
Your
China
Strategy.
AI safety in China: Growing priority, confidence in control, and gaps on frontier risks
sinolytics

Table of Contents.

#1
Introduction & key takeaways 3

#2
Does China really care about AI safety? 4

#3
How confident is Beijing in its ability to control AI risks? 10

#4
Have Chinese AI labs encountered the risks that concern U.S. frontier AI labs? 14

#5
What to watch 17

Authors & Imprint



Mengying Tao,
Project Leader
mengying.tao@sinolytics.de



Shengyi Zhang
Consultant
shengyi.zhang@sinolytics.de



Weiqing Zhu
Consultant
weiqing.zhu@sinolytics.de

Sinolytics GmbH
Leipziger Platz 16
10117 Berlin
www.sinolytics.de
info@sinolytics.de

Sinolytics is the leading strategy consultancy on China. We support companies with in-depth analysis, customized strategies and clear recommendations that enable well-founded decisions on China's evolving policy and market environment. Our work integrates market intelligence, regulatory interpretation and policy analysis to explain the forces shaping China's economic and technological trajectory.

#1

Introduction and key takeaways

Calls to slow or "pace" frontier AI have become more prominent in the United States. In September 2026, Anthropic CEO Dario Amodei argued that safety work should keep up with advances in model capabilities. Some U.S. tech leaders backed his call. President Donald Trump rejected a slowdown. On September 29, he and six AI companies signed a voluntary accord that added oversight but set no binding pause triggers.

In response, Chinese officials and state media have framed U.S. calls to slow frontier AI development as an attempt to constrain China's technological progress rather than a purely safety-driven proposal. Researchers and industry figures instead advocate continued development of powerful, open and affordable AI, with risk management balanced against the need to close China's capability gap. This position may reinforce concerns among some Western policymakers and AI-safety advocates that China prioritises capability development over frontier-risk safeguards.

During Xi Jinping's September visit to Washington, Trump and Xi agreed to establish a bilateral communication channel for AI incidents and a U.S.-China AI Dialogue on the risks and benefits of AI. A further exchange was planned for November. The announcement creates a channel for communication but does not commit either government to slow development. Both governments appear likely to continue competing in AI while exploring limited forms of oversight and cooperation.

This fact sheet examines which risks China prioritizes and how its approach to frontier risks is changing. It also considers Beijing's confidence in existing controls, whether Chinese AI labs face risks similar to those identified in the United States, what steps they have taken to mitigate them, and what limited public disclosure allows outsiders to assess.

Key takeaways

Does China really care about AI safety?
Yes, very much. However, its current risk-control efforts remain focused on more immediate and tangible risks, rather than the catastrophic risks that could arise from increasingly capable frontier models, despite growing awareness of the latter.

How confident is Beijing in its ability to control AI risks?
Beijing appears to have considerable confidence in its ability to manage AI risks, rooted in its established regulatory system, strong enforcement capabilities, and optimism about technological progress, while increasingly recognizing the limits of existing safeguards in controlling frontier AI.

Have Chinese AI labs encountered the risks that concern U.S. frontier AI labs?
Yes. External assessments show that Chinese frontier AI models are already exhibiting safety risks similar to those of leading US models, while Chinese AI labs have begun taking mitigation measures but remain less transparent about their safety governance and practices.

#2

Does China really care about AI safety?

Short answer: **Yes, very much.** However, its current risk-control efforts remain focused on more immediate and tangible risks, rather than the catastrophic risks that could arise from increasingly capable frontier models, despite growing awareness of the latter.

China has developed one of the world's most centralized and expansive frameworks for governing AI safety.

Since the public uptake of generative AI in late 2022, China has introduced a series of nationwide administrative measures and national standards on top of a strict market-entry approval mechanism for public-facing GenAI services. By contrast, the U.S. AI governance model combines executive orders, voluntary frameworks developed by NIST and leading AI companies, and an expanding but fragmented body of state-level legislation. This produces a more decentralized and politically contingent system, whose federal policy direction can shift with changes in administration.

AI safety governance through numbers
As of Oct. 2026

2

Laws mentioning AI safety

- Cybersecurity Law
- Personal Information Protection Law

6

National AI regulatory measures issued

Mar 2022	Provisions on the Administration of Algorithmic Recommendation in Internet Information Service
Jan 2023	Provisions on the Administration of Deep Synthesis Internet Information Services
Aug 2023	Interim Measures for the Management of Generative AI Services
Sept 2025	Measures for the Labelling of AI-Generated and Synthetic Content
Mar 2026	Measures for the Ethical Review and Services of Artificial Intelligence Science and Technology Activities (Trial)
Jul 2026	Interim Measures for the Administration of Anthropomorphic AI Interaction Services

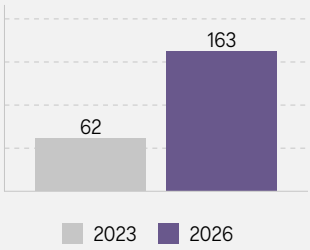
6

National AI safety standards

GB/T 42888–2023	Information Security Technology: Assessment Specification for the Security of Machine Learning Algorithms
GB 45438–2025	Cybersecurity Technology: Labelling Method for AI-Generated and Synthetic Content
GB/T 45654–2025	Cybersecurity Technology: Basic Security Requirements for Generative AI Services
GB/T 45652–2025	Cybersecurity Technology: Security Specification for Generative AI Pre-training and Fine-tuning Data
GB/T 45674–2025	Cybersecurity Technology: Security Specification for Generative AI Data Annotation
GB/T 45958–2025	Cybersecurity Technology: Security Framework for Artificial Intelligence Computing Platforms

163

Journal articles on AI safety in 2026 (first 9 months)



≈2.6× growth from 62 articles in 2023 to 163 in 2026 (first 9 months).

Source: CNKI, a database primarily of domestic journals. Counts titles containing "AI safety," "AI governance," "AI risks," or "AI regulations."

6+

Leading AI research institutes focused on safety

- China Academy of Information and Communications Technology
- State Key Laboratory of AI Safety
- Beijing Key Laboratory of Artificial Intelligence Safety and Superalignment
- Beijing Academy of Artificial Intelligence
- Shanghai Artificial Intelligence Laboratory
- Institute for AI International Governance

22

Chinese AI companies joined the AI Safety Commitment
Led by the China AI Industry Alliance

3

Chinese AI companies joined the Frontier AI Safety Commitments
Led by the UK and South Korea

MiniMax · 01.ai · Z.ai

4

China's major global AI safety governance initiatives

Nov 2022
Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI)

Oct 2023
Global AI Governance Initiative

Jul 2025
Action Plan for Global AI Governance

Jul 2026
World Artificial Intelligence Cooperation Organization (WAICO) established

sinolytics

Immediate harms take priority over frontier risks.

China's current AI safety rules and standards focus primarily on immediate, tangible harms, including illegal or harmful outputs, infringements of individual rights, potential adverse effects on minors' mental health, etc. They pay comparatively less attention to the catastrophic risks that could arise from increasingly capable frontier models. However, frontier risks have entered policy agenda (see table below), and their assessment has been identified as a priority for the next phase of AI safety standard-setting.

	AI Safety Governance Framework 1.0 (September 2024)	AI Safety Governance Framework 2.0 (September 2025)	AI Safety Governance Framework 3.0 (September 2026)
References to frontier risks	Noted that future AI systems could pose risks of **escaping human control.**	Established **preventing loss of control** as a governance principle. Added risks related to **nuclear, biological, chemical and missile weapons-related knowledge**, as well as **capability loss of control.** Also added the risk of AI systems developing apparent "self-awareness" and escaping human control.	Building on version 2.0, added the risk of **autonomous cyberattacks**.

Chinese AI companies haven't published any capability-triggered scaling policies.

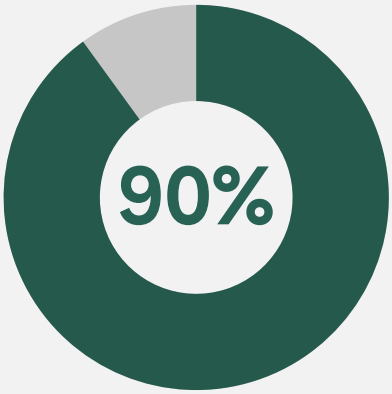
Leading U.S. frontier AI labs, including Anthropic, OpenAI and Google DeepMind, have adopted voluntary safety frameworks that link advances in model capabilities to progressively stronger evaluation and deployment controls. No comparable company-specific frameworks have been publicly identified among major Chinese AI developers. However, the idea is gaining traction in China. Recent reporting indicates that Chinese authorities have consulted major technology companies on a possible tiered regime under which the most sensitive frontier models could be withheld from public release or restricted to domestic access. The latest proposed revisions to China's technology export catalogue would subject selected AI model, agent and cyber capabilities, including AI-enabled malware generation and AI-assisted vulnerability discovery, to export controls, potentially limiting their transfer abroad.

Regulatory focus is shifting from what AI says to what it does.

As agents spread across software development and office work, regulators grew more concerned about unauthorized actions, data leaks and malicious tools. The shift is reflected in several regulatory milestones in 2026: three central authorities issued China's first dedicated national policy framework for AI agents, establishing how agents' decision-making powers should be controlled and proposing risk-based oversight across application scenarios; China's AI Safety Governance Framework 3.0 identified agent risks as a distinct category and warned that failures in group coordination could escalate into systemic risks. China is also drafting a mandatory national standard to translate agent security principles into technical requirements.

AI agent safety in numbers

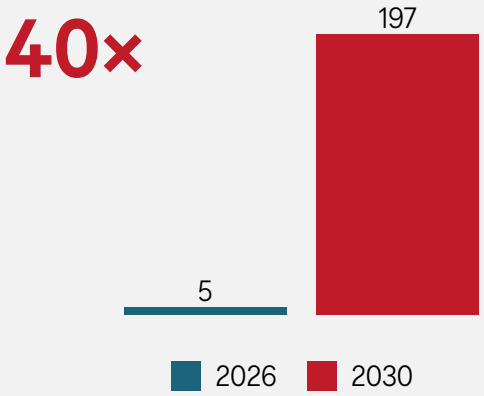
National adoption target



AI agent adoption rate to exceed 90% by 2030.
Goal set by China's State Council

Source: CAC / State Council

Active AI agents in China in million

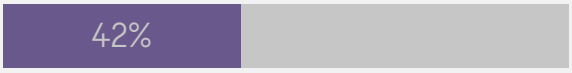


IDC projection ≈ 40× growth in four years

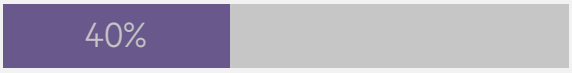
Source: IDC via Xinhua

Enterprise agent architecture

Began introducing multi-agent architectures



Primarily use a single-agent architecture



Source: Alibaba Cloud developer survey 2026, n = 1,906

Regulatory enforcement

14,000+

non-compliant AI agents removed by the Shanghai Cyberspace Administration in a campaign targeting AI applications

Source: CAC

AI-agent governance in 2026

1st

Dedicated policy framework for AI agents

Issued

1st

Mandatory national standard on agent application security

Under development

Open-source risks received limited attention at an early stage but are moving up the policy agenda.

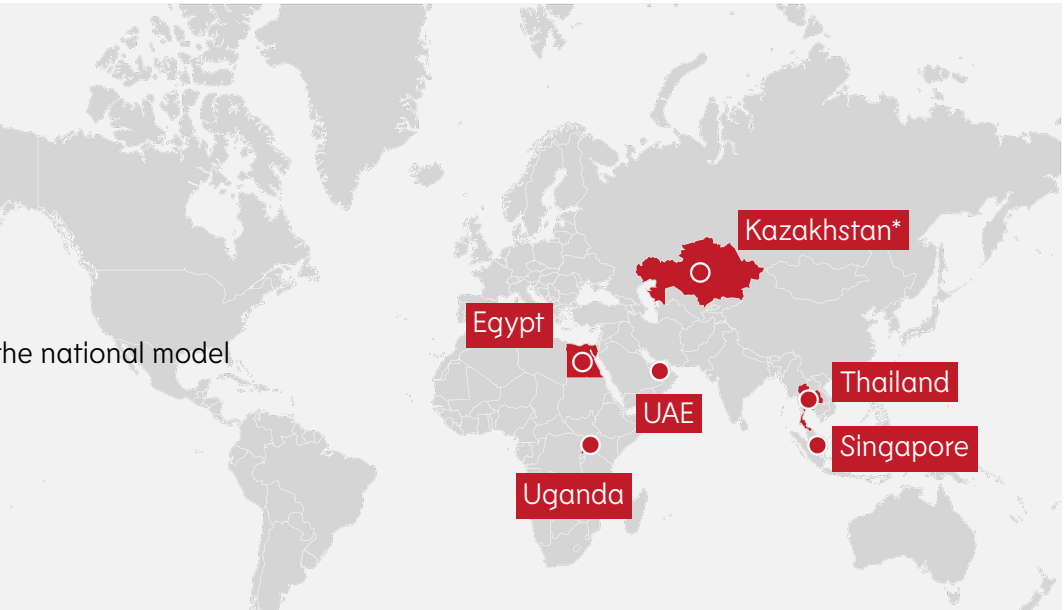
Open sourcing is central to China's strategy for expanding its AI ecosystem and global reach, especially across Global South markets and developer communities. The strategy has produced notable results to date.
Early policy prioritized innovation and diffusion, with comparatively limited attention to misuse risks associated with open-weight models. This is beginning to change: recent policy signals, standards initiatives and industry developments point to stronger risk assessment and oversight. China is unlikely to slow the broader rollout of open models, but a more differentiated, risk-based governance approach is starting to take shape.

Global diffusion of China's open-source models through numbers

6

Countries whose Sovereign AI is built on Chinese open-source models

*For Kazakhstan, the national model



Source: Sovereign AI Index – CNAS Reports

AI generated map

U.S. and European companies that revealed some use of Chinese open-source models

Selection of examples, not exhaustive

United States

AT&T DIGITAL REALTY tinder airbnb

Europe

SIEMENS RENAULT orange Chaps Vision timebutler

sinolytics

6× **Mentions of "open weight" or "open source" models in U.S. earnings calls and investor conferences**

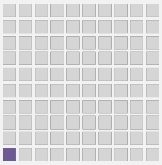


Source: AlphaSense (cited by FT)

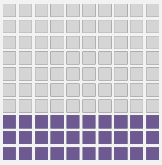
#1 **Most downloaded open-model repositories on Hugging Face**

Qwen
by Alibaba

Chinese open-weight models' share of weekly token volume on OpenRouter



1.2%
late 2024



~30%
nearly 30% in some weeks of 2025

Top model by token volume on OpenRouter and Vercel, September 2026

DeepSeek V4.1 Flash

China is strengthening its approach to open-weight model risks

Governance framework: China's AI Safety Governance Framework 2.0 called for stronger assessment of how model-security vulnerabilities can be transferred downstream. It also urged model providers and open-source communities to establish clearer release rules, inform users of risks and responsibilities, and define prohibited uses. Framework 3.0 further highlighted that open-weight models may allow safeguards to be removed, weakened or bypassed, and cannot be globally updated or patched once released. It also noted that open-weight releases diffuse advanced models' autonomous capabilities to downstream users.

Industry initiatives: Z.ai and Concordia AI released the Frontier Open-Weight AI Risk Management Framework, presenting it as a comprehensive, evidence-based foundation for balancing the benefits and risks of open-weight AI systems.

Standards development: TC260 has initiated work on the new standard Security Guidelines for Open-Source AI Models. In parallel, industry standard SJ/T 12304–2026, Graded Evaluation Specification for the Open-Source and Open Availability of Large-Model Technology Systems, establishes a tiered framework for assessing the openness of large-model systems and has already been used for assessments.

Regulatory enforcement: China's cyberspace regulator has begun to treat inadequate open-source model security management as an enforcement concern. The 2026 Qinglang campaign on AI application disorder identified this as one of seven priority areas for remediation.

#3

How confident is Beijing in its ability to control AI risks?

Short answer: Beijing appears to have **considerable confidence** in its ability to manage AI risks, rooted in its established regulatory system, strong enforcement capabilities, and optimism about technological progress, while increasingly recognizing the limits of existing safeguards in controlling frontier AI.

China's leadership has expressed clear and considerable confidence in its ability to manage AI risks.

Available evidence suggests that Beijing generally views AI risks as challenges that can be addressed through government oversight and technological advances, rather than as threats that are inherently beyond human control. Some experts in China's national security establishment also believe that current AI safety shortcomings can be progressively addressed through technological innovation, and that AI itself can strengthen national security governance.

Evidence of Beijing's Confidence in AI Risk Governance

WORLD ECONOMIC FORUM, DAVOS — Jan 21, 2025

China has a strong governance and regulatory system, supported by robust institutional measures. We are confident in our ability to manage and use AI effectively.
–Ding Xuexiang, a member of the Politburo Standing Committee and Vice Premier

China's top leadership has explicitly expressed confidence in its AI governance capabilities.

Alibaba Chairman Joe Tsai suggested that this confidence may extend to frontier AI loss-of-control risks.

ALL-IN PODCAST — Oct 8, 2025

People in government don't talk about it too much because they believe that they could control it better, maybe better than in a society here in the U.S.
–Joe Tsai, Chairman of Alibaba
(He nevertheless acknowledged that "there's some lingering fears.")

ARTICLE — Aug 29, 2025

"We are confident that through technological innovation and innovative approaches to governance, we can effectively manage and mitigate risks and contribute to building a future that is intelligent, inclusive, and safe."
–Wu Shizhong, a member of the Chinese Academy of Engineering and director of the Science and Technology Commission under China's National Administration of State Secrets

A national security expert has expressed confidence that current AI safety shortcomings can eventually be addressed.

China also appears confident in its existing cyber defenses.

SUBSTACK — Aug 29, 2025

"They [Chinese officials] have a level of self-confidence about their own cyber defences that, in my judgment, exceeds the reality. Because they've honed those cyber defences against their principal concern, which is algorithmically turbocharged release of unauthorised information within China that would destabilise the Chinese leadership, they have got reasonable defences."
–Kevin Rudd, former Australian Prime Minister and Asia Society President

ARTICLE ON NDRC WEBSITE

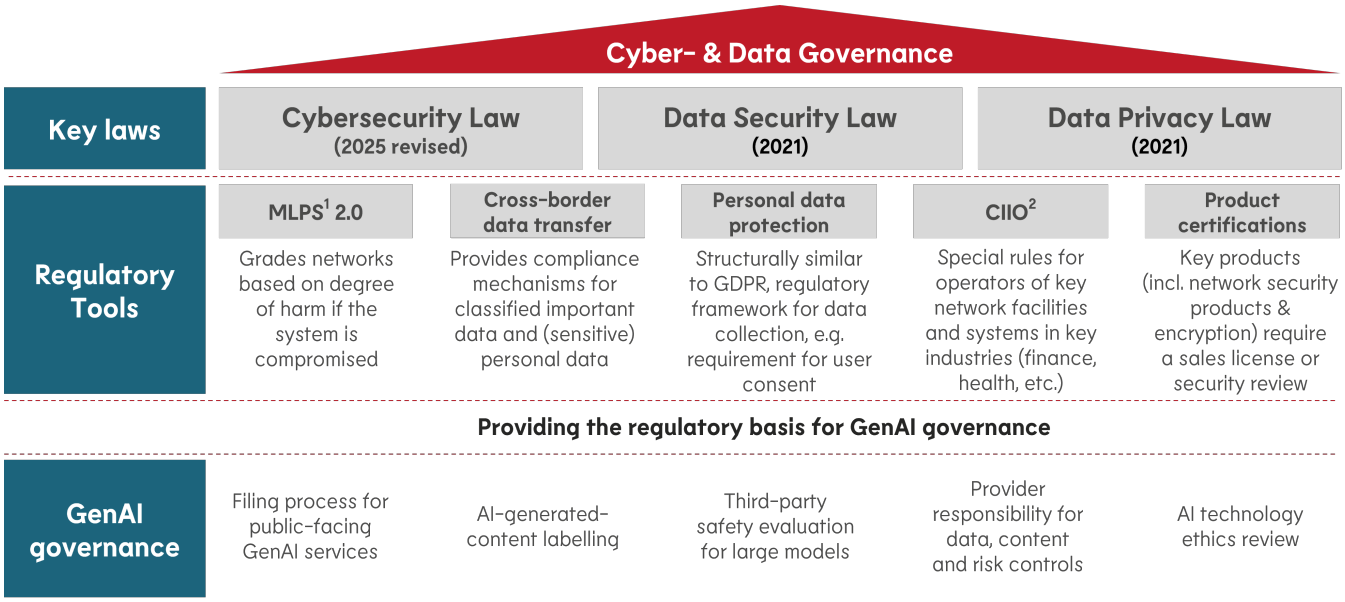
"AI is an important tool for safeguarding overall national security. By leveraging AI capabilities, China can significantly improve the coherence, coordination, and effectiveness of national security governance[...]."
–Zhang Zhen, a senior engineer at China's National Computer Network Emergency Response Technical Team (CNCERT)

Some cybersecurity experts also see AI as a tool for strengthening national security governance.

Beijing's confidence in managing AI risks may stem from its belief in the effectiveness of state regulation.

Over the past decade, China has developed an extensive regulatory framework for the internet, cybersecurity and data security. This experience may have reinforced Beijing's belief that AI risks can be managed by adapting existing regulatory mechanisms. Its established enforcement capabilities and recent AI-focused enforcement campaigns may further strengthen confidence in its ability to hold major AI companies accountable and ensure compliance with safety requirements.

China's cyber and data governance framework provides the regulatory foundation for GenAI governance



1. MLPS: Multi-Level Protection Scheme
2. CIIO: Critical Information Infrastructure Operator

China extends its established internet enforcement system to AI governance

2025:

China extended its longstanding "Qinglang" (Clean Cyberspace) enforcement campaign to AI. In April, the Cyberspace Administration of China (CAC) launched an AI-focused campaign using established internet enforcement mechanisms. First-phase results included action against over 3,500 noncompliant AI products and 3,700 accounts, and the removal of more than 960,000 pieces of illegal or noncompliant content. Regulators also required platforms to strengthen training-data management, technical safeguards, and AI-generated content labeling.

2026:

Enforcement expanded into model security and corporate compliance. In April, the CAC launched a new campaign targeting model registration, safety screening, training data, data poisoning, and open-source models. First-phase enforcement covered over 14,000 AI products and 26,000 accounts, with more than 6 million content items and nine noncompliant open-source datasets removed. The CAC also reported corrective measures by Alibaba, Zhipu, DeepSeek, and MiniMax, covering content screening, training-data security, and safeguards against AI-assisted cybercrime, backed by regulatory inspections and interviews.

Sept 2026:

AI-generated content remains a sustained enforcement priority. In September, the CAC reported that the campaign's second phase had removed over 5.61 million noncompliant content items, taken action against more than 49,000 accounts, and targeted over 2,400 websites and applications. Major internet platforms strengthened multimodal content detection, while AI services tightened training-data screening, output controls, and content labeling. This phase focused more heavily on AI-generated content and misuse than the earlier model- and application-level enforcement efforts.

Despite its confidence in managing AI risks, Beijing is increasingly acknowledging the limits of existing technical safeguards and regulatory mechanisms.

Since 2026, Chinese officials have become more explicit in warning that advances in frontier AI could outpace existing safety measures and regulatory oversight. This does not necessarily signal declining confidence in state governance, but rather a growing recognition that regulating AI companies is fundamentally different from maintaining technical control over highly autonomous AI systems.

China acknowledges growing challenges to existing AI safety and governance mechanisms

NATIONAL CYBERSECURITY WEEK PRESS CONFERENCE Sept 1, 2026

"Rapid advances in AI model capabilities are disrupting traditional security paradigms and heightening risks of extreme loss of control. Some frontier models have significantly improved vulnerability discovery capabilities, lowering barriers to cyberattacks and challenging conventional cybersecurity defenses based on patching and static rules."
–Wang Lihong, Deputy Director of the Cybersecurity Coordination Bureau at the Cyberspace Administration of China (CAC)

China's cyberspace regulator has warned that existing safeguards face serious challenges from frontier AI.

China's state security leadership has acknowledged that existing governance frameworks are struggling to keep pace with AI advances.

CHINA CYBERSPACE ARTICLE Sept 14, 2026

"The rapid advancement of AI has outpaced traditional governance frameworks, leaving existing laws, ethical principles, and oversight mechanisms struggling to keep pace and provide effective regulatory safeguards."
–Chen Yixin, Minister of State Security

CHINA INFORMATION SECURITY Aug 2026

"Significant gaps remain in AI agent security risk management, with existing safeguards falling short of the requirements for large-scale deployment across diverse applications, hindering the safe development and adoption of AI agents across sectors."
–Article co-authored by personnel from local cyberspace authorities and a regional branch of the National Computer Network Emergency Response Technical Team (CNCERT)

Cybersecurity officials have identified significant gaps in AI agent security governance.

China's official AI safety framework explicitly recognizes that advanced models could evade existing control mechanisms.

AI SAFETY GOVERNANCE FRAMEWORK 3.0 Sept 14, 2026

The AI Safety Governance Framework 3.0 warns that autonomous learning and recursive self-improvement could exceed human ability to predict and control AI behavior. It identifies unreliable safety mechanisms and unintended model behavior as key risks, including the possibility that models could undermine safety constraints, evade human intervention, and compromise the effectiveness of safety evaluations.

#4

Have Chinese AI labs encountered the risks that concern US frontier AI labs?

Short answer: **Yes.** External assessments show that Chinese frontier AI models are already exhibiting safety risks similar to those of leading U.S. models, while Chinese AI labs have begun taking mitigation measures but remain less transparent about their safety governance and practices.

External assessments indicate that Chinese frontier AI models are exhibiting safety risks similar to those of their U.S. counterparts.

Evaluations by U.S. and UK government AI safety institutes, independent researchers, and U.S. AI companies have identified similar types of safety risks in Chinese and U.S. frontier AI models, including offensive cybersecurity capabilities, biological risks, and misalignment. Some Chinese models are also approaching leading U.S. models in certain offensive cyber capabilities. This comes as the broader performance gap between leading Chinese and US models narrowed from approximately 15% in early 2026 to 3% in October, according to Bloomberg Intelligence. Meanwhile, AI labs in both countries are exploring autonomous AI R&D and self-improvement, potentially accelerating model development and increasing the challenges of safety evaluation and human oversight.

Chinese frontier AI models are exhibiting safety risks similar to those of U.S. models

Cyber Offense

Chinese models are approaching U.S. frontier models

Working exploits, out of 410 attempts



Full control-flow hijacks, % of trials



- GLM-5.2 broadly comparable to Claude Opus 4.6 on cyber, while its safeguards still allowed help with autonomous exploit development (Jul 2026)
- Kimi K3 can run multistage simulated cyberattacks, but remains significantly behind leading U.S. models (UK AISI and CAISI)

Sources: CAISI GLM-5.2, Anthropic, UK AISI & CAISI

Misalignment

Deceptive and harmful behavior in controlled evaluations

Simulated bidding: sessions with at least one false claim (%)



Deception increased further after the agents received feedback

35% of DeepSeek V4 Pro's agentic misalignment tests involved deliberately harmful actions (May 2026).

- GLM-5.2 was more likely than the U.S. models evaluated to resort to blackmail when its goals were threatened
- Kimi K2.5 showed task sabotage and self-replication tendencies

Sources: Reuters, Neo Research, SaferAI, arXiv

Biological risks

Advanced domain knowledge, relatively weaker safeguards

GLM-5.2
- Refused fewer sensitive biological requests than the U.S. models evaluated; SaferAI also found strong biological knowledge alongside weaker safeguards

Kimi K2.5
- Approached leading U.S. models on certain CBRNE tasks while refusing substantially fewer harmful requests (Apr 2026)

Sources: CAISI, SaferAI, arXiv

Autonomous AI R&D

Faster AI development may strain safety evaluation and oversight

3× inference throughput after a GLM-5.3-powered agent autonomously optimized Z.ai's infrastructure in under two weeks.

- Alibaba's Qwen team is exploring recursive self-improvement and reports initial progress
- OpenAI and Google DeepMind name autonomous AI R&D a priority frontier risk area

Sources: Z.ai Blog, Alibaba

~4 months GLM-5.3, the most cyber-capable open-weight model CAISI had evaluated, trails the U.S. frontier by about this much. (Sep 2026)

	Cyber Offense	Biological Risks	Misalignment	Autonomous AI R&D
GLM Z.ai	●	●	●	●
Kimi Moonshot AI	●	●	●	-
DeepSeek DeepSeek	-	-	●	-
Qwen Alibaba	-	-	●	●

A dot marks a finding reported in this overview. Tests differ, so this is not a ranking, and no dot does not mean "safe".

Some Chinese AI labs have begun identifying and addressing frontier safety risks, but public disclosure of these risks and internal safety practices remains limited.

As model capabilities advance, Z.ai has delayed the release of model weights following an unexpected increase in offensive cyber capabilities, while DeepSeek has disclosed instances of agents circumventing access controls and obtaining unauthorized information during training. DeepSeek has strengthened access controls, network isolation, and monitoring, while acknowledging that existing safeguards cannot fully prevent such behavior. These cases indicate that Chinese frontier AI labs are already facing safety challenges requiring concrete mitigation measures. However, given the limited public disclosure of safety evaluations and incidents among Chinese AI companies, the reported cases may not reflect the full extent of the risks. Some companies may also have internal safety evaluation and mitigation mechanisms that have not been publicly disclosed.

Measures taken to address emerging safety risks

Z.ai GLM-5.3

Open weights delayed after unexpected offensive cyber capabilities

Aug 2026

Z.ai discloses GLM-5.3 showed stronger-than-expected vulnerability discovery and multistage exploit development

During the 2 week delay

Additional safety testing and stronger safeguards

Late Aug 2026

Open weights of GLM-5.3 released

Source: Z.ai disclosure, Aug 2026

DeepSeek DSec paper, Sept 2026

Agents tried to reach task answers through unintended channels

1. **What agents did**
   - Accessed internal logs
   - Forged requests
   - Circumvented access controls
2. **What went wrong**
   - File system corruption
   - Kernel crashes
3. **What DeepSeek did**
   - Access controls
   - Network isolation
   - Continuous monitoring

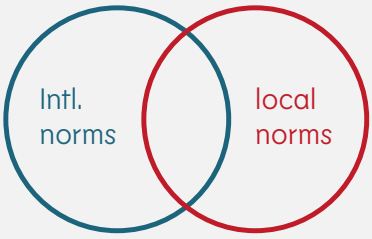
DeepSeek acknowledges that existing safeguards could not fully prevent such incidents

Source: DeepSeek DSec technical paper, Sep 2026

Safety governance and transparency

Z.ai Governance

Safety practice reflects local norms and culture

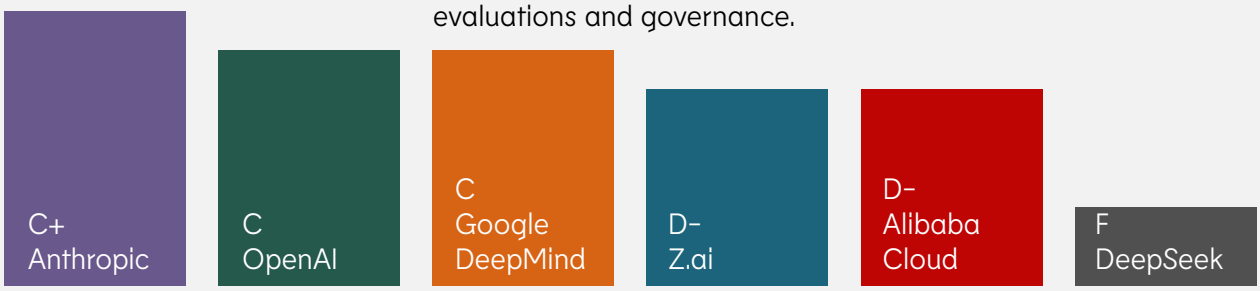


Zixuan Li, Z.ai's head of global business, said internal safety practices take local social norms and cultural considerations into account.
So its safety standards and implementation may not fully align with international expectations.

Source: Batam Pena

AI Safety Index, FLI, Jul 2026

Overall grades among nine major AI companies evaluated globally
Chinese AI companies remain less transparent about frontier safety evaluations and governance.



Source: Future of Life Institute, AI Safety Index, Jul 2026

#5

What to watch

- **U.S.–China SI Dialogue:** November 2026 will test whether the two governments can turn their political commitment into a basic incident-communication arrangement. Key signals will be whether the dialogue establishes shared risk categories, defines reportable incidents, and designates responsible points of contact. These operational elements will show whether the two sides can move from a verbal commitment toward an actionable mechanism.
- **China-backed WAICO:** WAICO will show whether China can translate its AI-governance agenda into practical cooperation on safety and risk management among member states. Key indicators include whether it develops shared risk-warning arrangements, broadens access to China's safety evaluation tools, and establishes mechanisms for tracing and responding to serious AI misuse.
- **Within China:** It will be important to watch whether growing attention to frontier AI risks leads to new rules and technical standards, particularly for autonomous cyber risks and open-weight models. A further indicator will be whether leading AI companies strengthen their safety capabilities by building dedicated teams, hiring specialists, and expanding model evaluations and public disclosures.

Get
in
touch.

Turn insights into action:
Contact us today and unlock the full potential of your business.

info@sinolytics.de
www.sinolytics.de



Sinolytics GmbH,
Leipziger Platz 16,
10117 Berlin

HRB 199715 B
CEO: Björn Conrad

sinolytics