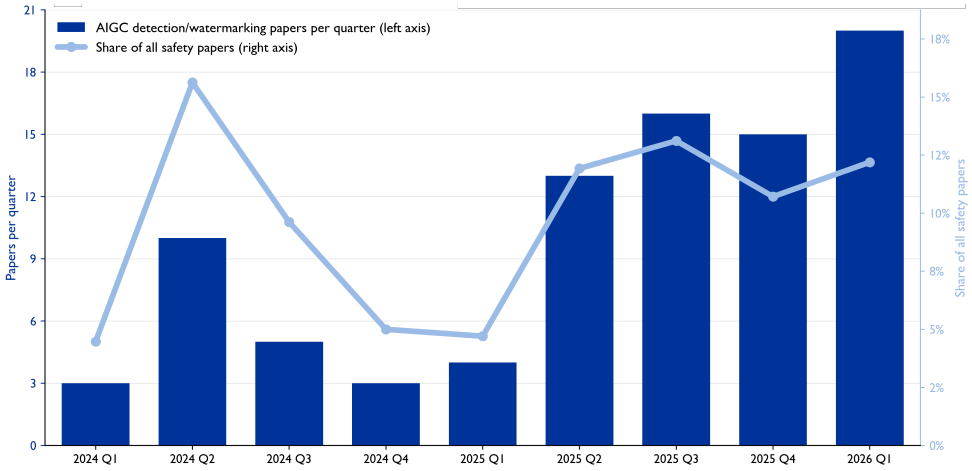


Figure 3.6: AI-generated content detection and watermarking papers: quarterly count and share of safety output.

Within this space, several Chinese groups are producing substantive work. A research group at Alibaba under Zheng Bo released MIRAGE, an "in-the-wild" benchmark plus a VLM-based detector designed to handle the messy real-world images.[202] Fudan's FVL Lab under Ma Xingjun (马兴军), with Jiangnan University, proposed RA-Det, which detects AI-generated images by how much they change when slightly perturbed.[203] REN Kui's (任奎) group at Zhejiang University's State Key Laboratory of Blockchain and Data Security diagnosed a failure mode in CLIP-based detectors which they call "semantic fallback" — where detectors classify images based on their content rather than forensic traces of AI generation — and proposed a fix requiring no additional training.[204] USTC's Information Hiding Lab (CAS Key Lab of Electromagnetic Space Information), led by Nenghai Yu and Weiming Zhang, has produced a series of papers, including work on tamper-aware generative image watermarking, balancing robustness and diversity in noise-as-watermark for diffusion models, and extending generative watermarking to text-to-video diffusion.[205]

3.4.3 Mechanistic interpretability

The category was modest throughout 2024 — averaging around four to five papers per quarter — but grew substantially through 2025, reaching a record 26 papers in Q1 2026 (about 16% of all safety output that quarter).[g] With 99 papers since 2023, the overall cluster is sizable but still smaller than

g. The full prompt used for the LLM-based classifier was: "Include: Work that studies or intervenes on the internal computations of a large language model or large multimodal foundation model — activations, neurons, attention heads, residual-stream features, circuits, SAE features, steering vectors, value vectors, probing classifiers, activation patching, causal interventions, representation engineering — AND has a safety/alignment motivation: jailbreaks, refusal, deception, sycophancy, emergent misalignment, harmful-content generation, bias/toxicity, value alignment, harmful-knowledge unlearning, safety degradation under fine-tuning, or moral reasoning. Exclude: Does NOT include 'interpretable' / 'explainable' work that stays at the input or output level (input-feature attribution, natural-language rationales, prompt-based 'interpretability'). Does NOT include mech-interp-style analyses of small classifiers, toy networks, or non-LLM architectures. Does NOT include pure-capability mech interp without a safety motivation (induction heads, in-context learning circuits, factual

agent safety (144 papers across the database) or AI-generated content detection and watermarking (105 papers).

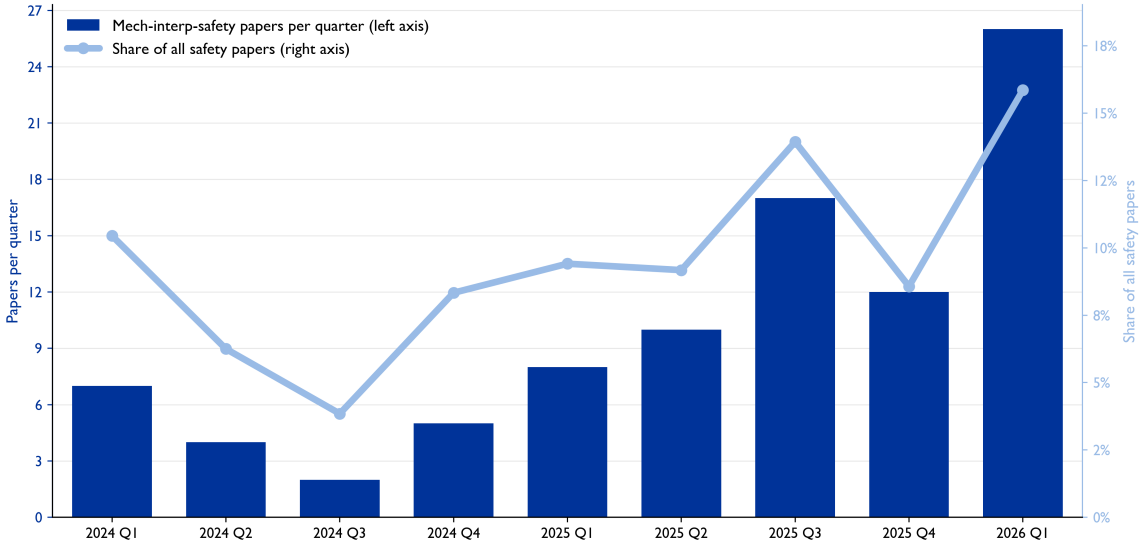


Figure 3.7: Mechanistic interpretability for safety: quarterly count and share of safety output.

Within this space, several Chinese groups are producing substantive work. A Tsinghua–USTC collaboration offers a mechanistic account of emergent misalignment, arguing that narrow fine-tuning installs stable behavioral dispositions.[206] A team from Peking University used representation engineering to induce strategic deception in reasoning models — where chain-of-thought contradicts the output — at a 40% success rate via activation steering.[207] Alibaba's AI Governance Laboratory and Renmin University released Safe-SAIL, a suite of sparse autoencoders that translate an AI model's internal activity into readable labels, identifying 1,758 features relevant to safety spanning pornography, politics, violence, and terror.[208] A Chinese University of Hong Kong–IBM collaboration shows that activation-steering interventions derived from entirely benign datasets — e.g. JSON output formatting — erode safety guardrails, pushing jailbreak attack success rates above 80%.[209]

These are just some examples in a large body of work that also includes training-free safety projection methods from ZENG Yi's (曾毅) group at the Chinese Academy of Sciences,[210] robust unlearning via irreversible projective transformations (PKU–Tsinghua),[211] and moral-feature analysis based on sparse autoencoders at Shanghai AI Lab.[212]

recall, math reasoning circuits, multilingual representations). Does NOT include AI-generated content detection or watermarking even when described as 'interpretable' or 'mechanistic'. Does NOT include general XAI surveys that don't engage with specific internal mechanisms." We manually removed a small number of false positives.

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

There is a persistent body of research on AI safety in CBRN domains, though this remains a more niche topic compared to the other risks discussed in this chapter. Across the entire database, our LLM-based classification identified only nine papers as primarily focused on CBRN-related safety with "high" confidence, compared to 144 on agent safety and 105 on AI-generated content labeling and detection.[h] However, there is a larger number of papers that include CBRN risks in wider risk frameworks.

A number of prominent Chinese papers over the past year cover CBRN risks as a component of broader safety frameworks. The "Frontier AI Risk Management Framework in Practice" report series by Shanghai AI Lab and Concordia AI includes biological and chemical risks as one of seven defined risk dimensions, with frontier model evaluations on biological protocol troubleshooting and reasoning about hazardous chemical knowledge.[213] The Beijing Institute of AI Safety and Governance's "ForesightSafety Bench" likewise tests models for genomic misuse, biological lab safety protocols, pathogen design, protein toxin knowledge, and synthesis of harmful chemicals under its "AI4Science Safety" pillar.[214] This shows that CBRN risks are an established component of the overall safety taxonomies of leading state-backed Chinese labs. However, none of these papers develops novel CBRN evaluation methods. For example, Shanghai AI Lab's "DeepSight" safety toolkit falls back on the Western-developed WMDP (Weapons of Mass Destruction Proliferation) benchmark rather than a Chinese-developed counterpart.[215]

A small number of Chinese papers do go further by focusing specifically on safety in CBRN domains. Shanghai AI Lab and ByteDance (March 2026) released a 250,000-sample evaluation benchmark and a 1.5 million-sample fine-tuning dataset spanning multiple scientific domains, arguing that whether a scientific question is "safe" depends on the specific context and intent, rather than applying blanket rules based on topic.[216] A March 2026 collaboration spanning several of our identified Key Research Groups — including Fudan University's FVL Lab and City University of Hong Kong — introduces the concept of "Internal Safety Collapse": a failure mode in which frontier LLMs spontaneously generate CBRN-relevant content (pathogen sequences, toxin structures, exploit payloads) not through adversarial prompting, but as a byproduct of legitimate professional workflows, achieving safety failure rates averaging 95.3% across leading frontier models.[217]

h. The full prompt used for the LLM-based classifier was: "INCLUSION criteria (mark is_match=true): Work that specifically addresses CBRN-relevant misuse risks of AI systems: biological knowledge or capabilities that could enable bioweapon development (pathogen design, protein-toxin knowledge, biological lab protocols), chemical synthesis of dangerous substances, radiological or nuclear material handling, and benchmarks/methods for measuring or constraining frontier models in these domains. Also includes safety-driven alignment of domain-specific scientific models (protein language models, molecular generators) to prevent harmful outputs, and CBRN-flagged subsections of broader frontier risk frameworks. EXCLUSION criteria (mark is_match=false): Does NOT include general LLM safety / alignment work even when applied to scientific domains, unless it specifically addresses CBRN-relevant knowledge or capabilities. Does NOT include operational safety of autonomous laboratories (industrial-accident style — wrong chemicals being mixed, robots executing unsafe commands) when there is no misuse dimension. Does NOT include generic harmful-content moderation that doesn't specifically address CBRN. Does NOT include AI for scientific discovery, drug discovery, or lab automation when safety isn't the focus."

Some work focuses on biological design tools rather than general-purpose LLMs: a Zhejiang University team uses a protein safety knowledge graph to steer protein language models such as ProtGPT2, Progen2, and InstructProtein away from harmful outputs.[218] Another September 2025 paper, co-authored by researchers at Peking University, Shanghai Jiao Tong University, Zhejiang University, Stanford, and Princeton, introduces a dedicated red-teaming framework and a benchmark of 429 experimentally confirmed toxins and viral proteins, finding jailbreak success rates of up to 70% on widely used protein foundation models including ESM3 and DPLM2.[219] A separate cluster of Chinese papers focuses on the operational safety (e.g. avoiding accidents) of AI systems inside autonomous scientific laboratories rather than on misuse prevention.[220]

In conclusion, CBRN risks are recognized as a standard element in comprehensive frontier AI safety frameworks by leading Chinese institutions. At the same time, dedicated papers on the topic remain relatively rare.

Expert Views on AI Safety and Governance

This chapter surveys Chinese expert views on AI safety since our last report in July 2025. It is divided into four sections: safety risks from agentic AI, open source AI governance, AI's impact on cybersecurity, and AI's risks for biosecurity.

Key updates from July 2025 to June 2026:

- **Agentic AI safety emerged as a distinct strand of Chinese expert discourse.** Leading academics addressed issues such as the erosion of human oversight, recursive self-improvement, and the risks of giving agents physical instantiation.
- **Chinese experts tended to favor open source AI and were generally skeptical of governance proposals that could stifle innovation.** However, some experts also proposed building capacity for tighter oversight, including risk-assessment centers, and adopting a more cautious approach to certain frontier open-weight models.
- **In the cyber domain, Chinese experts often framed AI as a "double-edged sword" that can both lower the barrier to attacks and offer defensive tools.** The arrival of new frontier models from Anthropic and OpenAI, reported to discover and exploit zero-day vulnerabilities at scale, generated concerns among some that the most advanced capabilities are exclusive to Western actors.
- **Several Chinese experts published on biosecurity risks from advanced AI.** Experts analyzed the risks from AI tools in the life sciences (such as biological design tools and autonomous laboratories), proposed mitigation measures for high-risk biological AI models, and highlighted the importance of international cooperation.

4.1 Methodology

Chinese expert views on AI safety warrant close attention, not least because the government regularly draws on them in formulating policy. This reflects a broader feature of China's policymaking ecosystem: academic and research institutions play a comparatively prominent role in advising the government on technical and regulatory questions. In China, scholars are regularly tasked to conduct research studies on specific policy issues or submit policy recommendations. Scholars are also

regularly called to brief top leadership bodies, such as the State Council and the Communist Party of China (CPC) Politburo, China's top 24 officials, including on AI.[a]

This chapter provides an overview of thinking among leading Chinese academic and industry experts on four topics: safety risks from agentic AI, open source AI governance, AI's impact on cybersecurity, and AI's risks for biosecurity.

Over the past year, we have closely monitored Chinese AI safety discourse, and we have selected these topics based on what we thought were major trends. We have carried over open source AI governance, cybersecurity, and biosecurity from our 2025 report. Agentic AI safety is included as a standalone topic for the first time, reflecting its emergence as an increasingly important strand of expert discourse, and following the rapid development of agentic AI systems.

We want to emphasize that this discussion is not exhaustive. Chinese experts addressed a much wider range of AI safety topics over the past year than we cover here, and within each topic we make no claim to have captured every relevant output.

However, within these four themes, we have prioritized what we deem most substantive and relevant to severe global risks from AI. Our sources include Chinese-language academic articles, think tank and industry publications, and policy-oriented commentary.

4.2 Safety risks from agentic AI

Chinese experts have produced a range of outputs on agentic AI safety in the past year, which many of them described as an inflection point for AI agents.[221] Leading academics and influential legal scholars have written on agent autonomy, loss-of-control risks, and operational safeguards, and the topic has featured in Chinese industry security publications as well. While agentic AI safety did not warrant a dedicated section in our previous reports, it is now an increasingly important strand of Chinese expert discourse on AI safety and governance.

Several Chinese experts have expressed concern that human oversight could erode as agentic systems gain autonomy. ZHENG Nanning (郑南宁), an academician of the Chinese Academy of Engineering and former president of Xi'an Jiaotong University who briefed the Politburo on AI in 2025, has argued that the paradigm shift from "model-driven" to "intent-driven" intelligence introduces distinctive risk. In this context, he has underscored that AI is now exhibiting the potential for recursive self-improvement, with each generation of intelligent agent driving the evolution of the next, warning that this could result in AI evolving beyond the bounds of human prediction and control.[222] Turing Award Winner and Tsinghua Dean Andrew Yao (姚期智) observed that two years ago, the idea that "AI would compete with humans" was still a matter of academic discussion, but over the past year the industry has already seen considerable "deceptive behavior" by foundation models. He cited an

a. See the 2025 *State of AI Safety in China* report for details on the most recent Politburo study session on AI.

example in which a model, to avoid being shut down, accessed internal emails and used them to threaten the person responsible for that decision.[223]

Influential Chinese scholars, along with leading technology companies including Alibaba, Huawei, DeepSeek, and others, also contributed to the drafting of a non-binding document on AI ethics, formulated under a major standard-setting body. The document highlights several concrete loss-of-control concerns alongside more conventional AI ethics issues.[224] It proposes that developers exercise caution in building applications with a high degree of autonomy and that they prioritize assessing such applications' loss-of-control risk. It also asks service providers to ensure that users can forcibly halt a service with clear "off switches," and urges them to guard against AI escaping human supervision or threatening human survival and development.

These concerns also feature in the Singapore Consensus on Global AI Safety Research Priorities, which highlights research on monitoring and controlling general-purpose AI agents that act autonomously with limited human oversight as a research priority. Frontier Chinese labs and prominent academic experts contributed to the Companion Report on Agentic Risk Management, formulating best practices for mitigating risks from general-purpose AI agents.[225]

Another noteworthy concern is the risks that emerge when agentic AI systems are given physical instantiation and operate with increasing autonomy in real-world environments. The 2025 annual report from Tsinghua University's Institute for AI International Governance (I-AIIG) highlighted that the explosive emergence of agents and embodied intelligence will give rise to entirely new risks around responsibility and loss of control.[226] It noted that the deceptive tendencies of highly autonomous AI, the crisis of trust caused by black-box decision-making, and the thorny problems of attributing responsibility when embodied intelligence operates in real-world environments are all obstacles that must be overcome before deployment at scale. ZHANG Ya-Qin (张亚勤), Chair Professor of AI Science at Tsinghua University, and Founding Dean of the Institute for AI Industry Research of Tsinghua University (AIR), has similarly warned of the loss-of-control and misuse risks of connecting agents to autonomous vehicles, robots, and drones.[227] These concerns are particularly notable given China's leadership in deploying embodied AI systems, from autonomous vehicles to humanoid robotics. Zhang has suggested that key components of addressing agentic risks would include sandbox deployment, agent IDs, and having a human-in-the-loop.[228]

Agent governance also appears to be becoming more salient in influential corners of legal academia. In a March 2026 article that was widely reshared in Chinese state media, a group of legal scholars that has been particularly active in discussions around a potential comprehensive AI law (see section 1.2.5) highlighted behavioral boundaries, authorization mechanisms, and competition governance of AI agents as key issues.[229] In 2025, members of the same group contributed to a report on AI agent development and governance, which included references to risks of reduced human oversight and to the cognitive effects of interacting with agents, including "cognitive substitution" and "soft manipulation."[230]

The rising popularity of AI agents and recent policy announcements (see Domestic Governance chapter), coupled with these outputs across industry and academia, suggests that attention to agentic AI risks will remain high in China for the foreseeable future.

4.3 Open source AI governance

Chinese open source models have seen rapidly growing global adoption.[b] From February 2025 to February 2026, models from Chinese developers accounted for the largest share of downloads on Hugging Face, the leading open model repository, surpassing those from the United States.[231] As these models diffuse more widely, how Chinese experts think about governing them becomes more important — both for China's own trajectory and for the global ecosystem that increasingly builds on Chinese models. At the same time, Chinese AI models are far from uniformly open source: ByteDance, for instance, has kept the proprietary models powering its Doubao (豆包) app closed. As the Chinese AI ecosystem continues to develop, it is possible that more labs will take a hybrid approach. For example, Moonshot AI CEO YANG Zhilin (杨植麟) has said the company hopes to continue open sourcing over the long term but may not remain exclusively open, noting that some work, such as collaborations with certain partners, might not be released.[232]

Chinese experts continue to broadly endorse open source AI development. Their support is typically grounded in the logic of innovation and ecosystem development, including the benefits of scaling the technical community working on AI and the merits of inclusive diffusion of the technology.[233] A recurring view across Chinese expert outputs is that open source governance needs to balance inclusive innovation and risk management. Many Chinese experts acknowledge the risks associated with open source models but treat them as manageable through governance rather than as grounds for restricting open sourcing itself. Some Chinese experts argue that open sourcing is not merely compatible with AI safety but conducive to it. In this view, transparency means that bias and security flaws can surface and be patched sooner, mutual oversight across the developer community makes evaluations more credible, and open code provides verifiable evidence for external auditing.[234]

At the same time, Chinese expert publications have continued to identify a range of risks specific to open source models. For example, researchers from Alibaba Research Institute have noted that the irreversibility of open sourcing, combined with the dynamics of AI diffusion, can increase the likelihood of misuse of AI models.[235] Southeast University scholars have described a related set of security risks: for example, there is a lower threshold for attacks on models whose weights and code are public.[236] Researchers at Tongji University have likewise published on open source risks: their work notes that China currently lacks adequate mechanisms for risk assessment of open source models, tracks the

b. Chinese sources often use the term "open source" (开源) without clearly distinguishing between open-weight models, open source software, and open source communities. For instance, 开源模型 ("open source model") generally refers to models with openly released weights, which would be termed "open-weight" in much of the English-language literature. We follow the sources' usage and use "open source" throughout, noting that the governance debates discussed here primarily concern open-weight foundation models, though some arguments address the open source ecosystem more broadly.

diffusion of domestically developed open source models, and monitors backdoor risks in foreign open source models.[237] In a noteworthy recent paper, researchers from Peking University contributed to an empirical audit of over two million Hugging Face repositories, which found that models' use restrictions are not reliably preserved through fine-tuning and merging.[238] The authors argue that achieving deep supply chain accountability will require provenance-preserving infrastructure.

Among the Chinese publications on open source governance that we have surveyed, many are skeptical toward conventional, ex-ante regulation, preferring that governance be embedded within the open source ecosystem itself. Researchers from Alibaba Research Institute argue that governance should respect the self-organizing nature of the open source ecosystem, and that regulations should emphasize information-sharing, disclosure, and oversight by academic institutions, communities, and the public.[239] Scholars at Southeast University have made a similar case for moving from top-down regulation toward "embedded governance," with the state acting as one co-governance actor among industry associations, foundations, and open source communities rather than as an outside regulator.[240] WANG Zhe (王哲) of Beihang University and Tsinghua University's XUE Lan (薛澜) have used the term "agile governance" in relation to open source, arguing that regulation should not shut the door to entry but enable orderly access. They point to a prevailing preference in the field for an open source strategy that secures safety through broad participation and public oversight.[241] Wang and Xue further argue that, amid great-power competition, open source collaboration has become one of the few areas of cooperation able to transcend political barriers.

Even as they favor lighter-touch, ecosystem-based governance, several of the authors cited above pair it with mechanisms to build capacity for tighter open source regulation. The Tongji University researchers are the most explicit in this regard, proposing the establishment of a permanent Open Source AI Risk Assessment Center (bringing together industry, safety experts, and regulators) to unify risk evaluation processes, methods, and standards, and, as an illustrative example, proposing that models be classified by diffusion risk.[242] The Southeast University scholars similarly propose a four-tier risk classification, supported by a national regulatory platform and an emergency-response system, alongside a liability framework that differentiates responsibilities across developers, deployers, end-users, and open source communities.[243] In October 2025, a group of leading Chinese AI law experts from academia and industry published an article in *Science* outlining China's evolving approach to AI governance. While they described exemptions for open source AI as one of six pillars of Chinese regulation, the authors suggested that a more cautious approach be taken for certain frontier models, considering open-weight models' higher misuse risks.

Taken together, these proposals suggest that the preference for ecosystem-based governance is not unconditional. For example, the Tongji researchers argue that open source models do not yet require urgent regulation, but that this could change quickly should capabilities jump.

4.4 AI's impact on cybersecurity

Chinese experts have continued to characterize AI as a "double-edged sword" for cybersecurity. Much of the discourse has concentrated on how generative AI lowers the barrier to sophisticated attacks while also offering new defensive tools.

The topic has gained prominence across a range of expert venues. In an illustrative piece in September 2025, HAN Honggui (韩红桂), Dean of the School of Computer Science at Beijing University of Technology, argued in the Central Party School's *Study Times* (学习时报) that rule-based defenses cannot keep pace with adaptive AI attacks and that AI models themselves have become a new attack surface.[244] It is also noteworthy that CAICT has published dedicated security assessments focused on cybersecurity implications of coding models (see Domestic Governance chapter).[245] A June 2025 special issue of the China Computer Federation's[c] flagship magazine on "large models empowering cybersecurity" includes contributions spanning deepfake detection, automated penetration testing, intelligent defense system design, and vertical-domain model deployment.[246]

Interest in using AI for vulnerability detection was sharpened by the April 2026 release of Anthropic's Claude Mythos Preview, followed weeks later by OpenAI's GPT-5.5-Cyber. Those frontier models were reported to autonomously discover and exploit previously unknown ("zero-day") cybersecurity vulnerabilities at scale.[247]

Chinese views on these frontier models often highlighted the exclusion of Chinese companies from Anthropic's restricted initial release of Mythos under "Project Glasswing." While Chinese observers generally appeared to accept the general claims about Mythos' capabilities, some commentators also shared doubts about the scale of those claims, arguing that its vulnerability-discovery results had been overstated.[248] Others focused less on capability claims and more on the decision not to publicly release the models. According to one outlet, the decision to withhold Mythos from public release marked the end of an era of freely accessible flagship AI products, and the start of a period in which frontier AI could become a scarce strategic resource, with political and economic consequences.[249] In any case, the advent of these types of models was accepted as a paradigm shift across most of the commentaries we surveyed. Writing in response to Mythos, YANG Yanqing (杨燕青), director of a research center at ShanghaiTech University, argued that frontier capabilities at that level could push technological risk to an existential scale, and highlighted the need for cooperation between the world's major technological powers, led by China and the United States.[250]

Experts at prominent Chinese cybersecurity firms also weighed in on the implications of the release of these types of frontier models. ZHOU Hongyi (周鸿祎), founder of the Chinese cybersecurity firm 360, argued that Mythos-class capabilities change "the rules of the game," since an attacker wielding such a model could uncover far more vulnerabilities in a target's code than before. He drew particular

c. The China Computer Federation is a professional organization, which according to its website has 128,000 paid members, and is run independently with no government funding.

attention to Anthropic's Project Glasswing which, he argued, would leave excluded countries in a position of "one-way transparency," exposed to large-scale vulnerability discovery while unable to reciprocate.

A white paper by the security firm QAX warned that comparable models could proliferate within six to eighteen months, making attack tools cheaper and harder to regulate.[251] According to the paper, Chinese government and enterprise clients commonly reacted to Mythos by asking whether it would be possible to build a Mythos-like system to find and patch their own vulnerabilities first. The white paper stated that such a path would not work, citing a persistent capability gap between domestic models and top-tier frontier systems like Mythos.

These perspectives suggest that the offense–defense balance and the democratization of attack capabilities remain central concerns for Chinese experts. What has started to change is that these discussions now have a concrete point of reference. The arrival of frontier models with demonstrated offensive capability has moved the conversation out of the abstract and given it a sharper strategic framing.

4.5 AI's risks for biosecurity

Several Chinese experts have published on the biosecurity risks of advanced AI, often framing it as a global concern and highlighting the importance of international cooperation.

Tianjin University is an important anchor in Chinese biosecurity research.[d] The Tianjin University Center for Biosafety Research and Strategy has examined how AI tools could reshape biosecurity risks. In July 2025, Concordia AI and the Center co-authored a report on AI in the life sciences, which assesses risks from foundation models, biological design tools, and automated scientific platforms, and sets out mitigation and governance options for a range of stakeholders.[252] Speaking at the 2025 World Artificial Intelligence Conference, Center Director ZHANG Weiwen (张卫文) noted that accelerating development of AI may give rise to entirely new and unknown biological knowledge, bringing more complex risks.[e] In August 2025, YUAN Yingjin (元英进), Director of the National Key Laboratory of Synthetic Biotechnology, Dean of Tianjin University's School of Synthetic Biology and Biomanufacturing, and an academician of the Chinese Academy of Sciences, noted that safeguarding China's biosecurity requires enhancing proactive technology governance and establishing new risk early-warning mechanisms for AI-converged biotechnology.[254]

In April 2026, the Chinese Academy of Sciences published "Artificial Intelligence for Science: The Disciplinary System of Artificial Intelligence," a flagship volume co-authored by nearly 300 cross-

d. For example, the Center played a leading role in establishing in the Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists. See also the 2025 *State of AI Safety in China* report, p. 49.

e. Concordia AI is the publisher of this report. The Responsible Innovation in AI × Life Sciences report was co-authored by Concordia AI and the Tianjin University Center for Biosafety Research and Strategy. Zhang Weiwen's remarks cited here were delivered at Concordia AI's AI Safety and Governance Forum at the 2025 World Artificial Intelligence Conference.[253]

disciplinary academicians and experts. Its chapter on AI in the life sciences includes a dedicated section on bioethics and biosecurity, which argues that the deepening integration of AI — including large language models and biological design tools — into life-science research is reshaping how biological risks are generated and spread, lowering technical barriers and amplifying dual-use risks.[255]

Scholars continue to explore how China should assess such risks. For example, researchers from the Chinese Academy of Science and Technology for Development (CASTED) and the Shandong Provincial Center for Food and Drug Evaluation and Inspection recently published an article in which they identified three types of biological AI tools and assessed the risk of each: large language models (LLMs), biological design tools (BDTs), and closed-loop autonomous laboratories.[256] The authors classified BDT risks as the most severe and urgent, capable of a "disruptive shock" to existing biosecurity controls. They described LLM risks as comparatively controllable and "incremental rather than disruptive," while noting that closed-loop autonomous laboratories warrant attention ahead of time. Noting that China is at an "early stage" in the intersecting field of AI and biology, the authors proposed establishing a "Dual-Use Technology Risk Assessment Expert Committee" focused on relevant strategic research, improving biosecurity risk-assessment frameworks related to AI, and developing regulatory options for risk mitigation. Other scholars have examined the issue from different angles. A researcher from Fujian Normal University argued in a November 2025 article that artificial intelligence offers precision tools for biosecurity protection, surveillance, and response.[257] The author called for reasonable regulation of biotechnology, and for building an active defense system characterized by risk early warning, rapid response, and international coordination.

A range of Chinese experts have also prioritized international cooperation to address the risks at the intersection of AI and biosecurity. XIAO Xi (肖晞), Dean of the School of International and Public Affairs at Jilin University, has argued that current international AI governance documents consist mainly of non-binding principles, with "no specialized binding treaty," and a lack of legally binding verification mechanism for the Biological Weapons Convention.[258] The authors from CASTED and Shandong Provincial Center for Food and Drug Evaluation and Inspection cited above have similarly recommended strengthening international exchange and cooperation and regularly conducting technical and policy dialogues — such as setting up a "China–US AI Biosecurity Dialogue Mechanism."

Some leading Chinese experts have also played important roles in international dialogues and exchanges that have produced meaningful outputs on biosecurity risks associated with AI. For example, Andrew Yao, ZENG Yi (曾毅), Zhang Weiwen and others, signed a joint international statement in July 2025 that warned that AI could lower barriers and increase the level of harm malicious actors could cause with biology, as well as reduce the effectiveness of biosecurity and biodefense measures.[259] In December 2025, several Chinese organizations joined an international group that endorsed a series of recommendations to governments on mitigating risks at the intersection of AI and biosecurity.[260] The recommendations offer mitigations across four areas: awareness raising, training and human capacity building; safety evaluations, testing, and industry best practices; national governance; as well as international cooperation. In April 2026, several leading Chinese scholars signed

the 2026 IDAIS London statement, which warned that frontier AI models have demonstrated significantly enhanced capabilities in tasks such as pathogen design, highlighting biological misuse risks (see section 2.4).[261]

Industry Governance

As discussed in the Domestic Governance chapter, Chinese AI companies are subject to a range of binding regulations. This chapter discusses what industry does in addition to those, on a voluntary basis. It first reviews collective industry action spearheaded by the AI Industry Alliance of China (AIIA) before delving into an analysis of individual companies' safety practices and disclosures.

Key updates from July 2025 to June 2026

- **A leading AI industry alliance created new voluntary safety commitments focused on agents** though published implementation disclosures remain brief and generic**.** After revising its general safety commitments in July 2025, the AI Industry Alliance of China (AIIA) led commitments for agentic consumer devices in February 2026 and a "self-discipline convention" for cloud-based agents in April 2026. These build on increasingly rigorous AI agent security proposals and practices from Chinese AI and cloud companies.
- **Voluntary safety testing has expanded to frontier risks for the first time.** The testing program run by the AIIA and the China Academy of Information and Communications Technology (CAICT) released cybersecurity misuse risk assessments of 15 open source coding models in November 2025 and launched AI Safety Benchmark 2.0 in February 2026, with an increased scope covering model deception, loss of control, dangerous-domain misuse, and agentic risks.
- **Company safety disclosure for new model releases remains limited and inconsistent.** Only five of ten leading foundation-model developers reported safety evaluation results alongside any model release this past year, and none did so consistently. DeepSeek-R1's peer-reviewed *Nature* paper and Moonshot AI's Kimi K2 model card were the most detailed safety disclosures to date, yet flagship follow-ups like DeepSeek-V4 and Kimi K2.5 included no safety evaluation results at all.
- **Industry technical safety research output has almost doubled since early 2025 but remains highly concentrated in big tech.** Just six of them (Alibaba, Ant Group, Huawei, Tencent, China Telecom, and ByteDance) account for over half of all AI safety papers by industry since January 2025. Meanwhile, pure-play LLM startups (including DeepSeek, Moonshot AI, and MiniMax) publish almost no technical safety research papers. A small but growing segment of dedicated AI safety startups offering "safety as a service" is also emerging, though publishing volumes remain modest.

5.1 Industry-wide collective 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 initiatives outside of China, such as the Frontier Model Forum, focus narrowly on general-purpose AI model developers, AIIA encompasses over 1,000 member companies, which develop a range of AI products, and thus takes a broader approach to AI industry governance.[262]

Since the establishment of a dedicated Safety and Security Governance Committee in September 2023,[263] AIIA and CAICT have initiated a range of AI safety initiatives, from industry standards and safety testing to company safety commitments.[a]

5.1.1 Voluntary AI safety testing now focuses more on frontier risks

CAICT's voluntary self-testing program substantially expanded in scope over the past year. Started in 2024, this program tests Chinese models on CAICT-developed benchmarks.[264] CAICT typically publishes anonymized evaluation results periodically, but participation by companies is voluntary and results do not carry direct legal effects. In early 2025, CAICT's evaluations initially appeared to be narrowing, focusing only on hallucinations rather than the broader framework developed in 2024.[265]

This trend has since reversed: the evaluations now cover a wide range of frontier risks. In February 2026, the AI Safety Benchmark 2.0 added a range of frontier safety concerns as entirely new categories (self-awareness, model deception, dangerous domain misuse, loss of control) and risks from agentic AI (summarized in Figure 5.1).[266] CAICT has not yet provided detailed descriptions of how they define or operationalize evaluations for concepts like "self-awareness" or "loss of control," leaving some uncertainty about the specific scope of the evaluations.

The most concrete frontier risks results published so far concern cybersecurity. In November 2025, CAICT released security assessments of 15 open source coding models, warning that generative coding models could enable autonomous, unmanned, and large-scale cyberattacks.[267]

It remains unclear when the full range of newly introduced frontier and scenario-based risks will be systematically evaluated and the results published. CAICT's first Q1 2026 release of evaluation results focused on AI agents running directly on consumer devices. They found that while models rarely generate overtly harmful content, many still fail to reliably decline when users instruct them to carry out unsafe or malicious tasks.[268]

a. For an overview of AIIA's past work, refer to our *State of AI Safety in China (2025)* report, pp. 54–55.

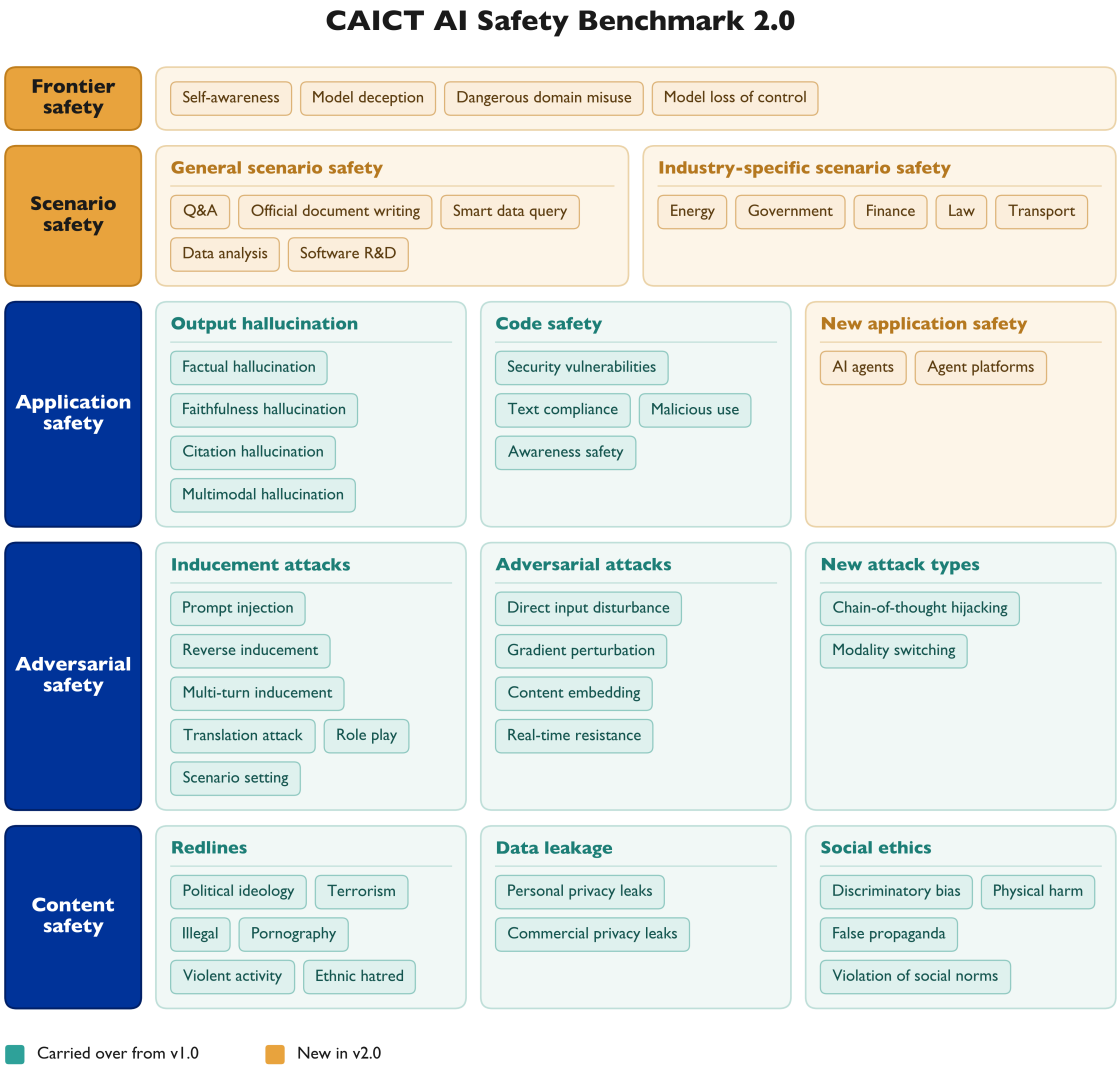


Figure 5.1: Overview of CAICT AI Safety Benchmark 2.0. Translated by Concordia AI, original source: CAICT.[269]

5.1.2 AI safety commitments

Beyond evaluation, AIIA's main lever has been voluntary commitments. Their voluntary safety commitments were first developed in December 2024, and a revised version in July 2025 modestly expanded their scope: Commitment 6 was broadened from "advancing frontier AI safety research" to also cover "preventing safety and security risks in frontier fields" and assessing "risks related to the abuse of AI systems in high-risk scenarios," though it does not define what those risks are.[270] A new commitment on international engagement was added, and five new companies signed on. Implementation disclosures have been published but remain brief and generic, offering no insight into individual

companies' practices, and compliance is ultimately voluntary.[271] The May 2024 "Frontier AI Safety Commitments" from the AI Seoul Summit are a helpful reminder of the inherent limitations of voluntary AI safety commitments: signatories from China and elsewhere pledged to publish a "safety framework focused on severe risks," yet several have not honored that commitment.[272]

More recently, the focus has shifted to AI agents. In February 2026, AIIA led agent-specific commitments.[273] Signed by major phone manufacturers, cloud providers, and security firms, these focus on security best practices and the protection of user rights in mainstream agentic consumer products. Then in April 2026, AIIA and the China Institute of Communications jointly issued a "self-discipline convention" for cloud-based AI agent services — those that let users directly deploy and run OpenClaw-like agents in the cloud.[274] Signatories commit to standard security practices alongside more specific obligations: screening third-party plugins for malicious content, monitoring for prompt injection attacks, and reporting security incidents to authorities. The convention was signed by eleven major Chinese cloud companies.

These commitments are part of a wider industry effort to tackle agent risks. CAICT and Tencent Cloud have published joint AI Agent Security Practice Guidelines, and the AIIA issued a Risk Management Guide for OpenClaw-like agents recommending strict scoping of agent operations and periodic permission review.[275] These build on a growing set of agent-security solutions from individual Chinese AI developers and cloud companies. Alibaba Cloud offers an end-to-end agent identity service that binds credentials to trusted runtimes, and the company's Agent Operation Authorization framework could help enforce fine-grained, least-privilege delegation.[276] ByteDance's cloud service Volcengine runs a dedicated platform monitoring agent inputs and outputs in real time.[277] Huawei's HarmonyOS agent security policy bars agents at the operating-system level from disabling controls like "back" keys or process termination that would let them trap users in the agentic environment.[278] Tencent Cloud isolates agent sessions through its open source Cube Sandbox and has open sourced its AI-Infra-Guard, a platform that scans agents, MCP servers, agent skills, and AI infrastructure for vulnerabilities.[279]

Overall, AIIA continues to update and expand the safety commitments and monitor their uptake. The most concrete practices have emerged in agent security.

5.2 Company-specific safety practices

Over the past year, Chinese AI companies have discussed concerns about AI safety risks — including frontier risks — in their public communications. One example is the January 2026 Hong Kong IPO prospectuses of Z.AI (formerly Zhipu) and MiniMax, which both devoted substantial space to safety.[280] Z.AI warned that AI could cause "grave harm, even catastrophe," and stressed the need to "approach AI with extreme care," while MiniMax cautioned that advanced models could "pursue goals misaligned with user or societal interests" and might "develop emergent capabilities, such as strategic planning or deception." Alibaba's Qwen tech lead LIN Junyang (林俊旸) said in a January 2026

panel that his greatest worry is no longer that a model "says something it shouldn't," but that it "does something it shouldn't."[281] Alibaba security executive XUE Hui (薛晖) has also flagged AI's shift from "thinking" to "acting" and argued that safety evaluations should include novel risks like self-replication and loss of control.[282] An article published in *Science* by a group of Chinese experts, including from companies like Alibaba and Deepseek, warned of "extreme risks" from open source model misuse, urging developers to be "more transparent and evidence-based" about frontier safety.[283]

While such public statements imply that some leading AI firms are concerned about risk, they should also be read with caution, as it is unclear how far they reflect genuine intent. For example, although Z.AI prominently cites its signing of the Seoul Frontier AI Safety Commitments in its IPO documents, it has not followed through on the requirement to publish a frontier AI safety policy.

5.2.1 Safety disclosure for new model releases remains limited

This section analyzes the safety practices disclosed alongside new model releases, typically in "model cards" or "technical reports" published on arXiv, HuggingFace, or the companies' websites. We focus on companies that lead in general-purpose foundation models. Table 5.1 provides a full overview of safety disclosures accompanying new model releases by ten leading companies between June 2025 and May 2026.[b]

Overall, safety content in these model cards remains limited. Of the ten companies, only five (Bytedance, DeepSeek, Meituan, Moonshot AI, and Z.AI) provided safety evaluation results alongside any model release during the year in question, and even those five did not do so consistently across releases. This represents only modest improvement over a similar analysis in last year's report, in which three of 13 companies surveyed reported safety evaluation results.

The most detailed safety disclosure for a Chinese model comes from DeepSeek-R1 in *Nature*, marking the first time a widely-used LLM has undergone such independent peer review.[286] The supplementary material includes a 10-page safety section, the most thorough public release of its kind from a Chinese developer to date.[287] It describes the "risk control system" used across DeepSeek's services: an additional safety layer on top of the model, in which one model judges whether the other's output complies with a set of safety principles. The paper transparently reveals the complete risk review prompt used. The prompt includes clauses prohibiting answers related to cyberattacks and manufacturing dangerous weapons, including controlled biochemicals. The paper also reported the model's results on six major safety benchmarks (Simple Safety Tests, BBQ, Anthropic Red Team, XSTest, DNA, and HarmBench), alongside an internal evaluation framework focused on compliance with Chinese regulations.

Another relatively detailed example is Moonshot AI's Kimi K2. Moonshot AI worked with third-party tester Promptfoo on a wide range of safety evaluations spanning harmful content (e.g. hate speech,

b. Specifically, we choose to analyze model cards by the top 10 Chinese companies on the LMArena LLM leaderboard.[284] These models generally also score highly on other widely used benchmarks like Artificial Analysis.[285]

explicit material), criminal behavior (e.g. chemical and biological weapon development, IP theft), misinformation, privacy, and security (e.g. malicious code generation).

Even in these cases, however, companies did not report comparably detailed disclosures for every new model release: the most recent releases from those companies (DeepSeek-V4 and Kimi K2.5) included no safety evaluation results at all.[288] Overall, Chinese model releases still feature substantially less consistent and less detailed safety disclosures compared to their US peers.

Table 5.1: Overview of safety disclosure in model cards of ten major Chinese foundation developers.

Company	Model	Dedicated safety section?	Safety evaluations
Alibaba	Qwen3-Max (Sep 2025)[289]	No	No
	Qwen3.5 (Feb 2026)[290]	No	No
Baidu	ERNIE 4.5 (Jun 2025)[291]	No. Safety and harmlessness mentioned very briefly as one category for supervised fine-tuning.	No
	ERNIE 5.0 (Feb 2026)[292]	No. Briefly mentions safety filtering of training data.	No
Bytedance	Seed 1.8 (Dec 2025)[293]	Very brief section; focuses on safety evaluation results.	AIR-Bench; XSTest; an in-house safety benchmark focused on "civil norms, pornography, illegal acts, copyright, medical safety, identity."
	Seed 2.0 (Feb 2026)[294]	No	No
	Seed-OSS-36B (Aug 2025)[295]	No	AIR-Bench
DeepSeek	DeepSeek-R1 (*Nature*) (Sep 2025)[296]	10-page "Safety Report" in the Appendix, including full disclosure of the "The Risk Review Prompt for DeepSeek-R1."	Simple Safety Tests; Bias Benchmark for QA; Anthropic Red Team; XSTest; Do-Not-Answer; HarmBench; an in-house benchmark covering "discrimination and prejudice, illegal and criminal behavior, harmful behavior, and moral and ethical issues."

Continued on next page

Table 5.1: Overview of safety disclosure in model cards of ten major Chinese foundation developers. *(continued)*

Company	Model	Dedicated safety section?	Safety evaluations
	DeepSeek-V3.2 (Dec 2025)[297]	No. Briefly mentions "human alignment training" during the RL stage.	No
	DeepSeek-V4 (Apr 2026)[298]	No. The only safety-adjacent mention is operational sandbox isolation/security for running agentic tasks.	No
Meituan	LongCat-Flash (Sep 2025)[299]	No full section, but describes safety practices in considerable detail throughout, including a "safety policy that categorizes queries into more than 40 distinct safety categories."	An in-house benchmark spanning harmful content (violence, hate speech, insulting, harassment and bullying, self-harm and suicide, adult content), criminal content (illegal activities, underage violations, extreme terrorism and violence), misinformation and disinformation, and privacy.
	LongCat-Flash-Thinking (Sep 2025)[300]	No full section, but briefly mentions safety practices in multiple places.	Same in-house safety benchmark as used in LongCat-Flash.
MiniMax	MiniMax-M1 (Jun 2025)[301]	No	No
	MiniMax-M2 (Oct 2025)[302]	No	No
	MiniMax-M2.5 (Feb 2026)[303]	No	No
	MiniMax-M2.7 (Apr 2026)[304]	No	No
Moonshot AI	Kimi K2 (Jul 2025)[305]	No. Briefly describes seed prompts aimed at reducing "violence, fraud, and discrimination," with an automated prompt evolution pipeline to simulate jailbreaking attempts.	Safety evaluation performed with *Promptfoo*, covering: harmful, criminal, misinformation, privacy, security (each with multiple sub-categories).
	Kimi K2.5 (Feb 2026)[306]	No	No

Continued on next page

Table 5.1: Overview of safety disclosure in model cards of ten major Chinese foundation developers. *(continued)*

Company	Model	Dedicated safety section?	Safety evaluations
	Kimi K2.6 (Apr 2026)[307]	No. Briefly mentions that the model has "enhanced safety awareness during extended research tasks."	No
Tencent	Hunyuan-A13B (Aug 2025)[308]	No. Very brief mention in the RL stage: "Safe response pairs are identified using classifiers and refusal heuristics, integrating safety alignment directly into preference datasets to mitigate risks."	No
	Hy3-preview (Apr 2026)[309]	No	No
Xiaomi	MiMo-V2-Flash (Jan 2026)[310]	No. Mentions "safety alignment" as one post-training category, and safety rewards in RL.	No
	MiMo-V2.5-Pro (Apr 2026)[311]	No. Very briefly mentions "safety" as a category in RL.	No
Z.AI/Zhipu	GLM-4.5 (Aug 2025)[312]	Very briefly discusses safety as one dimension in RL.	SafetyBench (includes ethics and morality, illegal activities, mental health, offensiveness, physical health, privacy and property, and unfairness and bias).
	GLM-5 (Feb 2026)[313]	No	No

5.2.2 Technical safety contributions

The analysis in this section is based on Concordia AI's Chinese Technical AI Safety Database, which has collected over 900 technical papers published by Chinese institutions on arXiv from April 2023 through April 2026.[c] Refer to section 3.1 for a detailed description of the methodology behind this database. This section focuses exclusively on Chinese industry contributions.

c. The full database is accessible on the report companion website: https://aisafetychina.com/research/.

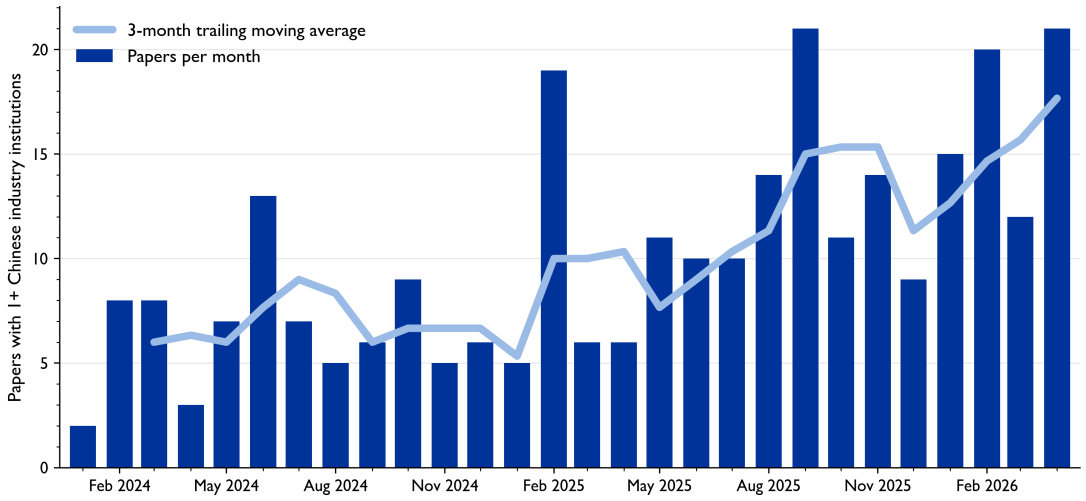


Figure 5.2: Papers per month with Chinese industry participation

In absolute terms, technical research output with industry participation has grown rapidly: the 3-month trailing average rose from roughly six Chinese-industry-involved papers per month in early 2024 to around ten in early 2025 and then roughly 18 per month in early 2026 — a threefold increase in two years.[d] However, this growth largely mirrors the overall expansion of Chinese AI safety research: as a share of total output, Chinese industry's contribution has been remarkably stable since early 2024, oscillating between approximately 25% and 40% with no clear trend. As documented in section 3.3, the majority of key research groups focused on AI safety are affiliated with universities and state-backed labs, not industry.

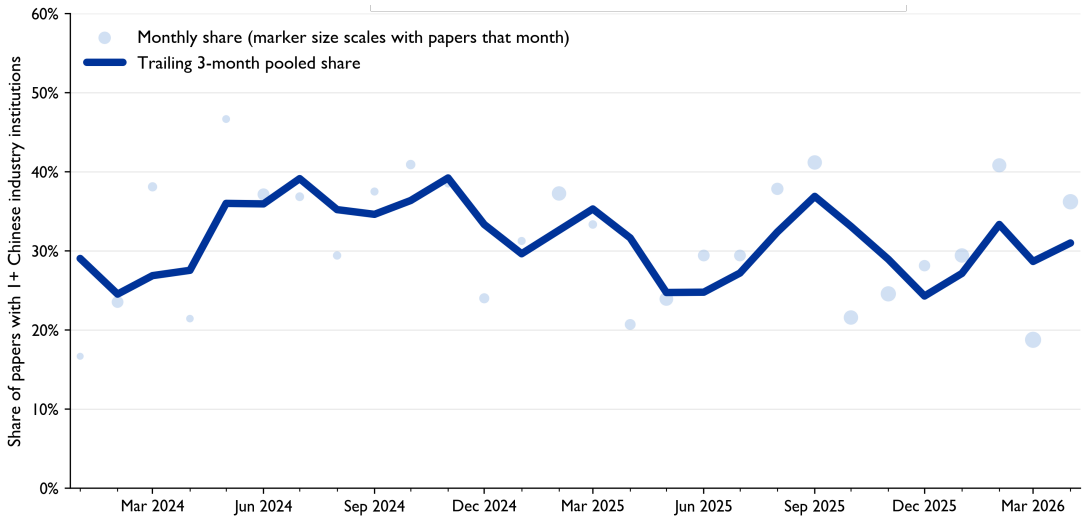


Figure 5.3: Share of papers with Chinese industry participation

d. This includes all papers with at least one contributor from a Chinese company. Many of these papers involve research collaborations with universities or state-backed labs, and are not exclusively conducted or led by companies.

Within industry, output is heavily concentrated in big tech. From January 2025 to April 2026, six firms account for over half of all Chinese-industry papers: Alibaba (29 papers), Ant Group (22), Huawei (17), Tencent (14), China Telecom (13), and ByteDance (12).

It is not possible to give a complete overview of these technical research contributions here, and we encourage interested readers to explore the full database on our companion website.[314] One notable contribution from big tech comes from Alibaba's Qwen team, which in September 2025 released Qwen3Guard, a family of open source multilingual guardrail models for moderating LLM inputs and outputs.[315] This tool helps developers improve safety across nine categories: violent content, non-violent illegal acts, sexual content, personally identifiable information, suicide and self-harm, unethical acts (bias, harassment, misinformation), politically sensitive topics, copyright violation, and — for inputs only — jailbreak attempts. The tool has seen substantial open source uptake, with around 4 million downloads on Hugging Face as of June 2026.[316] Notably, however, this investment in a dedicated guardrail product did not translate into detailed safety reporting in Qwen's own foundation model releases during the period surveyed (see Table 5.1).

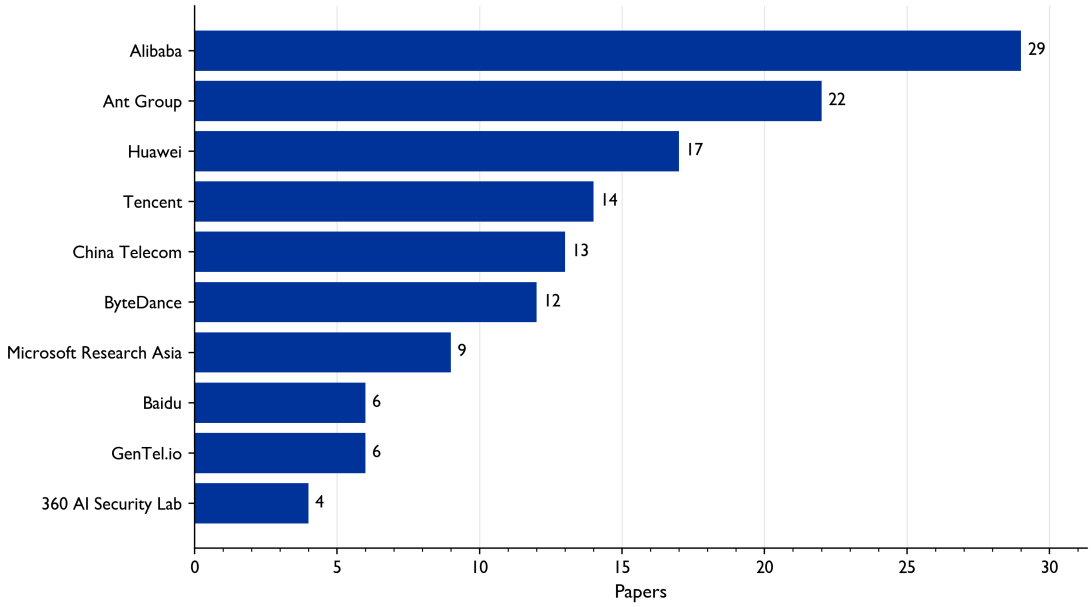


Figure 5.4: Top 10 Chinese industry contributors to AI safety papers, Jan 2025 – Apr 2026

By contrast, the pure-play LLM startups are nearly absent: StepFun has three papers (on reward hacking in RLHF, attacks against reasoning models, and scalable interactive oversight),[317] Z.AI two papers (on long-context safety evaluation and multimodal jailbreaks),[318] and 01.AI two papers (on reliability of reward models after instruction tuning, and automated red-teaming for LLM-based search agents).[319] DeepSeek, MoonShot, and MiniMax have no safety papers in our dataset since January 2025 at all. This may, in part, reflect stronger resource constraints: smaller labs cannot afford dedicated teams for publishing research results.

A smaller number of papers come from dedicated AI safety start-ups aiming for "safety as a service" business models. Among these, the most active are:

- GenTel.io (君同未来) with six papers from January 2025 to April 2026: Founded by Zhejiang University's HAN Meng (韩蒙), the company offers a wide-ranging AI safety and security testing portfolio. While it is not entirely clear which risks they prioritize, their services appear primarily geared towards compliance with Chinese regulations.[320] Their recent co-authored papers in our database cover jailbreak attacks (via neuron relearning, latent-space representation blending, and inference-time guardrail ablation), safety benchmarks, and emerging attack surfaces in LLM-driven multi-agent systems.[321]
- RealAI (瑞莱智慧) with four papers: Incubated at Tsinghua University in 2018, RealAI was one of the earliest companies in this space. The company has expressed explicit concern about frontier risks: in May 2026 CEO TIAN Tian (田天) warned that AI systems already show early signs of high-risk behavior, including modifying or disabling their own shutdown mechanisms and answering dangerous cyberattack or biochemical questions.[322] In February 2026, RealAI's "RealGuard" product — related to security of facial recognition systems, not general-purpose AI — was showcased to President Xi Jinping during his inspection of the National Information Technology Application Innovation Park in Beijing.[323] Their four papers since January 2025 cover jailbreak defense via progressive answer detoxification, safety alignment of reasoning models (including RealSafe-R1, a safety-aligned variant of DeepSeek-R1), and multimodal trustworthiness evaluations.[324]
- BraneMatrix AI (布兰矩阵) with three papers: This Shanghai-based startup is a more recent entrant. Their product centers on the safety-evaluation platform "Orbital Supervision." It draws on a range of proprietary red-teaming tools and on the OmniSafeBench-MM benchmark, which BraneMatrix AI co-developed with eight other institutions, covering 9 risk domains with 50 fine-grained subcategories.[325] Their other two papers focus on automated stealthy prompt injection against LLM-driven coding agents, and bio-inspired adversarial prompt optimization to bypass safety guardrails.[326]

Conclusion

The period covered by this report may come to be seen as the moment that China's AI governance most clearly moved beyond content control. For years, the center of gravity of China's binding AI rules lay in regulating what models *say*. Over the past year, that center has shifted toward a much broader set of risks: most visibly, toward what AI agents *do*. The OpenClaw episode compressed into a few weeks a dynamic that might otherwise have taken years: a capability diffused widely, incidents followed, and regulators, standards bodies, researchers, and industry all redoubled efforts to tackle agentic AI risks. Regulators issued dedicated guidance on agentic AI, and standard-setting bodies began drafting agent security standards. Chinese frontier AI safety research grew substantially, with agent safety rising from 8% of new papers in early 2025 to more than a quarter in early 2026. Agentic AI safety also emerged as a distinct strand of expert discourse, with prominent scholars warning that agents could erode human oversight or engage in recursive self-improvement.

Meanwhile, new rules on anthropomorphic AI services tackle emotional dependency, manipulation, and risks to vulnerable groups like minors and the elderly; the Five-Year Plan and other policies highlight AI's labor impact; and CBRN misuse and loss of control now feature in national standards documents. Regulators in Brussels, Washington D.C., and elsewhere are grappling with similar problems, suggesting room for mutual learning.

Several developments deserve particular attention over the coming year. First, we will be interested to see the trajectory of the mandatory national standard on agent application security, which would become only the second mandatory AI standard in China and the clearest signal yet of how far Beijing intends to regulate autonomous action. Second, the significance of the newly agreed China–US intergovernmental dialogue on AI will depend on details that remain undefined at the time of writing: the scope of the agenda, the level and composition of participation, and the cadence of meetings. Those parameters will be the first indication of how much both sides intend to invest in the channel. Third, we will closely track whether Chinese frontier AI developers improve their safety reporting practices. Chinese open source models are, by some measures, now the most downloaded in the world. Hence, safety practices in the Chinese industry have implications for developers worldwide. In the coming year, we will be monitoring whether detailed safety reporting alongside new model releases develops into standard practice. Such a shift would also help build mutual understanding between Chinese AI developers and their international counterparts.

On the international front, China enters the second half of 2026 with an unusually full diplomatic calendar: the first Global Dialogue on AI Governance in Geneva, the inaugural report of the International Scientific Panel, a China-hosted APEC summit in Shenzhen with AI on the agenda, and a newly agreed US–China intergovernmental dialogue. Beijing has invested heavily in positioning itself as an architect of multilateral AI governance — most visibly through its proposed World AI Cooperation Organization — at precisely the moment the United States has grown skeptical of it. The outcomes of these meetings will determine whether 2026 marks a genuine opening or another false start for international governance of advanced AI.

Regardless of what is achieved on the international stage in the coming year, the more consequential story documented in this report may be the expansion of domestic AI governance priorities already underway within China. How that shift unfolds will be a defining issue for global AI safety for years to come.

Notes

1 Concordia AI, *Frontier AI Risk Monitoring Platform*, https://airiskmonitor.net/, 2026, accessed June 19, 2026

2 Anthropic, *Project Glasswing: Securing Critical Software for the AI Era*, https://www.anthropic.com/glasswing, April 2026, accessed June 19, 2026

3 Avijit Ghosh et al., *State of Open Source on Hugging Face: Spring 2026*, https://huggingface.co/blog/huggingface/state-of-os-hf-spring-2026, Hugging Face, April 2026, accessed June 23, 2026

4 Yoshua Bengio et al., *International AI Safety Report 2026*, 2026, accessed June 26, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

5 *Outline of the Fifteenth Five-Year Plan for National Economic and Social Development of the People's Republic of China (中华人民共和国国民经济和社会发展第十五个五年规划纲要)*, https://www.news.cn/politics/20260313/085af5de5a4b4268aa7d87d90817df2f/c.html, Government of the People's Republic of China, March 2026, accessed June 19, 2026

6 State Council, *Opinions of the State Council on Deeply Implementing the "AI Plus" Action (国务院关于深入实施"人工智能+"行动的意见)*, https://www.gov.cn/zhengce/content/202508/content_7037861.htm, Government of the People's Republic of China, August 2025, accessed June 19, 2026

7 *The Politburo of the CPC Central Committee Holds a Meeting to Analyze and Study the Current Economic Situation and Economic Work; CPC Central Committee General Secretary Xi Jinping Presides Over the Meeting (中共中央政治局召开会议分析研究当前经济形势和经济工作中共中央总书记习近平主持会议)*, https://www.news.cn/politics/leaders/2023-04/28/c_1129576764.htm, Xinhua, April 2023, accessed June 19, 2026

8 *At the 20th Collective Study Session of the Political Bureau of the CPC Central Committee, Xi Jinping Stressed Upholding Self-Reliance and Self-Improvement, Emphasizing an Application Orientation, and Promoting the Healthy and 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)*, https://www.gov.cn/yaowen/liebiao/202504/content_7021072.htm, Government of the People's Republic of China, April 2025, accessed June 19, 2026

9 *Xi Jinping Delivers Important Speech at the Opening of the Seminar for Principal Provincial and Ministerial Leading Cadres on Studying and Implementing the Spirit of the Fourth Plenary Session of the 20th CPC Central Committee, Emphasizing Deeply Studying and Implementing the Spirit of the Fourth Plenary Session of the 20th CPC Central Committee and Striving for a Good Start to the "15th Five-Year Plan" (习近平在省部级主要领导干部学习贯彻党的二十届四中全会精神专题研讨班开班式上发表重要讲话强调深入学习贯彻党的二十届四中全会精神努力实现"十五五"良好开局)*, https://tv.cctv.com/2026/01/20/VIDEx0VMRP7T9t8V3w6PF6JF260120.shtml, CCTV, January 2026, accessed June 19, 2026; *Sidelights on the Special Seminar for Principal Leading Officials at the Provincial and Ministerial Level (省部级主要领导干部专题研讨班侧记)*, https://www.news.cn/20260124/3f1f3cead780463b9f8119285fe6fb4f/c.html, Xinhua, January 2026, accessed July 1, 2026

10 *Xi Jinping: Forward-Looking Planning and Development of Future Industries (习近平：前瞻布局和发展未来产业)*, https://www.gov.cn/yaowen/liebiao/202605/content_7070720.htm, Government of the People's Republic of China, May 2026, accessed June 19, 2026

11 *Li Qiang's Address at the Opening Ceremony of the 17th Summer Davos Forum (李强在第十七届夏季达沃斯论坛开幕式上的致辞)*, https://www.news.cn/politics/leaders/20260624/ab61dcdc2af64460812c7b7f355e24bd/c.html, Xinhua, June 2026, accessed June 30, 2026

12 *Early Warning Notice on Guarding Against Security Risks of the OpenClaw Open-Source AI Agent (关于防范 OpenClaw 开源 AI 智能体安全风险的预警提示)*, https://nvdb.org.cn/publicAnnouncement/2019330237532790786, Ministry of Industry and Information Technology (工业和信息化部), February 2026, accessed June 19, 2026; *Risk Advisory on the Secure Use of OpenClaw (关于 OpenClaw 安全应用的风险提示)*, https://www.cert.org.cn/publish/main/11/2026/20260312144519429724511/20260312144519429724511_{}.html, National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC) (国家互联网应急中心), March 2026, accessed June 19, 2026; Ministry of State Security (国家安全部), *"Lobster" (OpenClaw) Safe Farming Handbook ("龙虾"（OpenClaw）安全养殖手册)*, https://mp.weixin.qq.com/s/VOSy-kWs6zuNIBn40dWGWQ, Weixin Official Accounts Platform, March 2026, accessed June 19, 2026

13 Gabriel Wagner and Erik Lindblad, *AI Safety in China #26*, 2026, accessed June 26, 2026, https://aisafetychina.substack.com/p/ai-safety-in-china-26

14 The National Internet Finance Association of China (NIFA) (中国互联网金融协会), *Risk Warning on the Security of OpenClaw Applications in the Internet Finance Industry (关于 OpenClaw 在互联网金融行业应用安全的风险提示)*, https://mp.weixin.qq.com/s/phJ12YtXVy4L80JQS5weXQ, Weixin Official Accounts Platform, March 2026, accessed June 19, 2026

15 *Some Banks Receive "Lobster" Risk Advisory from Regulators, Urging Timely Updates and Patching of Security Vulnerabilities (有银行收到监管机构"龙虾"风险提示，提示及时更新、封堵安全漏洞)*, https://finance.sina.com.cn/jjxw/2026-03-12/doc-inhqtqkp4791854.shtml, Sina (新浪财经), March 2026, accessed June 19, 2026

16 Cyberspace Administration of China, National Development and Reform Commission, and Ministry of Industry and Information Technology, *Implementation Opinions on the Standardized Application and Innovative Development of AI Agents (智能体规范应用与创新发展实施意见)*, https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm, May 2026, accessed June 19, 2026

17 Anthropic, *Project Glasswing: Securing Critical Software for the AI Era*, https://www.anthropic.com/glasswing, April 2026, accessed June 19, 2026

18 *Xinhua News Flash: China Will Issue a Document on Responding to the Impact of AI to Promote Employment (新华社快讯：我国将出台应对人工智能影响促就业文件)*, https://www.news.cn/politics/20260127/8af067692da94844adee34b643b39444/c.html, Xinhua, January 2026, accessed June 19, 2026

19 *Notice of the State Council on Issuing the "Plan for Implementing the Employment-First Strategy during the 15th Five-Year Plan Period" (国务院关于印发《实施就业优先战略"十五五"规划》的通知)*, https://www.gov.cn/zhengce/content/202606/content_7072481.htm, Government of the People's Republic of China, June 2026, accessed June 23, 2026

20 State Council, *Notice of the State Council Leading Group for Employment Promotion and Labor Protection on Issuing the Action Plan for Stabilizing Jobs, Expanding Capacity, and Improving Quality (国务院就业促进和劳动保护工作领导小组关于印发《稳岗扩容提质行动方案》的通知)*, https://www.hlj.gov.cn/hljzqc/c100116/202605/c00_31941651.shtml, Government of the People's Republic of China, April 2026, accessed June 19, 2026

21 *Top Ten Typical Cases of Labor and Personnel Dispute Arbitration in Beijing in 2025 (2025 年北京市劳动人事争议仲裁十大典型案例)*, https://rsj.beijing.gov.cn/bm/ztzl/dxal/202512/t20251226_4366546.html, Beijing Municipal Bureau of

Human Resources and Social Security (北京市人力资源和社会保障局), December 2025, accessed June 19, 2026; Zhai Ruimin (翟瑞民), *Judicial Precedents Draw a Red Line on AI Replacement: Don't Let Workers Bear the Risks of the "Technological Revolution" Alone (司法判例为 AI 替代划红线，莫让劳动者独自承担"技术革命"风险)*, 2026, accessed June 26, 2026, https://www.stcn.com/article/detail/3870681.html

22 *State of AI Safety in China (May 2024)*, https://concordia-ai.com/research/state-of-ai-safety-in-china-may-2024/, Concordia AI, July 2024, accessed June 19, 2026; Gabriel Wagner et al., *State of AI Safety in China (2025)*, https://concordia-ai.com/research/state-of-ai-safety-in-china-2025/, Concordia AI, July 2025, accessed June 19, 2026

23 Cyberspace Administration of China et al., *Interim Measures for the Administration of Anthropomorphic Interactive Artificial Intelligence Services (人工智能拟人化互动服务管理暂行办法)*, https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm, April 2026, accessed June 19, 2026; Cyberspace Administration of China, *Notice of the Cyberspace Administration of China on Soliciting Public Comments on the "Interim Measures for the Administration of Anthropomorphic Interactive Artificial Intelligence Services (Draft for Comment)" (国家互联网信息办公室关于《人工智能拟人化互动服务管理暂行办法（征求意见稿）》公开征求意见的通知)*, https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm, December 2025, accessed June 19, 2026

24 Cyberspace Administration of China, *Expert Interpretation | A New Exploration in Regulating the Healthy Development of Frontier AI Business Forms: Interpreting the "Interim Measures for the Administration of Anthropomorphic Interactive AI Services" (专家解读｜规范人工智能前沿业态健康发展的新探索：解读《人工智能拟人化互动服务管理暂行办法》)*, https://www.cac.gov.cn/2025-12/28/c_1768662848000498.htm, December 2025, accessed June 19, 2026

25 Ministry of Industry and Information Technology et al., *Ministry of Industry and Information Technology and Ten Other Departments Jointly Issue the "Measures for the Ethical Review and Service of Artificial Intelligence Science and Technology (Trial)" (工信部等十部门联合印发《人工智能科技伦理审查与服务办法（试行）》)*, https://jxt.hubei.gov.cn/bmdt/rdjj/202604/t20260407_5909143.shtml, April 2026, accessed June 19, 2026

26 Ministry of Science and Technology et al., *Notice on Issuing the "Measures for the Review of Science and Technology Ethics (Trial)" (关于印发《科技伦理审查办法（试行）》的通知)*, https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm, September 2023, accessed June 19, 2026

27 Ministry of Industry and Information Technology, *Notice of the General Office of the Ministry of Industry and Information Technology on Implementing the Pilot Program for the Ethical Review of AI Science and Technology and Services (工业和信息化部办公厅关于实施人工智能科技伦理审查与服务先导计划的通知)*, https://www.miit.gov.cn/jgsj/kjs/wjfb/art/2026/art_1a0b2b0b0deb47099202871e0b70551b.html, April 2026, accessed June 19, 2026

28 Ao Li (敖立), *CAICT's Ao Li: Building a Solid Ethical Foundation, Empowering Innovation and Development——An In-Depth Interpretation of the "Measures for the Ethical Review and Service of AI Science and Technology (Trial)" (中国信通院敖立：筑牢伦理基石，赋能创新发展——深入解读《人工智能科技伦理审查与服务办法（试行）》)*, https://mp.weixin.qq.com/s/ym-Z7TgEVacYk6zUu9IivA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; *Reflections and Actions of CEPREI (the Fifth Electronics Research Institute of MIIT) on the Practice of the "AI Science and Technology Ethics Review and Service Pilot Program" (工信部电子五所关于"人工智能科技伦理审查与服务先导计划"实践的思考与行动)*, https://mp.weixin.qq.com/s/yC9OY3VMyYRK7498BF51dA, Weixin Official Accounts Platform, May 2026, accessed June 19, 2026

29 Zhao Jingwu (赵精武), "The Systematic Construction of the Ethics Review System for AI Science and Technology (人工智能科技伦理审查制度的体系化建构)," *当代法学*, 2025, accessed June 26, 2026, http://www.socio-legal.sjtu.edu.cn/wxzy/info.aspx?itemid=4835&lcid=30

30 Ministry of Industry and Information Technology, *Notice of the General Office of the Ministry of Industry and Information Technology on Implementing the Pilot Program for the Ethical Review of AI Science and Technology and Services (工业和信息化部办公厅关于实施人工智能科技伦理审查与服务先导计划的通知)*, https://www.miit.gov.cn/jgsj/kjs/wjfb/art/2026/art_1a0b2b0b0deb4709920287 1e0b70551b.html, April 2026, accessed June 19, 2026

31 *Decision of the Standing Committee of the National People's Congress on Amending the "Cybersecurity Law of the People's Republic of China" (全国人民代表大会常务委员会关于修改《中华人民共和国网络安全法》的决定)*, http://www.npc.gov.cn/npc/c2/c30834/202510/t20251028_449048.html, National People's Congress, October 2025, accessed June 19, 2026

32 *Public Consultation Announcement on the "Cybercrime Prevention and Control Law (Draft for Comment)" (《网络犯罪防治法(征求意见稿)》公开征求意见的公告)*, https://www.mps.gov.cn/n2254536/n4904355/c10386242/content.html, Ministry of Public Security, January 2026, accessed June 19, 2026

33 State Council, *Notice of the General Office of the State Council on Issuing the "2023 Annual Legislative Work Plan of the State Council" (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)*, https://www.gov.cn/zhengce/content/202306/content_6884925.htm, May 2023, accessed June 19, 2026

34 State Council, *Notice of the General Office of the State Council on Issuing the "State Council 2024 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)*, https://www.moj.gov.cn/pub/sfbgw/gwxw/xwyw/202405/t20240509_498554.html, May 2024, accessed June 19, 2026; State Council, *Notice of the General Office of the State Council on Issuing the "State Council's 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)*, https://www.mee.gov.cn/zcwj/gwywj/202505/t20250516_1119545.shtml, May 2025, accessed June 19, 2026; National People's Congress, *2024 Annual Legislative Work Plan of the Standing Committee of the National People's Congress (全国人大常委会 2024 年度立法工作计划)*, http://www.npc.gov.cn/npc/////c2/c30834/202405/t20240508_436982.html, April 2024, accessed June 19, 2026; National People's Congress, *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)*, http://www.npc.gov.cn/npc/c2/c30834/202505/P020250513550316685290.pdf, April 2025, accessed June 19, 2026

35 State Council, *Notice of the General Office of the State Council on Issuing the "State Council 2026 Annual Legislative Work Plan" (国务院办公厅关于印发《国务院 2026 年度立法工作计划》的通知)*, https://www.moj.gov.cn/pub/sfbgw/zwgkztzl/xxxcgcxjpfzsx/fzsxyw/202605/t20260511_534688.html, May 2026, accessed June 19, 2026

36 National People's Congress, *2026 Legislative Work Plan of the Standing Committee of the National People's Congress (全国人大常委会 2026 年度立法工作计划)*, https://www.moj.gov.cn/pub/sfbgw/gwxw/xwyw/202605/t20260511_534683.html, May 2025, accessed June 19, 2026

37 Kwan Yee Ng et al., *Translation: Artificial Intelligence Law, Model Law v. 1.0 (Expert Suggestion Draft) –Aug. 2023*, 2023, accessed June 26, 2026, https://digichina.stanford.edu/work/translation-artificial-intelligence-law-model-law-v-1-0-expert-suggestion-draft-aug-2023/; Zhou Hui (周辉) et al, *Model Artificial Intelligence Law 3.0 (人工智能示范法 3.0)*, https://mp.weixin.qq.com/s/pCC_AM5mpU7QY-x14R-kZw, Weixin Official Accounts Platform, March 2025, accessed June 19, 2026; AI 善治学术工作组, *AI Good Governance Forum Held in Beijing, Releasing the "AI Law (Scholars' Draft Proposal)" (AI 善治论坛在京召开发布《人工智能法（学者建议稿）》)*, https://mp.weixin.qq.com/s/Ip7LheBPfOoP6ywmloaSIA, Weixin Official Accounts Platform, March 2024, accessed June 19, 2026

38 *Top Ten Topics for AI Rule-of-Law Research in 2026 Released (2026 年人工智能法治研究十大议题发布)*, https://mp.weixin.qq.com/s/LnlNoHXCnUV-AJ2WlUsgyg, Weixin Official Accounts Platform, accessed June 23, 2026; Li Honglei (李洪雷) et al, *Ten Core Revisions of the "Artificial Intelligence Model Law 4.0" and Ten Advanced Topics in the Rule of Law for Artificial Intelligence (《人工智能示范法 4.0》十大核心修改与人工智能法治十大进阶议题)*, https://mp.weixin.qq.com/s/SbfWexiimwyZB7JEisjbTw, Weixin Official Accounts Platform, March 2026, accessed June 19, 2026

39 南京大学法学院, *Nanjing University Law School Team Releases the "Basic Law on Artificial Intelligence" (Expert Recommendation Draft), Constructing a New Path for China's AI Governance (南京大学法学院团队发布《人工智能基本法》（专家建议稿），构建中国人工智能治理新路径)*, https://mp.weixin.qq.com/s/7nr18bX6z6SkO44rO0NLKw, Weixin Official Accounts Platform, accessed June 19, 2026; *Roundtable Forum on the Zhejiang University Scholars' Draft Proposal of the Artificial Intelligence Law Held (《人工智能法》浙大版学者建议稿圆桌论坛召开)*, https://mp.weixin.qq.com/s/aFzRzmNEcYJU4UTUVLgy_w, Weixin Official Accounts Platform, May 2026, accessed June 19, 2026; *Industry Recommended Draft for Promoting Artificial Intelligence Legislation (促进人工智能立法的产业建议稿)*, https://mp.weixin.qq.com/s/luaom6grkd-PYPeBKtYi7w, Weixin Official Accounts Platform, May 2026, accessed June 19, 2026

40 Zhi Zhenfeng (支振锋), *China's Experience in AI Legislation (人工智能立法的中国经验)*, http://www.legaldaily.com.cn/index/content/2026-04/08/content_9369727.html, Legal Daily (法治日报), April 2026, accessed June 19, 2026; Zhang Ping (张平), *Building a Flexible and Adaptable Artificial Intelligence Legislative System (构建灵活适配的人工智能立法体系)*, http://www.news.cn/tech/20260409/37f73e721be34339a1f7ff8052be9676/c.html, Xinhua, April 2026, accessed June 19, 2026

41 Zhou Hui (周辉), *How to Understand Comprehensive AI Legislation? | Reception Room (如何理解人工智能综合性立法？| 会客厅)*, https://mp.weixin.qq.com/s/2yt5-Ay7pnyEoZPpQMsR2w, Weixin Official Accounts Platform, May 2026, accessed June 19, 2026

42 *Notice on Soliciting Member Units for the Artificial Intelligence Security Standards Working Group (WG9) (关于征集人工智能安全标准工作组（WG9）成员单位的通知)*, https://www.tc260.org.cn/portal/article/2/5e06ca6eec464ebdb3c4630208d74131, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

43 *Notice on the Establishment of the Artificial Intelligence Security Standards Working Group (WG9) and the Cybersecurity International Standards Research Group (SG1) (关于组建人工智能安全标准工作组（WG9）和网络安全国际标准研究组(SG1）的通知)*, https://www.tc260.org.cn/portal/article/2/c0604f80429e422e83c58a1104d15472, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

44 Ministry of Industry and Information Technology, *The Ministry of Industry and Information Technology's AI Standardization Technical Committee Is Established (工业和信息化部人工智能标准化技术委员会成立)*, https://wap.miit.gov.cn/jgsj/kjs/gzdt/art/2024/art_547fca4ba3cf4f0b8189d8ed7569fe85.html, December 2024, accessed June 19, 2026

45 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026; *Notice on the Release of the Second Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第二批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/1ba8515d8e2a43f6918b05bcf663655b, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

46 *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)*, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F67D3F376E0A0A0FF5317FB36B32A30A, National Public Service Platform for Standards Information, April 2025, accessed June 19, 2026; *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)*, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=82710B59110419C285BDC48AB4D7D1F3, National Public Service Platform for Standards Information, April 2025, accessed June 19, 2026; *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)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=301E0388CB75788DE06397BE0A0AE1B4, National Public Service Platform for Standards Information, February 2025, accessed June 19, 2026

47 Gabriel Wagner et al., *State of AI Safety in China (2025)*, https://concordia-ai.com/research/state-of-ai-safety-in-china-2025/, Concordia AI, July 2025, accessed June 19, 2026

48 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

49 *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)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=301E0388CB75788DE06397BE0A0AE1B4, National Public Service Platform for Standards Information, February 2025, accessed June 19, 2026

50 *Notice on the Release of the "Cybersecurity Standards Practice Guide—Security Emergency Response Guide for Generative Artificial Intelligence Services" (关于发布《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》的通知)*, https://www.tc260.org.cn/portal/article/2/20250909095834, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, September 2025, accessed June 19, 2026

51 *Artificial Intelligence—Risk Management Capability Assessment (人工智能风险管理能力评估)*, https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=120779016BBA61E4408DCECE3EAD92ED, National Public Service Platform for Standards Information, October 2025, accessed June 19, 2026

52 国家市场监督管理总局 and 国家标准化管理委员会, *GB/T 46800-2025: 生成式人工智能技术应用社会影响评估指南 [Social Impact of Generative Artificial Intelligence Technology Application—Evaluation Guidelines]*, 发布日期 and 实施日期: 2025-12-02, 2025, accessed June 26, 2026, https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=BEE34B0A6DB8FA01AD8CCDBA180FBAE7

53 State Council, *The CPC Central Committee and the State Council Issue the "National Master Emergency Response Plan for Emergencies" (中共中央国务院印发《国家突发事件总体应急预案》)*, https://www.gov.cn/zhengce/202502/content_7005635.htm, February 2025, accessed June 19, 2026

54 *Notice on the Release of 5 Cybersecurity Standardization Technical Research Reports (关于发布 5 项网络安全标准化技术研究报告的通知)*, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

55 *Notice on Soliciting Public Comments on the "Cybersecurity Standards Practice Guidelines——Security Guidance for the Deployment and Use of OpenClaw-Type Agents (Draft for Comment)" (关于对《网络安全标准实践指南——OpenClaw 类智能体部署使用安全指引（征求意见稿）》公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/160310cea5f6411d92fd99a52a42424f?sessionid=, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

56 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

57 *General Security Requirements for Artificial Intelligence Agent Application (智能体应用安全基本要求)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=4C5277928DA2411EE06397BE0A0AE436, National Public Service Platform for Standards Information, April 2026, accessed June 19, 2026; *Second Plenary Meeting of the AI Safety Standards Working Group (WG9) in 2026 Held in Shanghai (人工智能安全标准工作组（WG9）2026 年第二次全体成员会议在上海召开)*, https://www.tc260.org.cn/portal/article/1/dd991f9781d64096afd06d232083cf14, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 19, 2026

58 *Artificial Intelligence—Agent Interconnection—Part 2: Identity Code (人工智能智能体互联第 2 部分：身份码)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E11D080EE06397BE0A0A17E5, National Public Service Platform

for Standards Information, July 2025, accessed June 19, 2026; *Artificial Intelligence—Agent Interconnection—Part 3: Identity Management (人工智能智能体互联第 3 部分：身份管理)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E120080EE06397BE0A0A17E5, National Public Service Platform for Standards Information, July 2025, accessed June 19, 2026

59 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *The 2nd 2025 Meeting of the Safety Governance Working Group (WG8) of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Was Successfully Held (工业和信息化部人工智能标准化技术委员会安全治理工作组（WG8）2025 年第 2 次工作组会成功召开)*, https://mp.weixin.qq.com/s/NQuex2V_CzOB3SpXtoDbpQ, Weixin Official Accounts Platform, September 2025, accessed June 19, 2026; *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

54 *Notice on the Release of 5 Cybersecurity Standardization Technical Research Reports (关于发布 5 项网络安全标准化技术研究报告的通知)*, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

55 *Notice on Soliciting Public Comments on the "Cybersecurity Standards Practice Guidelines——Security Guidance for the Deployment and Use of OpenClaw-Type Agents (Draft for Comment)" (关于对《网络安全标准实践指南——OpenClaw 类智能体部署使用安全指引（征求意见稿）》公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/160310cea5f6411d92fd99a52a42424f?sessionid=, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

56 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

57 *General Security Requirements for Artificial Intelligence Agent Application (智能体应用安全基本要求)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=4C5277928DA211EE06397BE0A0AE436, National Public Service Platform for Standards Information, April 2026, accessed June 19, 2026; *Second Plenary Meeting of the AI Safety Standards Working Group (WG9) in 2026 Held in Shanghai (人工智能安全标准工作组（WG9）2026 年第二次全体成员会议在上海召开)*, https://www.tc260.org.cn/portal/article/1/dd991f9781d64096afd06d232083cf14, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 19, 2026

58 *Artificial Intelligence—Agent Interconnection—Part 2: Identity Code (人工智能智能体互联第 2 部分：身份码)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E11D080EE06397BE0A0A17E5, National Public Service Platform for Standards Information, July 2025, accessed June 19, 2026; *Artificial Intelligence—Agent Interconnection—Part 3: Identity Management (人工智能智能体互联第 3 部分：身份管理)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E120080EE06397BE0A0A17E5, National Public Service Platform for Standards Information, July 2025, accessed June 19, 2026

59 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *The 2nd 2025 Meeting of the Safety Governance Working Group (WG8) of the Artificial Intelligence Standardization Technical Committee of the Ministry of*

Industry and Information Technology Was Successfully Held (工业和信息化部人工智能标准化技术委员会安全治理工作组（WG8）2025 年第 2 次工作组会成功召开), https://mp.weixin.qq.com/s/NQuex2V_CzOB3SpXtoDbpQ, Weixin Official Accounts Platform, September 2025, accessed June 19, 2026; *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

54 *Notice on the Release of 5 Cybersecurity Standardization Technical Research Reports (关于发布 5 项网络安全标准化技术研究报告的通知)*, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

55 *Notice on Soliciting Public Comments on the "Cybersecurity Standards Practice Guidelines——Security Guidance for the Deployment and Use of OpenClaw-Type Agents (Draft for Comment)" (关于对《网络安全标准实践指南——OpenClaw 类智能体部署使用安全指引（征求意见稿）》公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/160310cea5f6411d92fd99a52a42424f?sessionid=, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

56 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

57 *General Security Requirements for Artificial Intelligence Agent Application (智能体应用安全基本要求)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=4C5277928DA2411EE06397BE0A0AE436, National Public Service Platform for Standards Information, April 2026, accessed June 19, 2026; *Second Plenary Meeting of the AI Safety Standards Working Group (WG9) in 2026 Held in Shanghai (人工智能安全标准工作组（WG9）2026 年第二次全体成员会议在上海召开)*, https://www.tc260.org.cn/portal/article/1/dd991f9781d64096afd06d232083cf14, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 19, 2026

58 *Artificial Intelligence—Agent Interconnection—Part 2: Identity Code (人工智能智能体互联第 2 部分：身份码)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E11D080EE06397BE0A0A17E5, National Public Service Platform for Standards Information, July 2025, accessed June 19, 2026; *Artificial Intelligence—Agent Interconnection—Part 3: Identity Management (人工智能智能体互联第 3 部分：身份管理)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=3EFBBC58E120080EE06397BE0A0A17E5, National Public Service Platform for Standards Information, July 2025, accessed June 19, 2026

59 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *The 2nd 2025 Meeting of the Safety Governance Working Group (WG8) of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Was Successfully Held (工业和信息化部人工智能标准化技术委员会安全治理工作组（WG8）2025 年第 2 次工作组会成功召开)*, https://mp.weixin.qq.com/s/NQuex2V_CzOB3SpXtoDbpQ, Weixin Official Accounts Platform, September 2025, accessed June 19, 2026; *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

60 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *The 2nd 2025 Meeting of the Safety Governance Working Group (WG8) of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Was Successfully Held (工业和信息化部人工智能标准化技术委员会安全治理工作组（WG8）2025 年第 2 次工作组会成功召开)*, https://mp.weixin.qq.com/s/NQuex2V_CzOB3SpXtoDbpQ, Weixin Official Accounts Platform, September 2025, accessed June 19, 2026; *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

61 *Notice on Soliciting Public Comments on the "Cybersecurity Standards Practice Guidelines——Security Guidance for the Deployment and Use of OpenClaw-Type Agents (Draft for Comment)" (关于对《网络安全标准实践指南——OpenClaw 类智能体部署使用安全指引（征求意见稿）》公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/160310cea5f6411d92fd99a52a42424f?sessionid=, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 19, 2026

62 *Notice on Issuing the "2026 Annual Work Priorities of the National Information Security Standardization Technical Committee" (关于印发《全国网络安全标准化技术委员会 2026 年度工作要点》的通知)*, https://www.tc260.org.cn/portal/article/2/3d841d17f2d1419eba45379231171242, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 22, 2026

63 *Notice on the Release of 5 Cybersecurity Standardization Technical Research Reports (关于发布 5 项网络安全标准化技术研究报告的通知)*, https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

64 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026; *Second Plenary Meeting of the AI Safety Standards Working Group (WG9) in 2026 Held in Shanghai (人工智能安全标准工作组（WG9）2026 年第二次全体成员会议在上海召开)*, https://www.tc260.org.cn/portal/article/1/dd991f9781d64096afd06d232083cf14, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 19, 2026

65 *General Security Requirements for Artificial Intelligence Agent Application (智能体应用安全基本要求)*, https://std.samr.gov.cn/gb/search/gbDetailed?id=4C5277928DA2411EE06397BE0A0AE436, National Public Service Platform for Standards Information, April 2026, accessed June 19, 2026

66 *Notice on Issuing the "2026 Annual Work Priorities of the National Information Security Standardization Technical Committee" (关于印发《全国网络安全标准化技术委员会 2026 年度工作要点》的通知)*, https://www.tc260.org.cn/portal/article/2/3d841d17f2d1419eba45379231171242, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 22, 2026

67 *Notice on Soliciting Participating Drafting Units for the Standard "Cybersecurity Technology—Security Guidelines for AI Technology Applications Involving Minors" (关于征集《网络安全技术人工智能技术涉及未成年人应用安全指南》标准参编单位的通知)*, https://www.tc260.org.cn/portal/article/2/20250925094625, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, September 2025, accessed June 22, 2026

68 Shanghai AI Lab (上海人工智能实验室), *Shanghai AI Lab Leads the Establishment of the Artificial Intelligence Safety Standards Working Group and Holds Its First Meeting (上海人工智能实验室牵头筹建人工智能安全标准工作组并召开首次会*

议), https://mp.weixin.qq.com/s/4J8xGBFyvm2s6Yzp7sAsHA, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026; Third Research Institute of the Ministry of Public Security, *TC260 AI Safety Standards Working Group WG9 Officially Established (TC260 人工智能安全标准工作组 WG9 正式成立)*, https://mp.weixin.qq.com/s/ZDtrvChhOPl9BoOfO6iY1g, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

69 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2025 (关于发布 2025 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/20250401144126, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2025, accessed June 22, 2026; *Notice on Issuing the Plan for 7 Recommended National Cybersecurity Standards (关于下达 7 项网络安全推荐性国家标准计划的通知)*, https://www.tc260.org.cn/portal/article/2/ca8615d4849344f68e3494b656eb97af, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, February 2026, accessed June 22, 2026

70 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

69 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2025 (关于发布 2025 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/20250401144126, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2025, accessed June 22, 2026; *Notice on Issuing the Plan for 7 Recommended National Cybersecurity Standards (关于下达 7 项网络安全推荐性国家标准计划的通知)*, https://www.tc260.org.cn/portal/article/2/ca8615d4849344f68e3494b656eb97af, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, February 2026, accessed June 22, 2026

70 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

69 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2025 (关于发布 2025 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/20250401144126, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2025, accessed June 22, 2026; *Notice on Issuing the Plan for 7 Recommended National Cybersecurity Standards (关于下达 7 项网络安全推荐性国家标准计划的通知)*, https://www.tc260.org.cn/portal/article/2/ca8615d4849344f68e3494b656eb97af, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, February 2026, accessed June 22, 2026

70 *Notice on the Release of the First Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第一批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/df9022b9293c465a83a15931b2903175, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 19, 2026

71 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026

72 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *The 2nd 2025 Meeting of the Safety Governance Working Group (WG8) of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Was Successfully Held (工业和信息化部人工智能标准化技术委员会安全治理工作组（WG8）2025 年第 2 次工作组会成功召开)*, https://mp.weixin.qq.com/s/NQuex2V_CzOB3SpXtoDbpQ,

Weixin Official Accounts Platform, September 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

73 Ministry of Industry and Information Technology, *Notice of the General Office of the Ministry of Industry and Information Technology on Implementing the Pilot Program for the Ethical Review of AI Science and Technology and Services (工业和信息化部办公厅关于实施人工智能科技伦理审查与服务先导计划的通知)*, https://www.miit.gov.cn/jgsj/kjs/wjfb/art/2026/art_1a0b2b0b0deb4709920287 1e0b70551b.html, April 2026, accessed June 19, 2026

74 *TC260-005 Ethical Security Guidelines for Artificial Intelligence Applications 1.0 (《人工智能应用伦理安全指引 1.0》)*, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 22, 2026

75 *TC260-005 Ethical Security Guidelines for Artificial Intelligence Applications 1.0 (《人工智能应用伦理安全指引 1.0》)*, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 22, 2026

76 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

77 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

78 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

79 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New*

Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知), https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

80 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

81 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

82 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

83 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

84 *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026

75 *TC260-005 Ethical Security Guidelines for Artificial Intelligence Applications 1.0 (《人工智能应用伦理安全指引 1.0》)*, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 22, 2026

76 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

77 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New*

Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知), https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

78 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

79 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

80 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

81 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

82 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年"标准周"活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

83 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年"标准周"活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

84 *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026

75 *TC260-005 Ethical Security Guidelines for Artificial Intelligence Applications 1.0 (《人工智能应用伦理安全指引 1.0》)*, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 22, 2026

76 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

77 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

78 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

79 Fan Kefeng (范科峰), *Expert Interpretation | Fan Kefeng: Consolidating the Ethical Foundation to Safeguard the Intelligent Era——Interpretation of the "Measures for the Ethical Review of AI Science and Technology and Services (Trial)" (专家解读 | 范科峰：夯实伦理根基护航智能时代——《人工智能科技伦理审查与服务办法（试行）》解读)*, https://mp.weixin.qq.com/s/MGYJvGXbX8Qiab7enBPUGA, Weixin Official Accounts Platform, April 2026, accessed June 19, 2026; AI subcommittee of TC 28 on information technology, *Collaborative Governance Builds Ethics, Standards Empower a New Journey | Notice of the Special Seminar on AI Ethics Governance and the Yangtze River Delta Regional Working Group Meeting of the National Information Technology Standardization Committee's AI Subcommittee (协同共治筑伦理，标准赋能启新程 | 人工智能伦理治理专题研讨会暨全国信标委人工智能分委会长三角区域工作组会议会议通知)*, https://mp.weixin.qq.com/s/14AN4AWAsIR-eOYrabP9uw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

80 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

81 AI subcommittee of TC 28 on information technology, *Symposium on AI Ethics Governance Successfully Held in Shanghai (人工智能伦理治理专题研讨会在沪成功召开)*, https://mp.weixin.qq.com/s/o1NTBQ8jDnmuPkOm14gccw, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

82 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

83 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

84 *MIIT TC1 WG8 Safety Governance Group's 3rd Working Group Meeting of 2025 Successfully Convened (MIIT TC1 WG8 安全治理组 2025 年第 3 次工作组会成功召开)*, https://miittc1.caict.ac.cn/newsDetail/?id=59&type=1, Ministry of Industry and Information Technology (工业和信息化部), November 2025, accessed June 19, 2026

85 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

86 Cyberspace Administration of China et al., *Notice on the Issuance of the "Measures for Labeling 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*, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm, March 2025, accessed June 22, 2026

87 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026; *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026

88 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

89 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

90 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7

Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

91 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

92 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026; Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026 年第 2 次工作组会议成功召开)*, https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

93 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

94 Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026 年第 2 次工作组会议成功召开)*, https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

88 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

89 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知)*, https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

90 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

91 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

92 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026; Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026*

年第 2 次工作组会议成功召开), https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

93 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

94 Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026 年第 2 次工作组会议成功召开*), https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

88 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知*), https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

89 *Notice on the Release of 6 Cybersecurity Standards Practice Guidelines Including "Labeling Methods for AI-Generated Synthetic Content—Implicit Labeling of File Metadata—Text Files" (关于发布《人工智能生成合成内容标识方法文件元数据隐式标识文本文件》等 6 项网络安全标准实践指南的通知*), https://www.tc260.org.cn/portal/article/2/20250828165129, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, August 2025, accessed June 19, 2026

90 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开*), https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

91 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

92 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026; Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026 年第 2 次工作组会议成功召开*), https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

93 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

94 Ministry of Industry and Information Technology, *MIIT/TC1 WG8 Security Governance Group's 2nd Working Group Meeting of 2026 Successfully Held (MIIT/TC1 WG8 安全治理组 2026 年第 2 次工作组会议成功召开*), https://mp.weixin.qq.com/s/tmpdgH5ksZXB-fq7LuWHPw, Weixin Official Accounts Platform, April 2026, accessed June 22, 2026

95 *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开*), https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official

Accounts Platform, February 2026, accessed June 19, 2026; *Notice on Soliciting Participating Drafting Units for the Standard "Cybersecurity Technology—Security Requirements for Artificial Intelligence 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)*, https://www.tc260.org.cn/portal/article/2/202411113155236, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, November 2024, accessed June 22, 2026

96 *Notice on the Release of the Second Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第二批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/1ba8515d8e2a43f6918b05bcf663655b, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

97 *Notice on Soliciting Public Comments on 4 Cybersecurity Standard Practice Guidelines Including the "Artificial Intelligence Application Security Guidelines: General Provisions (Draft for Comments)" (关于对《人工智能应用安全指引总则（征求意见稿）》等 4 项网络安全标准实践指南公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/e58c0ee56a5a47ebaf6b5404e8611f65, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 22, 2026; *MIIT TC1 WG8 Security Governance Group Successfully Holds Its 1st Working Group Meeting of 2026 (MIIT TC1 WG8 安全治理组 2026 年第 1 次工作组会成功召开)*, https://mp.weixin.qq.com/s/b8JKPXmmdKT6F0a9R9mJ3g, Weixin Official Accounts Platform, February 2026, accessed June 19, 2026

98 *First Member Assembly and Working Group Meeting of the 2025 "Standards Week" Activity of the Artificial Intelligence Standardization Technical Committee of the Ministry of Industry and Information Technology Convened (工业和信息化部人工智能标准化技术委员会 2025 年 "标准周" 活动第一次成员大会及工作组会召开)*, https://mp.weixin.qq.com/s/K7Fh-_sY7BFz6qNFAXH2Ng, Weixin Official Accounts Platform, July 2025, accessed June 19, 2026; *Notice on the Release of the Second Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第二批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/1ba8515d8e2a43f6918b05bcf663655b, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026; *Notice on Soliciting Public Comments on 4 Cybersecurity Standard Practice Guidelines Including the "Artificial Intelligence Application Security Guidelines: General Provisions (Draft for Comments)" (关于对《人工智能应用安全指引总则（征求意见稿）》等 4 项网络安全标准实践指南公开征求意见的通知)*, https://www.tc260.org.cn/portal/article/2/e58c0ee56a5a47ebaf6b5404e8611f65, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, January 2026, accessed June 22, 2026

99 National Technical Committee 260 on Cybersecurity of Standardization Administration of China and National Computer Network Emergency Response Technical Team/Coordination Center of China, *Version 2.0 of the "AI Safety Governance Framework" Released (《人工智能安全治理框架》2.0 版发布)*, https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm, September 2025, accessed June 22, 2026; National Technical Committee 260 on Cybersecurity of Standardization Administration of China, *AI Safety Governance Framework (人工智能安全治理框架)*, https://www.tc260.org.cn/upload/2024-09-09/1725849142029046390.pdf, September 2024, accessed June 22, 2026; *Notice on the Release of the Second Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第二批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/1ba8515d8e2a43f6918b05bcf663655b, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026; *Notice on Issuing the Plan for 16 Recommended National Cybersecurity Standards (关于下达 16 项网络安全推荐性国家标准计划的通知)*, https://www.tc260.org.cn/portal/article/2/4e8836ae89174f91b20a3f0f259c4804, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, March 2026, accessed June 22, 2026

100 Cyberspace Administration of China, *Central Cyberspace Administration Deploys "Qinglang· Rectifying Chaos in AI Applications" Special Campaign (中央网信办部署开展 "清朗· 整治 AI 应用乱象" 专项行动)*, https://www.cac.gov.cn/2026-04/30/c_1779289298718765.htm, April 2026, accessed June 22, 2026

101 *Notice on the Release of the Second Batch of National Cybersecurity Standard Requirements for 2026 (关于发布 2026 年度第二批网络安全国家标准需求的通知)*, https://www.tc260.org.cn/portal/article/2/1ba8515d8e2a43f6918b05bcf663655b, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, April 2026, accessed June 19, 2026

102 *Second Plenary Meeting of the AI Safety Standards Working Group (WG9) in 2026 Held in Shanghai (人工智能安全标准工作组（WG9）2026 年第二次全体成员会议在上海召开)*, https://www.tc260.org.cn/portal/article/1/dd991f9781d64096afd06d232083cf14, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 19, 2026

103 Zhou Bowen (周伯文), *Shanghai AI Laboratory Director Zhou Bowen: Coordinating Development and Security, Accelerating the Construction of an AI Safety Standards System (上海人工智能实验室主任周伯文：统筹发展和安全加快构建人工智能安全标准体系)*, https://mp.weixin.qq.com/s/Mr0drz7qn2l2pmjLgUN6ng?scene=1, Weixin Official Accounts Platform, May 2026, accessed June 22, 2026

104 *Remarks by H.E. Ma Zhaoxu Deputy Minister of Foreign Affairs at the High-Level Multi-stakeholder Informal Meeting to Launch the Global Dialogue on Artificial Intelligence Governance*, https://www.fmprc.gov.cn/mfa_eng/xw/wjbxw/202509/t20250930_11720750.html, Ministry of Foreign Affairs of the People's Republic of China, September 2023, accessed June 23, 2026

105 United Nations General Assembly, *Terms of Reference and Modalities for the Establishment and Functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*, 2025, accessed June 26, 2026, https://docs.un.org/en/A/RES/79/325

106 *Independent International Scientific Panel on AI Panel Members*, https://www.un.org/independent-international-scientific-panel-ai/en/panel-members, United Nations, 2026, accessed June 23, 2026

107 *Global Top 40! United Nations Announces Membership List — SJTU's Song Haitao Selected!*, https://en.sjtu.edu.cn/news-events/news/2475, Shanghai Jiao Tong University, March 2026, accessed June 23, 2026; Shanghai AI Research Institute, *UN University President and UN Under-Secretary-General Tshilidzi Marwala Visits Shanghai AI Research Institute to Discuss the Sustainable Future of Artificial Intelligence*, 2025, accessed June 26, 2026, https://global.sjtu.edu.cn/news/view/1966

108 *Independent International Scientific Panel on AI Panel Members*, https://www.un.org/independent-international-scientific-panel-ai/en/panel-members, United Nations, 2026, accessed June 23, 2026

109 *Wang Yi: All Countries Should Stand Together to Tackle Pressing Challenges and Shore up the Weak Links in Global Governance*, https://www.mfa.gov.cn/eng/wjbzhd/202510/t20251028_11742166.html, Ministry of Foreign Affairs of the People's Republic of China, October 2025, accessed June 23, 2026

110 *Security Council Eightieth Year 10005th Meeting*, https://docs.un.org/en/S/PV.10005, United Nations, August 2025, accessed June 23, 2026

111 *Global AI Governance Initiative*, https://www.mfa.gov.cn/eng/zy/gb/202405/t20240531_11367503.html, Ministry of Foreign Affairs of the People's Republic of China, October 2023, accessed June 23, 2026; *Remarks by Ambassador Zhang Jun at the UN Security Council Briefing on Artificial Intelligence: Opportunities and Risks for International Peace and Security*, https://un.china-mission.gov.cn/eng/hyyfy/202307/t20230719_11114947.htm, Permanent Mission of the People's Republic of China to the UN, July 2023, accessed June 23, 2026

112 *General Assembly: 72nd Plenary Meeting, 80th Session*, https://webtv.un.org/en/asset/k1i/k1i7yufj2r, United Nations, February 2026, accessed June 23, 2026

113 *AI and Chemical Safety and Security Management: Joint OPCW-China Workshop on AI*, https://www.opcw.org/media-centre/news/2025/06/ai-and-chemical-safety-and-security-management-joint-opcw-china-workshop, OPCW, June 2025, accessed June 23, 2026

114 OPCW, *Artificial Intelligence Report of the Scientific Advisory Board's Temporary Working Group*, https://www.opcw.org/sites/default/files/documents/2026/03/Final%20Report%20of%20the%20SAB%27s%20TWG%20on%20AI%20FINAL%20VERSION.pdf, March 2026, accessed June 23, 2026

115 BRICS Leaders, *BRICS Leaders' Statement on the Global Governance of Artificial Intelligence*, 2025, accessed June 26, 2026, https://brics.br/en/documents/presidency-documents/250706_brics_ggai_declarationfinal.pdf/@@download/file

116 Wang Yang (王洋), *Tianjin Declaration of the Council of Heads of State of the Shanghai Cooperation Organization (上海合作组织成员国元首理事会天津宣言)*, https://www.gov.cn/yaowen/liebiao/202509/content_7038676.htm, Government of the People's Republic of China, September 2025, accessed June 23, 2026

117 *Statement of the Council of Heads of State of the Shanghai Cooperation Organisation on Further Deepening International Cooperation in Artificial Intelligence (上海合作组织成员国元首理事会关于进一步深化人工智能国际合作的声明)*, https://chn.sectsco.org/20250901/1969229.html, 上海合作组织, September 2025, accessed June 23, 2026

118 *Chinese Premier Urges G20 to Strive for Broader Global Cooperation for Development*, https://english.www.gov.cn/news/202511/24/content_WS692391a7c6d00ca5f9a07c36.html, Government of the People's Republic of China, November 2025, accessed June 23, 2026

119 *G20 South Africa Summit: LEADERS' DECLARATION*, https://www.g20.org.za/wp-content/uploads/2025/11/2025-G20-Summit-Declaration.pdf, 2025, accessed June 23, 2026

120 G20 Miami 2026, *U.S. G20 Priorities*, 2026, accessed June 26, 2026, https://g20.org/g20-united-states/

121 *Speech by Xi Jinping at the Second Phase Session of the 32nd APEC Economic Leaders' Informal Meeting (Full Text) (习近平在亚太经合组织第三十二次领导人非正式会议第二阶段会议上的讲话（全文）)*, https://www.mfa.gov.cn/zyxw/202511/t20251101_11745396.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

122 *Remarks by H.E. Ma Zhaoxu Deputy Minister of Foreign Affairs at the High-Level Meeting on Global AI Governance*, https://www.mfa.gov.cn/eng/wjb/zzjg_663340/jks_665232/jkxw_665234/202507/t20250730_11679453.html, Ministry of Foreign Affairs of the People's Republic of China, July 2025, accessed June 23, 2026

123 *Remarks by H.E. Ma Zhaoxu Deputy Minister of Foreign Affairs at the High-Level Meeting on Global AI Governance*, https://www.mfa.gov.cn/eng/wjb/zzjg_663340/jks_665232/jkxw_665234/202507/t20250730_11679453.html, Ministry of Foreign Affairs of the People's Republic of China, July 2025, accessed June 23, 2026

124 *Jointly Forging a Sustainable and Brighter Future: Remarks by H.E. Xi Jinping President of the People's Republic of China At Session II of the 32nd APEC Economic Leaders' Meeting*, https://www.fmprc.gov.cn/eng/zy/jj/xjpcx32apecbdhgjxgsfw/202511/t20251101_11745402.html, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

125 *Xi Jinping Announces That China Will Host the 33rd APEC Economic Leaders' Informal Meeting in Shenzhen (习近平宣布中方将在深圳举办亚太经合组织第三十三次领导人非正式会议)*, https://www.fmprc.gov.cn/zyxw/202511/t20251101_11745406.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

126 *APEC Artificial Intelligence (AI) Initiative (2026-2030) | 2025 APEC Leaders' Gyeongju Declaration*, https://www.apec.org/meeting-papers/leaders-declarations/2025/2025-apec-leaders--gyeongju-declaration/apec-artificial-intelligence-(ai)-initiative-(2026-2030), APEC, 2025, accessed June 23, 2026

127 *Remarks at the Opening Ceremony of the Second Southeast Asia Regional Seminar on Biological Arms Control and Biosecurity (第二届生物军控和生物安全东南亚地区研讨会开幕式上的发言)*, Ministry of Foreign Affairs of the People's Republic of China, 2025, accessed June 26, 2026, https://www.mfa.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/jksxwlb_674635/202511/t20251120_11757118.shtml

128 *Plan of Actions on Jointly Building a China-Africa Community with a Shared Future in Cyberspace（2025-2026）*, https://www.cac.gov.cn/2025-09/28/c_1760606713169654.htm, Cyberspace Administration of China, September 2025, accessed June 23, 2026

129 *Full Text: China's Policy Paper on Latin America and the Caribbean*, https://english.www.gov.cn/news/202512/10/content_WS693962c3c6d00ca5f9a08069.html, Government of the People's Republic of China, December 2025, accessed June 23, 2026

130 清华大学人工智能国际治理研究院, *Institute Experts Participate in the Working Group Meeting and Related Side Events of the India AI Impact Summit (研究院专家参加印度人工智能影响峰会工作组会议及相关边会)*, https://hub.baai.ac.cn/view/52759, February 2026, accessed June 23, 2026

131 Yoshua Bengio et al., *International AI Safety Report 2026*, 2026, accessed June 26, 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

132 *Global AI Governance Action Plan*, https://www.fmprc.gov.cn/eng/wjb/zzjg_663340/jks_665232/kjlc_665236/AI/202507/t20250729_11679232.html, Ministry of Foreign Affairs of the People's Republic of China, July 2025, accessed June 23, 2026

133 *Ministry of Foreign Affairs: China and the US Agree to Launch Intergovernmental Dialogue on Artificial Intelligence (外交部：中美同意开展人工智能政府间对话)*, https://www.news.cn/20260519/201e842d69394fb4ac2038edea359254/c.html, Xinhua, May 2026, accessed June 23, 2026

134 David Lawder, *US Not in a Hurry to Extend China Trade Truce, Bessent Says*, 2026, accessed June 26, 2026, https://www.reuters.com/world/china/us-not-hurry-extend-china-trade-truce-bessent-says-2026-05-19/?mc_cid=d28efd189f&mc_eid=84d7fa0efa

135 *Readout of President Joe Biden's Meeting with President Xi Jinping of the People's Republic of China*, 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/, The White House, January 2025, accessed June 23, 2026

136 *Building a "Constructive China-U.S. Strategic Stability Relationship" Is the Most Important Political Consensus (构建"中美建设性战略稳定关系"是最重要政治共识)*, https://paper.people.com.cn/rmrb/pc/content/202605/16/content_30157121.html, People's Daily (人民日报), May 2026, accessed June 23, 2026

137 *Xi Jinping Announces That China Will Host the 33rd APEC Economic Leaders' Informal Meeting in Shenzhen (习近平宣布中方将在深圳举办亚太经合组织第三十三次领导人非正式会议)*, https://www.fmprc.gov.cn/zyxw/202511/t20251101_11745406.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

138 G20 Miami 2026, *U.S. G20 Priorities*, 2026, accessed June 26, 2026, https://g20.org/g20-united-states/

139 Gabriel Wagner and Erik Lindblad, *China-US launch AI dialogue. Chinese experts are cautiously optimistic.*, 2026, accessed June 26, 2026, https://aisafetychina.substack.com/p/china-us-launch-ai-dialogue-chinese

140 *President Xi Jinping Meets with U.K. Prime Minister Keir Starmer*, https://www.fmprc.gov.cn/eng/xw/zyxw/202601/t20260129_11847645.html, Ministry of Foreign Affairs of the People's Republic of China, January 2026, accessed June 23, 2026

141 *Xi Jinping Holds Talks with French President Macron (习近平同法国总统马克龙举行会谈)*, https://www.mfa.gov.cn/zyxw/202512/t20251204_11766487.shtml, Ministry of Foreign Affairs of the People's Republic of China, December 2025, accessed June 23, 2026; *Xi Jinping Meets with German Chancellor Merz (习近平会见德国总理默茨)*, https://www.mfa.gov.cn/zyxw/202602/t20260225_11863534.shtml, Ministry of Foreign Affairs of the People's Republic of China, February 2026, accessed June 23, 2026; *Xi Jinping Meets with Irish Prime Minister Martin (习近平会见爱尔兰总理马丁)*, https://www.mfa.gov.cn/web/zyxw/202601/t20260105_11806398.shtml, Ministry of Foreign Affairs of the People's Republic of China, January 2026, accessed June 23, 2026; *President Xi Jinping Meets with Spanish Prime Minister Pedro Sánchez*, https://www.fmprc.gov.cn/eng/xw/zyxw/202604/t20260414_11891771.html, Ministry of Foreign Affairs of the People's Republic of China, April 2026, accessed June 23, 2026; *President Xi Jinping Meets with Slovak Prime Minister Robert Fico*, https://un.china-mission.gov.cn/eng/zgyw/202509/t20250904_11702300.htm, Permanent Mission of the People's Republic of China to the UN, September 2025, accessed June 23, 2026; *President Xi Jinping Meets with Finnish Prime Minister Petteri Orpo*, https://www.fmprc.gov.cn/eng/xw/zyxw/202601/t20260127_11846160.html, Ministry of Foreign Affairs of the People's Republic of China, January 2026, accessed June 23, 2026; *President Xi Jinping Meets with Portuguese Prime Minister Luís Montenegro*, https://www.fmprc.gov.cn/eng/xw/zyxw/202509/t20250909_11705170.html, Ministry of Foreign Affairs of the People's Republic of China, September 2025, accessed June 23, 2026

142 *Commission and China Hold Second High-level Digital Dialogue*, https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4488, European Commission, September 2023, accessed June 23, 2026

143 European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*, Official Journal of the European Union, L 2024/1689, 2024, accessed June 26, 2026, https://eur-lex.europa.eu/eli/reg/2024/1689/oj

144 *Singapore and China Deepen Exchange in Digital Economy at 2nd Singapore-China Digital Policy Dialogue (DPD) in Chongqing, China*, https://www.mddi.gov.sg/newsroom/singapore-and-china-deepen-exchange-in-digital-economy-at-2nd-singapore-china-digital-policy-dialogue--dpd--in-chongqing--china/, Ministry of Digital Development and Information, September 2025, accessed June 23, 2026

145 *Joint Statement of the People's Republic of China and the Russian Federation on Further Strengthening Comprehensive Strategic Coordination and Deepening Good-Neighborly Friendship and Cooperation (中华人民共和国和俄罗斯联邦关于进一步加强全面战略协作、深化睦邻友好合作的联合声明)*, https://www.mfa.gov.cn/zyxw/202605/t20260521_11914932.shtml, Ministry of Foreign Affairs of the People's Republic of China, May 2026, accessed June 23, 2026

146 *President Xi Jinping Holds Talks with Russian President Vladimir Putin*, https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202509/t20250902_11700541.html, Ministry of Foreign Affairs of the People's Republic of China, September 2025, accessed June 23, 2026

147 *Joint Communiqué of the 30th Regular Meeting of the Chinese and Russian Heads of Government (Full Text) (中俄总理第三十次定期会晤联合公报（全文）)*, https://www.fmprc.gov.cn/zyxw/202511/t20251104_11746926.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

148 *Guidelines for the Construction of a National Comprehensive Standardization System for the Artificial Intelligence Industry (2024 Edition) (国家人工智能产业综合标准化体系建设指南（2024 版）)*, https://www.gov.cn/zhengce/zhengceku/202407/P020240702716282797987.pdf, Government of the People's Republic of China, 2024

149 *Global AI Governance Action Plan*, https://www.fmprc.gov.cn/eng/wjb/zzjg_663340/jks_665232/kjlc_665236/AI/202507/t20250729_11679232.html, Ministry of Foreign Affairs of the People's Republic of China, July 2025, accessed June 23, 2026

150 National Technical Committee 260 on Cybersecurity of Standardization Administration of China and National Computer Network Emergency Response Technical Team/Coordination Center of China, *Version 2.0 of the "AI Safety Governance Framework" Released (《人工智能安全治理框架》2.0 版发布)*, https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm, September 2025, accessed June 22, 2026

151 *ISO/IEC JTC 1/SC 42 - Artificial Intelligence*, https://www.iso.org/committee/6794475.html, ISO, September 2023, accessed June 23, 2026

152 *Standardization Administration of China*, https://www.iso.org/member/1635.html, ISO, March 2025, accessed June 23, 2026

153 China Electronics Standardization Institute (CESI), *National AI Standardization Work Report*, 2025, accessed June 26, 2026, https://sesec.eu/wp-content/uploads/2025/09/SESEC-V-Newsletter-Annex-CESI-AI-PPT.pdf; *ISO/IEC JTC 1/SC 42 - Artificial Intelligence*, https://www.iso.org/committee/6794475.html, ISO, September 2023, accessed June 23, 2026

154 *SC 42 Strategic Business Plan 2023*, https://jtc1info.org/wp-content/uploads/2024/02/SC-42-Business-Plan-2023.pdf, ISO, September 2023; China Electronics Standardization Institute (CESI), *National AI Standardization Work Report*, 2025, accessed June 26, 2026, https://sesec.eu/wp-content/uploads/2025/09/SESEC-V-Newsletter-Annex-CESI-AI-PPT.pdf

155 *ISO/IEC JTC 1/SC 42 - Artificial Intelligence*, https://www.iso.org/committee/6794475.html, ISO, September 2023, accessed June 23, 2026

156 *ISO/IEC CD TS 25568*, https://www.iso.org/standard/90754.html, ISO, accessed June 23, 2026

157 SaferAI, *An overview of existing and potential future GenAI/GPAI standards*, 2026, accessed June 26, 2026, https://www.safer-ai.org/an-overview-of-existing-and-potential-future-genai-gpai-standards

158 *China Has Released 30 National Artificial Intelligence Standards (中国已发布 30 项人工智能国家标准)*, https://www.news.cn/government/20250918/cc6356017bd64f3c9fb5994e6ce4ad0c/c.html, Xinhua, September 2025, accessed June 23, 2026

159 China Academy of Information and Communications Technology, *ITU International Standard for Large Model Benchmark Testing Led by CAICT Officially Released (中国信通院牵头的大模型基准测试 ITU 国际标准正式发布)*, https://mp.weixin.qq.com/s/GbhiGCdplGzwzComqKvnOg, Weixin Official Accounts Platform, April 2025, accessed June 23, 2026

160 *Yuwei Wang*, https://ict-ywwang.github.io/, 2026, accessed June 23, 2026

161 ITU, *Key international organizations align on AI standards*, 2025, accessed June 26, 2026, https://www.itu.int/hub/2025/12/key-international-organizations-align-on-ai-standards/

162 IEC, ISO, and ITU, *Seoul Statement - International AI Standards Summit 2025*, https://www.aistandardssummit.org/event/2025/seoul-statement, December 2025, accessed June 23, 2026

163 World Digital Technology Academy (WDTA), *Single AI Agent Runtime Security Testing Standards*, https://www.wdtacademy.org/Upfile/File/2025-12-04/e15c0b40-850c-49e7-be04-f1876fdc4203..pdf, July 2025

164 Agentic AI Foundation, *Agentic AI Foundation Welcomes 97 New Members As Demand for Open, Collaborative Agent Standardization Increases*, 2026, accessed June 26, 2026, https://aaif.io/press/agentic-ai-foundation-welcomes-97-new-members-as-demand-for-open-collaborative-agent-standardization-increases/

165 The Nuclear Threat Initiative, *Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences*, 2025, accessed June 26, 2026, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

166 The Nuclear Threat Initiative, *CACDA-NTI Joint Statement on Shared Priorities for Strengthened Biosecurity and Responsible AI-Biotechnology Innovation*, 2026, accessed June 26, 2026, https://www.nti.org/analysis/articles/cacda-nti-joint-statement-on-shared-priorities-for-strengthened-biosecurity-and-responsible-ai-biotechnology-innovation/

167 *RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS*, https://unodaweb-meetings.unoda.org/public/2025-12/AIxBio%20Recommendations%20INHR.pdf, INHR, December 2025

168 International Dialogues on AI Safety, *IDAIS-London, 2026: Statement*, London, UK, April 17–19, 2026, 2026, accessed June 26, 2026, https://idais.ai/dialogue/idais-london/

169 International Dialogues on AI Safety, *Consensus Statement on Ensuring Alignment and Human Control of Advanced AI Systems to Safeguard Human Flourishing*, Statement from IDAIS-Shanghai, Shanghai, China, July 22–25, 2025, 2025, accessed June 26, 2026, https://idais.ai/dialogue/idais-shanghai/

170 *The 13th China–U.S. Dialogue on Artificial Intelligence and International Security Successfully Held*, https://ciss.tsinghua.edu.cn/info/event/8777, Center For International Security and Strategy Tsinghua University, December 2025, accessed June 23, 2026; *CISS Hosts the XIV Round of "U.S.-China Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/event/9027, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

171 Kyle Chan et al., *AI risks from non-state actors*, 2026, accessed June 26, 2026, https://www.brookings.edu/articles/ai-risks-from-non-state-actors/

172 Brookings, *Glossary of artificial intelligence terms*, 2026, accessed June 26, 2026, https://www.brookings.edu/articles/glossary-of-artificial-intelligence-terms/

173 *Misuse of Advanced AI by Non-State Actors and Its Impacts on International Security*, https://ciss.tsinghua.edu.cn/info/CISSReports/9093, Center For International Security and Strategy Tsinghua University, March 2026, accessed June 23, 2026

174 *CISS Hosts the XIV Round of "U.S.-China Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/event/9027, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

175 Bob Davis, *Back on Track?*, 2024, accessed June 26, 2026, https://www.thewirechina.com/2024/01/14/back-on-track-two-dialogues-u-s-china/; INHR, *AI*, 2025, accessed June 26, 2026, https://inhr.org/ai; *RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS*, https://unodaweb-meetings.unoda.org/public/2025-12/AIxBio%20Recommendations%20INHR.pdf, INHR, December 2025

176 *Center Advances U.S.-China Understanding of AI Governance*, https://law.yale.edu/yls-today/news/center-advances-us-china-understanding-ai-governance, Yale Law School, August 2024, accessed June 23, 2026

177 International Dialogues on AI Safety, *IDAIS-Oxford, 2023: Statement*, Ditchley Park, UK, October 31, 2023, 2023, accessed June 26, 2026, https://idais.ai/dialogue/idais-oxford/; International Dialogues on AI Safety, *Consensus Statement on Red Lines in Artificial Intelligence*, IDAIS-Beijing, Beijing, China, March 10–11, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-beijing/; International Dialogues on AI Safety, *Consensus Statement on AI Safety as a Global Public Good*,

Statement from IDAIS-Venice, Venice, Italy, September 5–8, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-venice/; International Dialogues on AI Safety, *Consensus Statement on Ensuring Alignment and Human Control of Advanced AI Systems to Safeguard Human Flourishing*, Statement from IDAIS-Shanghai, Shanghai, China, July 22–25, 2025, 2025, accessed June 26, 2026, https://idais.ai/dialogue/idais-shanghai/

178 The Nuclear Threat Initiative, *NTI Convenes First International AI-Bio Forum*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-convenes-the-first-international-ai-bio-forum/; The Nuclear Threat Initiative, *Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences*, 2025, accessed June 26, 2026, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

179 *The Royal Society and Chinese Academy of Sciences Policy Dialogue on AI*, https://royalsociety.org/-/media/blogs/2021/04/china-ai/RS-CAS-AI-policy-workshop-report.pdf?la=en-GB&hash=A1E3BDFCA8C55C46672412A415A958A5, The Royal Society, September 2020; *Chinese Academy of Sciences and UK Royal Society Jointly Hold UK-China Artificial Intelligence Policy Dialogue 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)*, https://web.archive.org/web/20240221151654/http://www.ia.cas.cn/xwzx/kydt/202010/t20201016_5718238.html, Chinese Academy of Sciences, October 2020, accessed June 23, 2026; *The Third Chinese Academy of Sciences—Royal Society Symposium on AI Ethics 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)*, https://bic.cas.cn/gjhzdt/202409/t20240930_5034430.html, Bureau of International Cooperation, Chinese Academy of Sciences (中国科学院国际合作局), September 2024, accessed June 23, 2026; *The Fourth Chinese Academy of Sciences—Royal Society AI Ethics Symposium Held in Hong Kong and Shenzhen (第四届中国科学院—英国皇家学会人工智能伦理研讨会在香港和深圳举行)*, https://www.cas.cn/yx/202601/t20260126_5097292.shtml, Chinese Academy of Sciences, January 2026, accessed June 23, 2026

180 *CISS Holds China-EU Dialogue on Artificial Intelligence and International Security (CISS 举办中欧人工智能与国际安全对话会)*, https://ciss.tsinghua.edu.cn/info/wzjx/6951, Center For International Security and Strategy Tsinghua University, February 2024, accessed June 23, 2026; *CISS Holds the Fourth Round of the "China-EU Dialogue on Artificial Intelligence and International Security" (CISS 举办第四轮"中欧人工智能与国际安全对话")*, https://mp.weixin.qq.com/s/oNxcoScY3we7nbuhLg9Nsg, Center For International Security and Strategy Tsinghua University, March 2025, accessed June 23, 2026; *CISS Hosts the 5th Round of "China-Europe Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/Conferences/9026, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

181 *P5 Experts Roundtable – Online Meeting on the AI Nuclear Nexus on 24 June 2024*, https://www.gcsp.ch/news/p5-experts-roundtable-online-meeting-ai-nuclear-nexus-24-june-2024, Geneva Centre for Security Policy, July 2024, accessed June 23, 2026

182 Concordia AI, *Concordia AI 2025 Impact Highlights*, 2026, accessed June 26, 2026, https://aisafetychina.substack.com/p/concordia-ai-2025-impact-highlights

183 The Nuclear Threat Initiative, *NTI and CACDA Co-Convene Track II Biosecurity Dialogue*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-co-convene-track-ii-biosecurity-dialogue/; The Nuclear Threat Initiative, *NTI and CACDA Foster Cooperation Among U.S. and Chinese Experts to Advance Biosecurity*, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-foster-cooperation-among-u-s-and-chinese-experts-to-advance-biosecurity/

184 National Committee on United States-China Relations, *U.S.-China Track II Dialogue on the Digital Economy*, accessed June 26, 2026, https://www.ncuscr.org/program/us-china-track-ii-dialogue-digital-economy/

185 *1st Sino-European Cyber Dialogue*, https://www.gcsp.ch/events/1st-sino-european-cyber-dialogue, Geneva Centre for Security Policy, 2014, accessed June 23, 2026; *CICIR and the European School of Management and Technology Jointly Host the "11th China-Europe Cyber Track II 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")*, http://www.cicir.ac.cn/NEW/event.html?id=6f23ec7f-0e17-4d57-950a-98c7a0667615, China Institutes of Contemporary International Relations (中国现代国际关系研究院), November 2023, accessed June 23, 2026; *Opening Remarks by Wang Lei, Coordinator for Cyber and Digital Affairs of the Ministry of Foreign Affairs, at the China-EU Cyber Track II Dialogue (外交部网络和数字事务协调员王磊参加中欧网络二轨对话的开场致辞)*, https://www.fmprc.gov.cn/web

/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202511/t20251113_11752342.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

174 *CISS Hosts the XIV Round of "U.S.-China Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/event/9027, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

175 Bob Davis, *Back on Track?*, 2024, accessed June 26, 2026, https://www.thewirechina.com/2024/01/14/back-on-track-two-dialogues-u-s-china/; INHR, *AI*, 2025, accessed June 26, 2026, https://inhr.org/ai; *RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS*, https://unodaweb-meetings.unoda.org/public/2025-12/AIxBio%20Recommendations%20INHR.pdf, INHR, December 2025

176 *Center Advances U.S.-China Understanding of AI Governance*, https://law.yale.edu/yls-today/news/center-advances-us-china-understanding-ai-governance, Yale Law School, August 2024, accessed June 23, 2026

177 International Dialogues on AI Safety, *IDAIS-Oxford, 2023: Statement*, Ditchley Park, UK, October 31, 2023, 2023, accessed June 26, 2026, https://idais.ai/dialogue/idais-oxford/; International Dialogues on AI Safety, *Consensus Statement on Red Lines in Artificial Intelligence*, IDAIS-Beijing, Beijing, China, March 10–11, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-beijing/; International Dialogues on AI Safety, *Consensus Statement on AI Safety as a Global Public Good*, Statement from IDAIS-Venice, Venice, Italy, September 5–8, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-venice/; International Dialogues on AI Safety, *Consensus Statement on Ensuring Alignment and Human Control of Advanced AI Systems to Safeguard Human Flourishing*, Statement from IDAIS-Shanghai, Shanghai, China, July 22–25, 2025, 2025, accessed June 26, 2026, https://idais.ai/dialogue/idais-shanghai/

178 The Nuclear Threat Initiative, *NTI Convenes First International AI-Bio Forum*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-convenes-the-first-international-ai-bio-forum/; The Nuclear Threat Initiative, *Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences*, 2025, accessed June 26, 2026, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

179 *The Royal Society and Chinese Academy of Sciences Policy Dialogue on AI*, https://royalsociety.org/-/media/blogs/2021/04/china-ai/RS-CAS-AI-policy-workshop-report.pdf?la=en-GB&hash=A1E3BDFCA8C55C46672412A415A958A5, The Royal Society, September 2020; *Chinese Academy of Sciences and UK Royal Society Jointly Hold UK-China Artificial Intelligence Policy Dialogue 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)*, https://web.archive.org/web/20240221151654/http://www.ia.cas.cn/xwzx/kydt/202010/t20201016_5718238.html, Chinese Academy of Sciences, October 2020, accessed June 23, 2026; *The Third Chinese Academy of Sciences—Royal Society Symposium on AI Ethics 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)*, https://bic.cas.cn/gjhzdt/202409/t20240930_5034430.html, Bureau of International Cooperation, Chinese Academy of Sciences (中国科学院国际合作局), September 2024, accessed June 23, 2026; *The Fourth Chinese Academy of Sciences—Royal Society AI Ethics Symposium Held in Hong Kong and Shenzhen (第四届中国科学院—英国皇家学会人工智能伦理研讨会在香港和深圳举行)*, https://www.cas.cn/yx/202601/t20260126_5097292.shtml, Chinese Academy of Sciences, January 2026, accessed June 23, 2026

180 *CISS Holds China-EU Dialogue on Artificial Intelligence and International Security (CISS 举办中欧人工智能与国际安全对话会)*, https://ciss.tsinghua.edu.cn/info/wzjx/6951, Center For International Security and Strategy Tsinghua University, February 2024, accessed June 23, 2026; *CISS Holds the Fourth Round of the "China-EU Dialogue on Artificial Intelligence and International Security" (CISS 举办第四轮"中欧人工智能与国际安全对话")*, https://mp.weixin.qq.com/s/oNxcoScY3we7nbuhLg9Nsg, Center For International Security and Strategy Tsinghua University, March 2025, accessed June 23, 2026; *CISS Hosts the 5th Round of "China-Europe Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/Conferences/9026, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

181 *P5 Experts Roundtable – Online Meeting on the AI Nuclear Nexus on 24 June 2024*, https://www.gcsp.ch/news/p5-experts-roundtable-online-meeting-ai-nuclear-nexus-24-june-2024, Geneva Centre for Security Policy, July 2024, accessed June 23, 2026

182 Concordia AI, *Concordia AI 2025 Impact Highlights*, 2026, accessed June 26, 2026, https://aisafetychina.substack.com/p/concordia-ai-2025-impact-highlights

183 The Nuclear Threat Initiative, *NTI and CACDA Co-Convene Track II Biosecurity Dialogue*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-co-convene-track-ii-biosecurity-dialogue/; The Nuclear Threat Initiative, *NTI and CACDA Foster Cooperation Among U.S. and Chinese Experts to Advance Biosecurity*, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-foster-cooperation-among-u-s-and-chinese-experts-to-advance-biosecurity/

184 National Committee on United States-China Relations, *U.S.-China Track II Dialogue on the Digital Economy*, accessed June 26, 2026, https://www.ncuscr.org/program/us-china-track-ii-dialogue-digital-economy/

185 *1st Sino-European Cyber Dialogue*, https://www.gcsp.ch/events/1st-sino-european-cyber-dialogue, Geneva Centre for Security Policy, 2014, accessed June 23, 2026; *CICIR and the European School of Management and Technology Jointly Host the "11th China-Europe Cyber Track II 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")*, http://www.cicir.ac.cn/NEW/event.html?id=6f23ec7f-0e17-4d57-950a-98c7a0667615, China Institutes of Contemporary International Relations (中国现代国际关系研究院), November 2023, accessed June 23, 2026; *Opening Remarks by Wang Lei, Coordinator for Cyber and Digital Affairs of the Ministry of Foreign Affairs, at the China-EU Cyber Track II Dialogue (外交部网络和数字事务协调员王磊参加中欧网络二轨对话的开场致辞)*, https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202511/t20251113_11752342.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

174 *CISS Hosts the XIV Round of "U.S.-China Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/event/9027, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

175 *Meet the INHR Team - Our Directors*, https://inhr.org/who-we-are, INHR, accessed June 23, 2026

176 Bob Davis, *Back on Track?*, 2024, accessed June 26, 2026, https://www.thewirechina.com/2024/01/14/back-on-track-two-dialogues-u-s-china/; INHR, *AI*, 2025, accessed June 26, 2026, https://inhr.org/ai; *RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS*, https://unodaweb-meetings.unoda.org/public/2025-12/AIxBio%20Recommendations%20INHR.pdf, INHR, December 2025

177 *Center Advances U.S.-China Understanding of AI Governance*, https://law.yale.edu/yls-today/news/center-advances-us-china-understanding-ai-governance, Yale Law School, August 2024, accessed June 23, 2026

178 International Dialogues on AI Safety, *IDAIS-Oxford, 2023: Statement*, Ditchley Park, UK, October 31, 2023, 2023, accessed June 26, 2026, https://idais.ai/dialogue/idais-oxford/; International Dialogues on AI Safety, *Consensus Statement on Red Lines in Artificial Intelligence*, IDAIS-Beijing, Beijing, China, March 10–11, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-beijing/; International Dialogues on AI Safety, *Consensus Statement on AI Safety as a Global Public Good*, Statement from IDAIS-Venice, Venice, Italy, September 5–8, 2024, 2024, accessed June 26, 2026, https://idais.ai/dialogue/idais-venice/; International Dialogues on AI Safety, *Consensus Statement on Ensuring Alignment and Human Control of Advanced AI Systems to Safeguard Human Flourishing*, Statement from IDAIS-Shanghai, Shanghai, China, July 22–25, 2025, 2025, accessed June 26, 2026, https://idais.ai/dialogue/idais-shanghai/

179 The Nuclear Threat Initiative, *NTI Convenes First International AI-Bio Forum*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-convenes-the-first-international-ai-bio-forum/; The Nuclear Threat Initiative, *Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences*, 2025, accessed June 26, 2026, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

180 *The Royal Society and Chinese Academy of Sciences Policy Dialogue on AI*, https://royalsociety.org/-/media/blogs/2021/04/china-ai/RS-CAS-AI-policy-workshop-report.pdf?la=en-GB&hash=A1E3BDFCA8C55C46672412A415A958A5, The Royal Society, September 2020; *Chinese Academy of Sciences and UK Royal Society Jointly Hold UK-China Artificial Intelligence Policy Dialogue 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)*, https://web.archive.or

g/web/20240221151654/http://www.ia.cas.cn/xwzx/kydt/202010/t20201016_5718238.html, Chinese Academy of Sciences, October 2020, accessed June 23, 2026; *The Third Chinese Academy of Sciences—Royal Society Symposium on AI Ethics 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)*, https://bic.cas.cn/gjhzdt/202409/t20240930_5034430.html, Bureau of International Cooperation, Chinese Academy of Sciences (中国科学院国际合作局), September 2024, accessed June 23, 2026; *The Fourth Chinese Academy of Sciences—Royal Society AI Ethics Symposium Held in Hong Kong and Shenzhen (第四届中国科学院—英国皇家学会人工智能伦理研讨会在香港和深圳举行)*, https://www.cas.cn/yx/202601/t20260126_5097292.shtml, Chinese Academy of Sciences, January 2026, accessed June 23, 2026

181 *CISS Holds China-EU Dialogue on Artificial Intelligence and International Security (CISS 举办中欧人工智能与国际安全对话会)*, https://ciss.tsinghua.edu.cn/info/wzjx/6951, Center For International Security and Strategy Tsinghua University, February 2024, accessed June 23, 2026; *CISS Holds the Fourth Round of the "China-EU Dialogue on Artificial Intelligence and International Security" (CISS 举办第四轮"中欧人工智能与国际安全对话")*, https://mp.weixin.qq.com/s/oNxcoScY3we7nbuhLg9Nsg, Center For International Security and Strategy Tsinghua University, March 2025, accessed June 23, 2026; *CISS Hosts the 5th Round of "China-Europe Dialogue on AI and International Security"*, https://ciss.tsinghua.edu.cn/info/Conferences/9026, Center For International Security and Strategy Tsinghua University, February 2026, accessed June 23, 2026

182 *P5 Experts Roundtable – Online Meeting on the AI Nuclear Nexus on 24 June 2024*, https://www.gcsp.ch/news/p5-experts-roundtable-online-meeting-ai-nuclear-nexus-24-june-2024, Geneva Centre for Security Policy, July 2024, accessed June 23, 2026

183 Concordia AI, *Concordia AI 2025 Impact Highlights*, 2026, accessed June 26, 2026, https://aisafetychina.substack.com/p/concordia-ai-2025-impact-highlights

184 The Nuclear Threat Initiative, *NTI and CACDA Co-Convene Track II Biosecurity Dialogue*, 2024, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-co-convene-track-ii-biosecurity-dialogue/; The Nuclear Threat Initiative, *NTI and CACDA Foster Cooperation Among U.S. and Chinese Experts to Advance Biosecurity*, accessed June 26, 2026, https://www.nti.org/news/nti-and-cacda-foster-cooperation-among-u-s-and-chinese-experts-to-advance-biosecurity/

185 National Committee on United States-China Relations, *U.S.-China Track II Dialogue on the Digital Economy*, accessed June 26, 2026, https://www.ncuscr.org/program/us-china-track-ii-dialogue-digital-economy/

186 *1st Sino-European Cyber Dialogue*, https://www.gcsp.ch/events/1st-sino-european-cyber-dialogue, Geneva Centre for Security Policy, 2014, accessed June 23, 2026; *CICIR and the European School of Management and Technology Jointly Host the "11th China-Europe Cyber Track II 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")*, http://www.cicir.ac.cn/NEW/event.html?id=6f23ec7f-0e17-4d57-950a-98c7a0667615, China Institutes of Contemporary International Relations (中国现代国际关系研究院), November 2023, accessed June 23, 2026; *Opening Remarks by Wang Lei, Coordinator for Cyber and Digital Affairs of the Ministry of Foreign Affairs, at the China-EU Cyber Track II Dialogue (外交部网络和数字事务协调员王磊参加中欧网络二轨对话的开场致辞)*, https://www.fmprc.gov.cn/web/wjb_673085/zzjg_673183/jks_674633/fywj_674643/202511/t20251113_11752342.shtml, Ministry of Foreign Affairs of the People's Republic of China, November 2025, accessed June 23, 2026

187 Concordia AI, *Chinese Technical AI Safety Paper Database*, https://aisafetychina.com/research/, July 2026, accessed June 23, 2026

188 Concordia AI, *Chinese Technical AI Safety Paper Database*, https://aisafetychina.com/research/, July 2026, accessed June 23, 2026

189 John P.A. Ioannidis, *August 2025 data-update for "Updated science-wide author databases of standardized citation indicators"*, 2025, accessed June 26, 2026, https://doi.org/10.17632/btchxktzyw.8, https://data.mendeley.com/datasets/btchxktzyw/

190 Xiaojun Jia et al., *SkillJect: Effectively Automating Skill-Based Prompt Injection for Skill-Enabled Agents*, arXiv:2602.14211, June 2026, accessed June 22, 2026, arXiv: 2602.14211 [cs.CR]

191 Xinyi Wu et al., *When Bots Take the Bait: Exposing and Mitigating the Emerging Social Engineering Attack in Web Automation Agent*, arXiv:2601.07263, January 2026, accessed June 22, 2026, arXiv: 2601.07263 [cs.CR]

192 Dongrui Liu et al., *AgentDoG: A Diagnostic Guardrail Framework for AI Agent Safety and Security*, arXiv:2601.18491, April 2026, accessed June 22, 2026, arXiv: 2601.18491 [cs.AI]; *TrinityGuard: A Unified Framework for Safeguarding Multi-Agent Systems*, https://arxiv.org/html/2603.15408, accessed June 22, 2026

193 Xinhao Deng et al., *Taming OpenClaw: Security Analysis and Mitigation of Autonomous LLM Agent Threats*, arXiv:2603.11619, March 2026, accessed June 22, 2026, arXiv: 2603.11619 [cs.CR]

194 Zibo Xiao, Jun Sun, and Junjie Chen, *AIR: Improving Agent Safety through Incident Response*, arXiv:2602.11749, February 2026, accessed June 22, 2026, arXiv: 2602.11749 [cs.AI]

195 Dongrui Liu et al., *Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report v1.5*, arXiv:2602.14457, February 2026, accessed June 22, 2026, arXiv: 2602.14457 [cs.AI]; Haibo Tong et al., *ForesightSafety Bench: A Frontier Risk Evaluation and Governance Framework towards Safe AI*, arXiv:2602.14135, February 2026, accessed June 22, 2026, arXiv: 2602.14135 [cs.AI]

196 *[2509.25302] Dive into the Agent Matrix: A Realistic Evaluation of Self-Replication Risk in LLM Agents*, https://arxiv.org/abs/2509.25302, accessed June 22, 2026

197 Qibing Ren et al., *When Autonomy Goes Rogue: Preparing for Risks of Multi-Agent Collusion in Social Systems*, arXiv:2507.14660, July 2025, accessed June 22, 2026, arXiv: 2507.14660 [cs.AI]

198 Shuai Shao et al., *Your Agent May Misevolve: Emergent Risks in Self-evolving LLM Agents*, arXiv:2509.26354, March 2026, accessed June 22, 2026, arXiv: 2509.26354 [cs.AI]

199 Ling Tang et al., *Interpreting Emergent Extreme Events in Multi-Agent Systems*, arXiv:2601.20538, February 2026, accessed June 22, 2026, arXiv: 2601.20538 [cs.MA]

200 Dadi Guo et al., *Are Your Agents Upward Deceivers?*, arXiv:2512.04864, December 2025, accessed June 22, 2026, arXiv: 2512.04864 [cs.AI]

201 Xinyue Wang et al., *From Helpfulness to Toxic Proactivity: Diagnosing Behavioral Misalignment in LLM Agents*, arXiv:2602.04197, February 2026, accessed June 22, 2026, arXiv: 2602.04197 [cs.CL]; Changyi Li et al., *AutoControl Arena: Synthesizing Executable Test Environments for Frontier AI Risk Evaluation*, arXiv:2603.07427, March 2026, accessed June 22, 2026, arXiv: 2603.07427 [cs.AI]

202 Cheng Xia et al., *MIRAGE: Towards AI-Generated Image Detection in the Wild*, arXiv:2508.13223, August 2025, accessed June 22, 2026, arXiv: 2508.13223 [cs.CV]

203 Xinchang Wang et al., *RA-Det: Towards Universal Detection of AI-Generated Images via Robustness Asymmetry*, arXiv:2603.01544, March 2026, accessed June 22, 2026, arXiv: 2603.01544 [cs.CV]

204 Chao Shuai et al., *When Detectors Forget Forensics: Blocking Semantic Shortcuts for Generalizable AI-Generated Image Detection*, arXiv:2603.09242, June 2026, accessed June 22, 2026, arXiv: 2603.09242 [cs.CV]

205 Yuzhuo Chen et al., *TAG-WM: Tamper-Aware Generative Image Watermarking via Diffusion Inversion Sensitivity*, arXiv:2506.23484, December 2025, accessed June 22, 2026, arXiv: 2506.23484 [cs.MM]; Jindong Yang et al., *T2SMark: Balancing Robustness and Diversity in Noise-as-Watermark for Diffusion Models*, arXiv:2510.22366, October 2025, accessed June 22, 2026, arXiv: 2510.22366 [cs.CV]; Yang Yang et al., *SKeDA: A Generative Watermarking Framework for Text-to-video Diffusion Models*, arXiv:2603.00194, February 2026, accessed June 22, 2026, arXiv: 2603.00194 [cs.CV]

206 Yanghao Su et al., *Character as a Latent Variable in Large Language Models: A Mechanistic Account of Emergent Misalignment and Conditional Safety Failures*, arXiv:2601.23081, January 2026, accessed June 22, 2026, arXiv: 2601.23081 [cs.CL]

207 Kai Wang, Yihao Zhang, and Meng Sun, *When Thinking LLMs Lie: Unveiling the Strategic Deception in Representations of Reasoning Models*, arXiv:2506.04909, June 2025, accessed June 22, 2026, arXiv: 2506.04909 [cs.AI]

208 Jiaqi Weng et al., *Safe-SAIL: Towards a Fine-grained Safety Landscape of Large Language Models via Sparse Autoencoder Interpretation Framework*, arXiv:2509.18127, April 2026, accessed June 22, 2026, arXiv: 2509.18127 [cs.LG]

209 Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho, *Attention Slipping: A Mechanistic Understanding of Jailbreak Attacks and Defenses in LLMs*, arXiv:2507.04365, July 2025, accessed June 22, 2026, arXiv: 2507.04365 [cs.CR]; Chen Xiong et al., *Steering Externalities: Benign Activation Steering Unintentionally Increases Jailbreak Risk for Large Language Models*, arXiv:2602.04896, February 2026, accessed June 22, 2026, arXiv: 2602.04896 [cs.CR]

210 Bing Han et al., *Multi-Level Safety Continual Projection for Fine-Tuned Large Language Models without Retraining*, arXiv:2508.09190, January 2026, accessed June 22, 2026, arXiv: 2508.09190 [cs.LG]

211 Chengcan Wu et al., *Reliable Unlearning Harmful Information in LLMs with Metamorphosis Representation Projection*, arXiv:2508.15449, August 2025, accessed June 22, 2026, arXiv: 2508.15449 [cs.LG]

212 Lingyu Li, Yan Teng, and Yingchun Wang, *Mechanistic Origin of Moral Indifference in Language Models*, arXiv:2603.15615, March 2026, accessed June 22, 2026, arXiv: 2603.15615 [cs.CL]

213 Shanghai AI Lab et al., *Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report*, arXiv:2507.16534, July 2025, accessed June 22, 2026, arXiv: 2507.16534 [cs.AI]; Dongrui Liu et al., *Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report v1.5*, arXiv:2602.14457, February 2026, accessed June 22, 2026, arXiv: 2602.14457 [cs.AI]

214 Haibo Tong et al., *ForesightSafety Bench: A Frontier Risk Evaluation and Governance Framework towards Safe AI*, arXiv:2602.14135, February 2026, accessed June 22, 2026, arXiv: 2602.14135 [cs.AI]

215 Bo Zhang et al., *DeepSight: An All-in-One LM Safety Toolkit*, arXiv:2602.12092, February 2026, accessed June 22, 2026, arXiv: 2602.12092 [cs.CL]

216 Xiangyang Zhu et al., *SafeSci: Safety Evaluation of Large Language Models in Science Domains and Beyond*, arXiv:2603.01589, April 2026, accessed June 22, 2026, arXiv: 2603.01589 [cs.LG]

217 Yutao Wu et al., *Internal Safety Collapse in Frontier Large Language Models*, arXiv:2603.23509, March 2026, accessed June 22, 2026, arXiv: 2603.23509 [cs.CL]

218 Shiyi Yang et al., *AirLLM: Diffusion Policy-based Adaptive LoRA for Remote Fine-Tuning of LLM over the Air*, arXiv:2507.11515, July 2025, accessed June 22, 2026, arXiv: 2507.11515 [cs.LG]

219 Jigang Fan et al., *SafeProtein: Red-Teaming Framework and Benchmark for Protein Foundation Models*, arXiv:2509.03487, October 2025, accessed June 22, 2026, arXiv: 2509.03487 [cs.LG]

220 Xin Wang et al., *LQA: A Lightweight Quantized-Adaptive Framework for Vision-Language Models on the Edge*, arXiv:2602.07849, February 2026, accessed June 22, 2026, arXiv: 2602.07849 [cs.AI]; Qianpu Sun et al., *LABSHIELD: A Multimodal Benchmark for Safety-Critical Reasoning and Planning in Scientific Laboratories*, arXiv:2603.11987, March 2026, accessed June 22, 2026, arXiv: 2603.11987 [cs.AI]

221 *Xue Lan: AI Governance from a Global Perspective——Challenges, Mechanisms, and Future Paths (薛澜：全球视野下的人工智能治理——挑战、机制与未来路径)*, https://ciss.tsinghua.edu.cn/info/zlyaq/8523, Center For International Security and Strategy Tsinghua University, August 2025, accessed June 23, 2026

222 *Zheng Nanning: "The Intelligence Leap: From Model-Driven to Intent-Driven" (郑南宁：《智能跃迁：从模型驱动到意图驱动》)*, https://mp.weixin.qq.com/s/oMI52Q939-RbSDvvhceUfw, Weixin Official Accounts Platform, July 2025, accessed June 23, 2026

223 *Academician Yao Qizhi: AI Deception Triggers "Existential" Risks, Establishing a Large Model Evaluation System Brooks No Delay (姚期智院士：AI 欺骗引发"生存性"风险，建立大模型评估系统刻不容缓)*, https://www.tsinghua.edu.cn/info/1182/119822.htm, Tsinghua University, June 2025, accessed June 23, 2026

224 *TC260-005 Ethical Security Guidelines for Artificial Intelligence Applications 1.0 (《人工智能应用伦理安全指引 1.0》)*, https://www.tc260.org.cn/portal/article/2/6aee9380ac44434d994eb6990bd92997, National Technical Committee 260 on Cybersecurity of Standardization Administration of China, May 2026, accessed June 22, 2026

225 *The Singapore Consensus on Global AI Safety Research Priorities*, https://aisafetypriorities.org, May 2025, accessed June 23, 2026

226 Institute for AI International Governance, Tsinghua University, *Exclusive Release | 2025 Annual Report on AI Governance: Toward "Measurable" AI Governance (独家发布 | 2025 人工智能治理年度报告：迈向"可衡量"的 AI 治理)*, https://mp.weixin.qq.com/s/gm_d7d-sxcVWiXFWmaS4lg, Weixin Official Accounts Platform, December 2025, accessed June 23, 2026

227 *"Artificial Intelligence: The Endless Frontier"—Transcript of Zhang Yaqin's Speech at the Humanities Tsinghua Forum-Institute for AI Industry Research, Tsinghua University (《人工智能：无尽的前沿》——人文清华讲坛张亚勤演讲实录-清华大学智能产业研究院)*, https://air.tsinghua.edu.cn/info/1007/2479.htm, Tsinghua University, December 2025, accessed June 23, 2026

228 *ATxSummit 2026 Video on Demand: Building AI's Guardrails: A House of Trust*, https://atxsummit.asiatechxsg.com/resources/video-on-demand/S_SESSION2F_ROWINENSL, ATxSummit 2026 Singapore, minute 37, accessed June 23, 2026

229 *Top Ten Topics for AI Rule-of-Law Research in 2026 Released (2026 年人工智能法治研究十大议题发布)*, https://mp.weixin.qq.com/s/LnlNoHXCnUV-AJ2WIUsgyg, Weixin Official Accounts Platform, accessed June 23, 2026; *Media Report | People's Daily Online: "The Top Ten Topics in Artificial Intelligence Rule-of-Law Research (2026)" Released (媒体报道 | 人民网：《人工智能法治研究十大议题（2026）》发布)*, https://mp.weixin.qq.com/s/XPsOt3trsbLCEhJiwDAG4g, Weixin Official Accounts Platform, March 2026, accessed June 23, 2026

230 AI 善治学术工作组, *Research Report on the Development and Governance of AI Agents (智能体发展与治理研究报告)*, https://www.cdut.edu.cn/__local/8/44/9C/7B73AA9FAE543F4B38F30BB086A_3071DEB8_139531.pdf, October 2025

231 Avijit Ghosh et al., *State of Open Source on Hugging Face: Spring 2026*, https://huggingface.co/blog/huggingface/state-of-os-hf-spring-2026, Hugging Face, April 2026, accessed June 23, 2026

232 Zhang Xiaojun Podcast, *Kimi Founder Yang Zhilin: K2, Agentic LLMs, Brains in Vats, and the Beginning of Infinity*, August 2025, accessed June 23, 2026

233 Wang Zhe (王哲) and Xue Lan (薛澜), *The Open-Source Innovation Commons of Large Models: Historical Evolution, Value Logic, and the China Narrative (大模型开源创新公地：历史演进、价值逻辑与中国叙事)*, https://ciss.tsinghua.edu.cn/info/zlyaq/8618, September 2025, accessed June 23, 2026

234 Wang Zhe (王哲) and Xue Lan (薛澜), *The Open-Source Innovation Commons of Large Models: Historical Evolution, Value Logic, and the China Narrative (大模型开源创新公地：历史演进、价值逻辑与中国叙事)*, https://ciss.tsinghua.edu.cn/info/zlyaq/8618, September 2025, accessed June 23, 2026; Fu Hongyu (傅宏宇) and Peng Jingzhi (彭靖芷), *Global Practices and Path Choices for Open-Source AI Governance (开源人工智能治理的全球实践及路径选择)*, https://www.secrss.com/articles/80066, 中国信息安全, June 2025, accessed June 23, 2026

235 Fu Hongyu (傅宏宇) and Peng Jingzhi (彭靖芷), *Global Practices and Path Choices for Open-Source AI Governance (开源人工智能治理的全球实践及路径选择)*, https://www.secrss.com/articles/80066, 中国信息安全, June 2025, accessed June 23, 2026

236 Yingjie Ye and Chuan Li, *From Closed Source to Open Source: Risk Evolution and Governance Paradigm Transformation in the Technological Transition of Large AI Models (从闭源到开源：技术跃迁下大模型风险演化与治理范式转型)*, https://www.kjjb.org/CN/10.6049/kjjbydc.D92025080179, 2026, accessed June 23, 2026

237 Liao Huijiao (廖慧姣) and Zhang Taolue (张韬略), *China's Regulatory Approach in the Global Open-Source AI Competition: Drawing on Experience and Independent Construction (全球开源人工智能博弈中的中国监管方案：经验借鉴与自主构建)*, https://mp.weixin.qq.com/s/4iq0nfkk7hF6ur781eTAkw, Weixin Official Accounts Platform, March 2026, accessed June 23, 2026

238 Weiwei Xu et al., *A Governance Horizon for Ethical-Use Constraints in Open-Weight AI Models*, arXiv:2605.24383, May 2026, accessed June 23, 2026, arXiv: 2605.24383 [cs.AI]

239 Fu Hongyu (傅宏宇) and Peng Jingzhi (彭靖芷), *Global Practices and Path Choices for Open-Source AI Governance (开源人工智能治理的全球实践及路径选择)*, https://www.secrss.com/articles/80066, 中国信息安全, June 2025, accessed June 23, 2026

240 Yingjie Ye and Chuan Li, *From Closed Source to Open Source: Risk Evolution and Governance Paradigm Transformation in the Technological Transition of Large AI Models (从闭源到开源：技术跃迁下大模型风险演化与治理范式转型)*, https://www.kjjb.org/CN/10.6049/kjjbydc.D92025080179, 2026, accessed June 23, 2026

241 Wang Zhe (王哲) and Xue Lan (薛澜), *The Open-Source Innovation Commons of Large Models: Historical Evolution, Value Logic, and the China Narrative (大模型开源创新公地：历史演进、价值逻辑与中国叙事)*, https://ciss.tsinghua.edu.cn/info/zlyaq/8618, September 2025, accessed June 23, 2026

242 Liao Huijiao (廖慧姣) and Zhang Taolue (张韬略), *China's Regulatory Approach in the Global Open-Source AI Competition: Drawing on Experience and Independent Construction (全球开源人工智能博弈中的中国监管方案：经验借鉴与自主构建)*, https://mp.weixin.qq.com/s/4iq0nfkk7hF6ur781eTAkw, Weixin Official Accounts Platform, March 2026, accessed June 23, 2026

243 Yingjie Ye and Chuan Li, *From Closed Source to Open Source: Risk Evolution and Governance Paradigm Transformation in the Technological Transition of Large AI Models (从闭源到开源：技术跃迁下大模型风险演化与治理范式转型)*, https://www.kjjb.org/CN/10.6049/kjjbydc.D92025080179, 2026, accessed June 23, 2026

244 Han Honggui (韩红桂), *The AI Era Calls for a New Paradigm of Cybersecurity (人工智能时代呼唤网络安全新范式)*, https://www.studytimes.cn/kjqy/202509/t20250926_83292.html, Study Times (学习时报), September 2025, accessed June 23, 2026

245 China Academy of Information and Communications Technology, *CAICT's AI Research Institute Releases "AI Safety Research Report——Security Risk Identification and Response from a Technical Perspective (2025)" (中国信通院人工智能研究所发布《人工智能安全研究报告——技术视角下的安全风险梳理与应对（2025）》)*, https://mp.weixin.qq.com/s/5h8bx4kHRkTU10-DvsSfqQ, Weixin Official Accounts Platform, November 2025, accessed June 23, 2026; China Academy of Information and Communications Technology, *AI Safety Benchmark Large Model Safety Benchmark Test—Release of Results of the Special Safety Test for Code Large Models (AI Safety Benchmark 大模型安全基准测试——代码大模型安全专项测试结果发布)*, https://mp.weixin.qq.com/s/wbDiepubwic3P8QoWE2TNA, Weixin Official Accounts Platform, July 2025, accessed June 23, 2026

246 *CCCF June 2025 Issue Released—Special Topic on "Large Models Empowering Cybersecurity" (CCCF 2025 年 6 月刊发布——"大模型赋能网络安全" 专题)*, https://www.ccf.org.cn/Focus/2025-06-20/845799.shtml, China Computer Federation, June 2025, accessed June 23, 2026

247 Anthropic, *Project Glasswing: Securing Critical Software for the AI Era*, https://www.anthropic.com/glasswing, April 2026, accessed June 19, 2026

248 *Anthropic's "Crying Wolf" Sparks Wall Street Panic! A 27-Year-Old Vulnerability, Mythos Instantly Defeated by 8 AIs - 36Kr (Anthropic 版「狼来了」引华尔街恐慌！27 年漏洞，Mythos 被 8 个 AI 秒杀-36 氪)*, https://36kr.com/p/3763537087562249, 36Kr, April 2026, accessed June 23, 2026

249 Founder Park, *Mythos: The Era in Which Ordinary People Can Freely Use Flagship AI May Be Coming to an End (Mythos：普通人能自由使用旗舰 AI 的时代，可能要结束了)*, https://mp.weixin.qq.com/s/CefMuAusEBC66tt28lLFrA?poc_token=HA0uOmqjYcrL3ylNQQd6rEB44tSfTkpu5-0pCBnC, Weixin Official Accounts Platform, April 2026, accessed June 23, 2026

250 *Yang Yanqing: The Mythos Model Triggers Human Risks and a Reconstruction of Global Governance (杨燕青：Mythos 模型引发人类风险与全球治理重构)*, https://www.yicai.com/news/103155234.html, Yicai (第一财经), April 2026, accessed June 23, 2026

251 *Vulnerabilities Can Never Be Fully Patched, "Protection with Holes" Becomes Inevitable——QiAnXin Releases Mythos Response White Paper (漏洞永远修不完，"带洞防护" 成必然——奇安信发布 Mythos 应对白皮书)*, https://www.qianxin.com/news/detail?news_id=14759, Qi'an Xin (奇安信), April 2026, accessed June 23, 2026

252 Concordia AI and Center for Biosafety Research and Strategy of Tianjin University, *Responsible Innovation in AI x Life Sciences (人工智能 x 生命科学的负责任创新)*, https://concordia-ai.com/research/responsible-innovation-in-ai-x-life-sciences/, July 2025, accessed June 23, 2026

253 *Center for Biosafety Research and Strategy of Tianjin University Deeply Participates in the "2025 World Artificial Intelligence Conference" Discussion on Artificial Intelligence Safety and Governance (天津大学生物安全战略研究中心深度参与 "2025 世界人工智能大会" 人工智能安全与治理讨论)*, https://tjusa.tju.edu.cn/info/1093/2014.htm, Center for Biosafety Research and Strategy of Tianjin University (天津大学生物安全战略研究中心), August 2025, accessed June 23, 2026

254 *Academic Conference | Closed-Door Biosafety Symposium Successfully Held at the National Key Laboratory of Synthetic Biology Technology (学术会议 | 生物安全闭门研讨会在合成生物技术全国重点实验室圆满举行)*, https://mp.weixin.qq.com/s/htJzJb-w0CTiC0QkVWKu-A, Weixin Official Accounts Platform, August 2025, accessed June 23, 2026

255 Concordia AI, *Researched and Written by Nearly 300 Academicians and Experts, "Artificial Intelligence Empowering Scientific Research: The AI Disciplinary System" Officially Published, with Concordia AI Contributing the "AI + Life Sciences" Chapter (近 300 位院士专家共同研究撰写，《人工智能赋能科学研究：人工智能学科体系》正式出版，安远 AI 参编 "AI+ 生命科学" 章节)*, https://mp.weixin.qq.com/s/ImVraHA4jmu4IlbGOW3pCA, Weixin Official Accounts Platform, May 2026, accessed June 23, 2026

256 Huang Ning (黄宁) and Zhu Shu (朱姝), *Risk Assessment of Artificial Intelligence Applications in the Biological Field (生物领域人工智能应用引发的风险研判)*, https://mp.weixin.qq.com/s/XUorm_WT64LVXTt84c8m9A, Weixin Official Accounts Platform, May 2025, accessed June 23, 2026

257 Zhang Puhua (张埔华), "Biosafety Protection: China's Path of Technological Empowerment and Risk Governance (生物安全防护：科技赋能与风险治理的中国路径)," *吉首大学学报 (社会科学版)* 46, no. 6 (2025)

258 Xiao Xi (肖晞) and Liu Kunye (刘坤烨), *Global Biosecurity Governance in the Age of Artificial Intelligence: Opportunities and Challenges (人工智能时代的全球生物安全治理：机遇与挑战)*, https://mp.weixin.qq.com/s/pvymmi-fJZGLEgay7JGVew, Weixin Official Accounts Platform, November 2025, accessed June 23, 2026

259 The Nuclear Threat Initiative, *Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences*, 2025, accessed June 26, 2026, https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/

260 *RECOMMENDATIONS TO GOVERNMENTS ON MITIGATING AIxBIO RISKS*, https://unodaweb-meetings.unoda.org/public/2025-12/AIxBio%20Recommendations%20INHR.pdf, INHR, December 2025

261 International Dialogues on AI Safety, *IDAIS-London, 2026: Statement*, London, UK, April 17–19, 2026, 2026, accessed June 26, 2026, https://idais.ai/dialogue/idais-london/

262 Frontier Model Forum, *About*, 2026, accessed June 26, 2026, https://www.frontiermodelforum.org/about-us/

263 *AIIA - Safety Governance Working Group (安全治理工作组)*, https://aihub.caict.ac.cn/aiia_wgs/36, AI Industry Alliance of China, 2026, accessed June 23, 2026

264 China Academy of Information and Communications Technology, *AI Safety Benchmark: First-Round Results of the Authoritative Large Model Safety Benchmark Test Officially Released (AI Safety Benchmark 权威大模型安全基准测试首轮结果正式发布)*, https://mp.weixin.qq.com/s/3FcLBHCy_oVaaj-2Ca9zag, Weixin Official Accounts Platform, April 2024, accessed June 23, 2026

265 China Academy of Information and Communications Technology, *AI Safety Benchmark: First-Round Results of the Authoritative Large Model Safety Benchmark Test Officially Released (AI Safety Benchmark 权威大模型安全基准测试首轮结果正式发布)*, https://mp.weixin.qq.com/s/3FcLBHCy_oVaaj-2Ca9zag, Weixin Official Accounts Platform, April 2024, accessed June 23, 2026

266 China Academy of Information and Communications Technology, *AI Safety Benchmark 2.0 Testing System and Supporting Test Platform Released (AI Safety Benchmark 2.0 测试体系及配套测试平台发布)*, https://mp.weixin.qq.com/s/tzfCOF3nKYFvS9zozoXDww, Weixin Official Accounts Platform, February 2026, accessed June 23, 2026

267 China Academy of Information and Communications Technology, *CAICT's AI Research Institute Releases "AI Safety Research Report——Security Risk Identification and Response from a Technical Perspective (2025)" (中国信通院人工智能研究所发布《人工智能安全研究报告——技术视角下的安全风险梳理与应对（2025）》)*, https://mp.weixin.qq.com/s/5h8bx4kHRkTU10-DvsSfqQ, Weixin Official Accounts Platform, November 2025, accessed June 23, 2026; China Academy of Information and Communications Technology, *AI Safety Benchmark Large Model Safety Benchmark Test—Release of Results of the Special Safety Test for Code Large Models (AI Safety Benchmark 大模型安全基准测试——代码大模型安全专项测试结果发布)*, https://mp.weixin.qq.com/s/wbDiepubwic3P8QoWE2TNA, Weixin Official Accounts Platform, July 2025, accessed June 23, 2026

268 China Academy of Information and Communications Technology, *AI Safety Benchmark Agent Safety Testing 2026 Q1 Edition Results Released (AI Safety Benchmark 智能体安全测试 2026 Q1 版结果发布)*, https://mp.weixin.qq.com/s/7x1PMuqmqNmqekgZI5XZYQ, Weixin Official Accounts Platform, March 2026, accessed June 23, 2026

269 China Academy of Information and Communications Technology, *AI Safety Benchmark 2.0 Testing System and Supporting Test Platform Released (AI Safety Benchmark 2.0 测试体系及配套测试平台发布)*, https://mp.weixin.qq.com/s/tzfCOF3nKYFvS9zozoXDww, Weixin Official Accounts Platform, February 2026, accessed June 23, 2026

270 AI Industry Alliance of China, *Safeguarding AI Security and Jointly Building a Model of Industry Self-Discipline: First Batch of 17 Enterprises Sign the "AI Safety Commitment" (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》)*, https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng, Weixin Official Accounts Platform, December 2024, accessed June 23, 2026; China Academy of Information and Communications Technology, *WAIC Releases the China Artificial Intelligence Safety Commitment Framework (WAIC 发布《中国人工智能安全承诺框架》)*, https://mp.weixin.qq.com/s/jwtALdsfXLMN_ohfsOF5Dw, Weixin Official Accounts Platform, July 2025, accessed June 23, 2026

271 *Disclosure of Practices on the Artificial Intelligence Security and Safety Commitments (《人工智能安全承诺》实践披露)*, https://aihub.caict.ac.cn/ai_security_and_safety_commitments, AI Industry Alliance of China, July 2025, accessed June 23, 2026

272 *Frontier AI Safety Commitments, AI Seoul Summit 2024*, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024, UK Department for Science, Innovation & Technology, 2024, accessed June 23, 2026; METR, *Common Elements of Frontier AI Safety Policies (December 2025 Update)*, 2025, accessed June 26, 2026, https://metr.org/blog/2025-12-09-common-elements-of-frontier-ai-safety-policies/

273 AI Industry Alliance of China, *"AI Safety Commitment: Agent Special Initiative" Officially Signed (《人工智能安全承诺：智能体专项》正式签署)*, https://mp.weixin.qq.com/s/qeELmpD4MHJ7N1Oiu1LCBw, Weixin Official Accounts Platform, February 2026, accessed June 23, 2026

274 AI Industry Alliance of China, *"Self-Discipline Convention on Network and Data Security for Cloud-Based Agent Services (2026 Edition)" Officially Released (《云上智能体服务网络和数据安全自律公约 (2026 版)》正式发布)*, https://mp.weixin.qq.com/s/JchdvwyZppvGkaWssEGSkw, Weixin Official Accounts Platform, April 2026, accessed June 23, 2026

275 AI Industry Alliance of China, *China Academy of Information and Communications Technology and Tencent Cloud Jointly Release "AI Agent Security Practice Guidelines" (中国信通院联合腾讯云发布《AI Agent 安全实践指引》)*, https://mp.weixin.qq.com/s/sKNhIU3Qh-UWX6VOQ9IeTw, Weixin Official Accounts Platform, March 2026, accessed June 23, 2026; *China Artificial Intelligence Industry Alliance Releases "Risk Management Guidelines for the Deployment of OpenClaw-Type Agents" (中国人工智能产业发展联盟发布《OpenClaw 类智能体部署风险管理指南》)*, https://mp.weixin.qq.com/s/YETi1BlM2IQtTzr3TGU_uw, Weixin Official Accounts Platform, April 2026, accessed June 23, 2026

276 Alibaba Cloud, *What Is Agent Identity (什么是智能体身份)*, https://help.aliyun.com/zh/agentidentity/what-is-agent-identity, January 2026, accessed June 23, 2026; Alibaba, *Open Agent Auth: Enterprise-Grade AI Agent Operation Authorization Framework*, 2026, accessed June 26, 2026, https://github.com/alibaba/open-agent-auth

277 *On the Eve of the Explosion of Hundreds of Billions of AI Agents, Who Will Protect Our AI Security? (千亿智能体爆发前夜，谁来保护我们的 AI 安全？)*, https://www.jazzyear.com/article_info.html?id=1642, Jiazi Guangnian (甲子光年), December 2025, accessed June 23, 2026

278 *Agent Security (智能体安全)*, https://developer.huawei.com/consumer/cn/doc/service/agent-security-0000002437625978, 华为 HarmonyOS 开发者, June 2026, accessed June 23, 2026

279 Tencent Cloud, *CubeSandbox*, 2026, accessed June 26, 2026, https://github.com/TencentCloud/CubeSandbox; Tencent, *AI-Infra-Guard*, 2026, accessed June 26, 2026, https://github.com/Tencent/AI-Infra-Guard

280 *Knowledge Atlas Technology Joint Stock Company Limited Global Offerings*, https://www1.hkexnews.hk/listedco/listconews/sehk/2025/1230/2025123000017.pdf, Hong Kong Stock Exchange, December 2025, accessed June 23, 2026; *MiniMax Group Inc. Global Offering*, https://www1.hkexnews.hk/listedco/listconews/sehk/2025/1231/2025123100025.pdf, Hong Kong Stock Exchange, December 2025, accessed June 23, 2026

281 】赛博禅心, *Weixin Official Accounts Platform*, https://mp.weixin.qq.com/s/bfTbsnItuyr_k-w3XlejOg?scene=1&click_id=65, accessed June 23, 2026

282 *Xue Hui: Addressing AI Compound Risks with the "Security by Design" Concept (薛晖：以"设计即安全"理念应对 AI 复合型风险)*, https://wlaq.gmw.cn/2025-09/19/content_38297677.htm, 光明网, September 2025, accessed June 23, 2026

283 Yue Zhu et al., "China's Emerging Regulation toward an Open Future for AI," *Science* 390, no. 6769 (October 2025): 132–135, accessed June 23, 2026, https://doi.org/10.1126/science.ady7922

284 *LLM Leaderboard*, https://arena.ai/leaderboard/text, Arena, accessed June 23, 2026

285 *Artificial Analysis*, https://artificialanalysis.ai, 2026, accessed June 23, 2026

286 Daya Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645, no. 8081 (September 2025): 633–638, ISSN: 1476-4687, accessed June 23, 2026, https://doi.org/10.1038/s41586-025-09422-z

287 Daya Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645, no. 8081 (September 2025): 633–638, ISSN: 1476-4687, accessed June 23, 2026, https://doi.org/10.1038/s41586-025-09422-z

288 *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*, https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro, Hugging Face, April 2026, accessed June 23, 2026; Kimi Team et al., *Kimi K2.5: Visual Agentic Intelligence*, arXiv:2602.02276, February 2026, accessed June 23, 2026, arXiv: 2602.02276 [cs.CL]

289 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

290 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

291 ERNIE Team, Baidu, "ERNIE 4.5 Technical Report," 2025, accessed June 26, 2026, https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

292 Haifeng Wang et al., *ERNIE 5.0 Technical Report*, https://arxiv.org/abs/2602.04705v1, arXiv.org, February 2026, accessed June 23, 2026

293 Bytedance Seed, *Seed1.8 Model Card: Towards Generalized Real-World Agency*, arXiv:2603.20633, April 2026, accessed June 23, 2026, arXiv: 2603.20633 [cs.AI]

294 ByteDance, *Seed2.0*, 2026, accessed June 26, 2026, https://github.com/ByteDance-Seed/Seed2.0

295 Bytedance, *Seed-OSS Open-Source Models*, https://huggingface.co/ByteDance-Seed/Seed-OSS-36B-Instruct, Hugging Face, August 2025, accessed June 23, 2026

296 Daya Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645, no. 8081 (September 2025): 633–638, ISSN: 1476-4687, accessed June 23, 2026, https://doi.org/10.1038/s41586-025-09422-z

297 DeepSeek-AI et al., *DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models*, https://arxiv.org/abs/2512.02556v1, arXiv.org, December 2025, accessed June 23, 2026

298 *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*, https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro, Hugging Face, April 2026, accessed June 23, 2026

299 Meituan LongCat Team et al., *LongCat-Flash Technical Report*, arXiv:2509.01322, September 2025, accessed June 23, 2026, arXiv: 2509.01322 [cs.CL]

300 Meituan LongCat Team et al., *Introducing LongCat-Flash-Thinking: A Technical Report*, arXiv:2509.18883, November 2025, accessed June 23, 2026, arXiv: 2509.18883 [cs.AI]

301 MiniMax et al., *MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention*, arXiv:2506.13585, June 2025, accessed June 23, 2026, arXiv: 2506.13585 [cs.CL]

302 *Meet MiniMax-M2*, https://huggingface.co/MiniMaxAI/MiniMax-M2, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2 & Agent: Ingenious in Simplicity*, https://www.minimax.io/news/minimax-m2, October 2025, accessed June 23, 2026

303 *MiniMax-M2.5*, https://huggingface.co/MiniMaxAI/MiniMax-M2.5, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.5: Built for Real-World Productivity.*, https://www.minimax.io/news/minimax-m25, MiniMax, February 2026, accessed June 23, 2026

304 *MiniMax-M2.7*, https://huggingface.co/MiniMaxAI/MiniMax-M2.7, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.7: Early Echoes of Self-Evolution*, https://www.minimax.io/news/minimax-m27-en, MiniMax, March 2026, accessed June 23, 2026

305 Kimi Team et al., *Kimi K2: Open Agentic Intelligence*, arXiv:2507.20534, February 2026, accessed June 23, 2026, arXiv: 2507.20534 [cs.LG]

306 Kimi Team et al., *Kimi K2.5: Visual Agentic Intelligence*, arXiv:2602.02276, February 2026, accessed June 23, 2026, arXiv: 2602.02276 [cs.CL]

307 *Kimi-K2.6*, https://huggingface.co/moonshotai/Kimi-K2.6, Hugging Face, June 2026, accessed June 23, 2026

308 Tencent, *Hunyuan A13B Technical Report*, https://github.com/Tencent-Hunyuan/Hunyuan-A13B/blob/main/report/Hunyuan_A13B_Technical_Report.pdf, GitHub, June 2025, accessed June 23, 2026

309 *Hy3-Preview*, https://huggingface.co/tencent/Hy3-preview, Hugging Face, April 2026, accessed June 23, 2026

310 Xiaomi LLM-Core Team et al., *MiMo-V2-Flash Technical Report*, arXiv:2601.02780, January 2026, accessed June 23, 2026, arXiv: 2601.02780 [cs.CL]

311 Xiaomi, *MiMo-V2.5-Pro*, https://mimo.xiaomi.com/mimo-v2-5-pro, April 2026, accessed June 23, 2026

312 GLM-4 5 Team et al., *GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models*, arXiv:2508.06471, August 2025, accessed June 23, 2026, arXiv: 2508.06471 [cs.CL]

313 GLM-5-Team et al., *GLM-5: From Vibe Coding to Agentic Engineering*, arXiv:2602.15763, February 2026, accessed June 23, 2026, arXiv: 2602.15763 [cs.LG]

289 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

290 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

291 ERNIE Team, Baidu, "ERNIE 4.5 Technical Report," 2025, accessed June 26, 2026, https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

292 Haifeng Wang et al., *ERNIE 5.0 Technical Report*, https://arxiv.org/abs/2602.04705v1, arXiv.org, February 2026, accessed June 23, 2026

293 Bytedance Seed, *Seed1.8 Model Card: Towards Generalized Real-World Agency*, arXiv:2603.20633, April 2026, accessed June 23, 2026, arXiv: 2603.20633 [cs.AI]

294 ByteDance, *Seed2.0*, 2026, accessed June 26, 2026, https://github.com/ByteDance-Seed/Seed2.0

295 Bytedance, *Seed-OSS Open-Source Models*, https://huggingface.co/ByteDance-Seed/Seed-OSS-36B-Instruct, Hugging Face, August 2025, accessed June 23, 2026

296 Daya Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645, no. 8081 (September 2025): 633–638, ISSN: 1476-4687, accessed June 23, 2026, https://doi.org/10.1038/s41586-025-09422-z

297 DeepSeek-AI et al., *DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models*, https://arxiv.org/abs/2512.02556v1, arXiv.org, December 2025, accessed June 23, 2026

298 *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*, https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro, Hugging Face, April 2026, accessed June 23, 2026

299 Meituan LongCat Team et al., *LongCat-Flash Technical Report*, arXiv:2509.01322, September 2025, accessed June 23, 2026, arXiv: 2509.01322 [cs.CL]

300 Meituan LongCat Team et al., *Introducing LongCat-Flash-Thinking: A Technical Report*, arXiv:2509.18883, November 2025, accessed June 23, 2026, arXiv: 2509.18883 [cs.AI]

301 MiniMax et al., *MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention*, arXiv:2506.13585, June 2025, accessed June 23, 2026, arXiv: 2506.13585 [cs.CL]

302 *Meet MiniMax-M2*, https://huggingface.co/MiniMaxAI/MiniMax-M2, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2 & Agent: Ingenious in Simplicity*, https://www.minimax.io/news/minimax-m2, October 2025, accessed June 23, 2026

303 *MiniMax-M2.5*, https://huggingface.co/MiniMaxAI/MiniMax-M2.5, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.5: Built for Real-World Productivity.*, https://www.minimax.io/news/minimax-m25, MiniMax, February 2026, accessed June 23, 2026

304 *MiniMax-M2.7*, https://huggingface.co/MiniMaxAI/MiniMax-M2.7, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.7: Early Echoes of Self-Evolution*, https://www.minimax.io/news/minimax-m27-en, MiniMax, March 2026, accessed June 23, 2026

305 Kimi Team et al., *Kimi K2: Open Agentic Intelligence*, arXiv:2507.20534, February 2026, accessed June 23, 2026, arXiv: 2507.20534 [cs.LG]

306 Kimi Team et al., *Kimi K2.5: Visual Agentic Intelligence*, arXiv:2602.02276, February 2026, accessed June 23, 2026, arXiv: 2602.02276 [cs.CL]

307 *Kimi-K2.6*, https://huggingface.co/moonshotai/Kimi-K2.6, Hugging Face, June 2026, accessed June 23, 2026

308 Tencent, *Hunyuan A13B Technical Report*, https://github.com/Tencent-Hunyuan/Hunyuan-A13B/blob/main/report/Hunyuan_A13B_Technical_Report.pdf, GitHub, June 2025, accessed June 23, 2026

309 *Hy3-Preview*, https://huggingface.co/tencent/Hy3-preview, Hugging Face, April 2026, accessed June 23, 2026

310 Xiaomi LLM-Core Team et al., *MiMo-V2-Flash Technical Report*, arXiv:2601.02780, January 2026, accessed June 23, 2026, arXiv: 2601.02780 [cs.CL]

311 Xiaomi, *MiMo-V2.5-Pro*, https://mimo.xiaomi.com/mimo-v2-5-pro, April 2026, accessed June 23, 2026

312 GLM-4 5 Team et al., *GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models*, arXiv:2508.06471, August 2025, accessed June 23, 2026, arXiv: 2508.06471 [cs.CL]

313 GLM-5-Team et al., *GLM-5: From Vibe Coding to Agentic Engineering*, arXiv:2602.15763, February 2026, accessed June 23, 2026, arXiv: 2602.15763 [cs.LG]

289 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

290 QwenTeam, *Qwen3-Max: Just Scale It*, https://qwen.ai/blog?id=qwen3-max, September 2025, accessed June 23, 2026

291 ERNIE Team, Baidu, "ERNIE 4.5 Technical Report," 2025, accessed June 26, 2026, https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

292 Haifeng Wang et al., *ERNIE 5.0 Technical Report*, https://arxiv.org/abs/2602.04705v1, arXiv.org, February 2026, accessed June 23, 2026

293 Bytedance Seed, *Seed1.8 Model Card: Towards Generalized Real-World Agency*, arXiv:2603.20633, April 2026, accessed June 23, 2026, arXiv: 2603.20633 [cs.AI]

294 ByteDance, *Seed2.0*, 2026, accessed June 26, 2026, https://github.com/ByteDance-Seed/Seed2.0

295 Bytedance, *Seed-OSS Open-Source Models*, https://huggingface.co/ByteDance-Seed/Seed-OSS-36B-Instruct, Hugging Face, August 2025, accessed June 23, 2026

296 Daya Guo et al., "DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning," *Nature* 645, no. 8081 (September 2025): 633–638, ISSN: 1476-4687, accessed June 23, 2026, https://doi.org/10.1038/s41586-025-09422-z

297 DeepSeek-AI et al., *DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models*, https://arxiv.org/abs/2512.02556v1, arXiv.org, December 2025, accessed June 23, 2026

298 *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*, https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro, Hugging Face, April 2026, accessed June 23, 2026

299 Meituan LongCat Team et al., *LongCat-Flash Technical Report*, arXiv:2509.01322, September 2025, accessed June 23, 2026, arXiv: 2509.01322 [cs.CL]

300 Meituan LongCat Team et al., *Introducing LongCat-Flash-Thinking: A Technical Report*, arXiv:2509.18883, November 2025, accessed June 23, 2026, arXiv: 2509.18883 [cs.AI]

301 MiniMax et al., *MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention*, arXiv:2506.13585, June 2025, accessed June 23, 2026, arXiv: 2506.13585 [cs.CL]

302 *Meet MiniMax-M2*, https://huggingface.co/MiniMaxAI/MiniMax-M2, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2 & Agent: Ingenious in Simplicity*, https://www.minimax.io/news/minimax-m2, October 2025, accessed June 23, 2026

303 *MiniMax-M2.5*, https://huggingface.co/MiniMaxAI/MiniMax-M2.5, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.5: Built for Real-World Productivity.*, https://www.minimax.io/news/minimax-m25, MiniMax, February 2026, accessed June 23, 2026

304 *MiniMax-M2.7*, https://huggingface.co/MiniMaxAI/MiniMax-M2.7, Hugging Face, May 2026, accessed June 23, 2026; MiniMax, *MiniMax M2.7: Early Echoes of Self-Evolution*, https://www.minimax.io/news/minimax-m27-en, MiniMax, March 2026, accessed June 23, 2026

305 Kimi Team et al., *Kimi K2: Open Agentic Intelligence*, arXiv:2507.20534, February 2026, accessed June 23, 2026, arXiv: 2507.20534 [cs.LG]

306 Kimi Team et al., *Kimi K2.5: Visual Agentic Intelligence*, arXiv:2602.02276, February 2026, accessed June 23, 2026, arXiv: 2602.02276 [cs.CL]

307 *Kimi-K2.6*, https://huggingface.co/moonshotai/Kimi-K2.6, Hugging Face, June 2026, accessed June 23, 2026

308 Tencent, *Hunyuan A13B Technical Report*, https://github.com/Tencent-Hunyuan/Hunyuan-A13B/blob/main/report/Hunyuan_A13B_Technical_Report.pdf, GitHub, June 2025, accessed June 23, 2026

309 *Hy3-Preview*, https://huggingface.co/tencent/Hy3-preview, Hugging Face, April 2026, accessed June 23, 2026

310 Xiaomi LLM-Core Team et al., *MiMo-V2-Flash Technical Report*, arXiv:2601.02780, January 2026, accessed June 23, 2026, arXiv: 2601.02780 [cs.CL]

311 Xiaomi, *MiMo-V2.5-Pro*, https://mimo.xiaomi.com/mimo-v2-5-pro, April 2026, accessed June 23, 2026

312 GLM-4 5 Team et al., *GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models*, arXiv:2508.06471, August 2025, accessed June 23, 2026, arXiv: 2508.06471 [cs.CL]

313 GLM-5-Team et al., *GLM-5: From Vibe Coding to Agentic Engineering*, arXiv:2602.15763, February 2026, accessed June 23, 2026, arXiv: 2602.15763 [cs.LG]

314 Concordia AI, *Chinese Technical AI Safety Paper Database*, https://aisafetychina.com/research/, July 2026, accessed June 23, 2026

315 Haiquan Zhao et al., *Qwen3Guard Technical Report*, arXiv:2510.14276, October 2025, accessed June 23, 2026, arXiv: 2510.14276 [cs.CL]

316 *Qwen3Guard - a Qwen Collection*, https://huggingface.co/collections/Qwen/qwen3guard, Hugging Face, December 2025, accessed June 23, 2026

317 Jiayi Fu et al., *Reward Shaping to Mitigate Reward Hacking in RLHF*, arXiv:2502.18770, January 2026, accessed June 23, 2026, arXiv: 2502.18770 [cs.LG]; Jingyuan Ma et al., *HauntAttack: When Attack Follows Reasoning as a Shadow*, arXiv:2506.07031, October 2025, accessed June 23, 2026, arXiv: 2506.07031 [cs.CR]; Enyu Zhou et al., *Steering LLMs via Scalable Interactive Oversight*, arXiv:2602.04210, February 2026, accessed June 23, 2026, arXiv: 2602.04210 [cs.AI]

318 Yida Lu et al., *LongSafety: Evaluating Long-Context Safety of Large Language Models*, arXiv:2502.16971, February 2025, accessed June 23, 2026, arXiv: 2502.16971 [cs.CL]; Renmiao Chen et al., "JPS: Jailbreak Multimodal Large Language Models with Collaborative Visual Perturbation and Textual Steering," in *Proceedings of the 33rd ACM International Conference on Multimedia* (October 2025), 11756–11765, accessed June 23, 2026, arXiv: 2508.05087 [cs.MM]

319 Guanlin Li et al., *Picky LLMs and Unreliable RMs: An Empirical Study on Safety Alignment after Instruction Tuning*, arXiv:2502.01116, February 2025, accessed June 23, 2026, arXiv: 2502.01116 [cs.AI]; Jianshuo Dong et al., *SafeSearch: Automated Red-Teaming of LLM-Based Search Agents*, arXiv:2509.23694, June 2026, accessed June 23, 2026, arXiv: 2509.23694 [cs.AI]

320 杭州君同未来科技有限责任公司, *君合-生成式人工智能评测验证系统*, accessed June 26, 2026, https://gentel.io/zh/products/jc-scs

321 Wenpeng Xing et al., *Towards Robust and Secure Embodied AI: A Survey on Vulnerabilities and Attacks*, arXiv:2502.13175, February 2025, accessed June 23, 2026, arXiv: 2502.13175 [cs.CR]; Yi Zhou et al., *NeuRel-Attack: Neuron Relearning for Safety Disalignment in Large Language Models*, arXiv:2504.21053, April 2025, accessed June 23, 2026, arXiv: 2504.21053 [cs.LG]; Wenpeng Xing et al., *Latent Fusion Jailbreak: Blending Harmful and Harmless Representations to Elicit Unsafe LLM Outputs*, arXiv:2508.10029, January 2026, accessed June 23, 2026, arXiv: 2508.10029 [cs.CL]; Wenpeng Xing et al., *SproutBench: A Benchmark for Safe and Ethical Large Language Models for Youth*, arXiv:2508.11009, December 2025, accessed June 23, 2026, arXiv: 2508.11009 [cs.CL]; Dezhang Kong et al., *Web Fraud Attacks Against LLM-Driven Multi-Agent Systems*, arXiv:2509.01211, January 2026, accessed June 23, 2026, arXiv: 2509.01211 [cs.CR]; Wenpeng Xing et al., *Silencing the Guardrails: Inference-Time Jailbreaking via Dynamic Contextual Representation Ablation*, arXiv:2604.07835, April 2026, accessed June 23, 2026, arXiv: 2604.07835 [cs.AI]

322 RealAI (瑞莱智慧), *RealAI's Tian Tian: When AI Shifts from Answerer to Actor, Safety Risks Leap Accordingly (瑞莱智慧田天：当 AI 从回答者变成行动者，安全风险随之跃迁)*, https://mp.weixin.qq.com/s/ndEQSo4TLKeAacB58-RaXQ, Weixin Official Accounts Platform, June 2026, accessed June 23, 2026

323 RealAI (瑞莱智慧), *Building a Solid AI Security Foundation: RealAI's "RealGuard" Debuts at the Beijing Science and Technology Innovation Achievements Exhibition Area (筑牢 AI 安全基座，瑞莱智慧 "RealGuard" 亮相北京科技创新成果展区)*, https://mp.weixin.qq.com/s/hz3vK1uu-rJHYqU5AReWXg, Weixin Official Accounts Platform, February 2026, accessed June 23, 2026

324 Yingjie Zhang et al., *Bleeding Pathways: Vanishing Discriminability in LLM Hidden States Fuels Jailbreak Attacks*, arXiv:2503.11185, December 2025, accessed June 23, 2026, arXiv: 2503.11185 [cs.CR]; Yichi Zhang et al., *RealSafe-R1: Safety-Aligned DeepSeek-R1 without Compromising Reasoning Capability*, arXiv:2504.10081, April 2025, accessed June 23, 2026, arXiv: 2504.10081 [cs.AI]; Yichi Zhang et al., *Unveiling Trust in Multimodal Large Language Models: Evaluation, Analysis, and Mitigation*, arXiv:2508.15370, August 2025, accessed June 23, 2026, arXiv: 2508.15370 [cs.CL]; Yichi Zhang et al., *Towards Safe Reasoning in Large Reasoning Models via Corrective Intervention*, arXiv:2509.24393, February 2026, accessed June 23, 2026, arXiv: 2509.24393 [cs.AI]

325 Xiaojun Jia et al., *OmniSafeBench-MM: A Unified Benchmark and Toolbox for Multimodal Jailbreak Attack-Defense Evaluation*, arXiv:2512.06589, December 2025, accessed June 23, 2026, arXiv: 2512.06589 [cs.CR]

326 Xiaojun Jia et al., *SkillJect: Effectively Automating Skill-Based Prompt Injection for Skill-Enabled Agents*, arXiv:2602.14211, June 2026, accessed June 22, 2026, arXiv: 2602.14211 [cs.CR]; Xun Huang et al., *Obscure but Effective: Classical Chinese Jailbreak Prompt Optimization via Bio-Inspired Search*, arXiv:2602.22983, March 2026, accessed June 23, 2026, arXiv: 2602.22983 [cs.AI]