

Between Development and Oversight: China's Approach to Addressing Artificial Intelligence Risks

Meny Vaknin | No. 2205 | October 25, 2026

The rapid development of artificial intelligence capabilities is broadening the way China defines the risks posed by this technology. Alongside espionage, cyber threats, and familiar national security threats, Beijing is increasingly addressing the risks originating from advanced AI systems themselves—including autonomous operation and loss of control. At the same time, technological lag and dependence on foreign technologies are also perceived as risks to national security. Official publications from September 2026 show that different approaches to these risks exist within the Chinese system, but all call for continued development of domestic AI capabilities alongside expanded mechanisms for safety, oversight, and control.

In mid-September 2026, Chinese officials issued three major statements on AI risks. On September 13, Minister of State Security Chen Yixin [published an article](#) addressing their implications for Chinese national security. The following day, the [Artificial Intelligence Safety Governance Framework 3.0](#) was released, mapping the risks and the tools for addressing them. The [Chinese Foreign Ministry also addressed](#) AI risks and the possibility of slowing the development of advanced models.

This sequence of statements reflects different perspectives among China's institutional bodies. Chen's article examines the technology through the lens of national security, while Framework 3.0, developed by cybersecurity and standardization bodies, gives greater weight to risks arising from the systems' evolving capabilities. The Foreign Ministry connects the two dimensions, presenting risk management as compatible with continued development. In any case, the boundaries between them are not clear-cut: the Chinese term *anquan* (安全) denotes both "safety" and "security," bringing technological safety, cybersecurity, political stability, and national security under a single conceptual framework.

Artificial Intelligence Through the Lens of National Security: The Minister of State Security's Article

Chen Yixin, Secretary of the Party Committee and Minister of State Security (MSS), published his article ahead of [National Cybersecurity Awareness Week](#). It draws heavily on earlier ideas and formulations of the Chinese leadership. A [comparative analysis](#) by Mark Parker Young, a former CIA analyst and senior official at the US National Intelligence Council, found that a significant portion of the minister's language is based on earlier documents and statements by President Xi Jinping and other Chinese officials. However, the section on AI risks contains a high proportion of material that could not be traced to previous texts, suggesting that the innovation lies primarily in applying existing security concepts to recent developments in the field.

Chen outlines six types of risks stemming from the development of artificial intelligence. For the purposes of this analysis, they can be grouped into three spheres. The first is the political and social sphere, which includes threats to political and ideological security, as well as challenges to governance and social order. In this context, Chen warns against the use of deepfakes, AI-generated content, and automated accounts to spread disinformation, inflame disputes, and wage cognitive warfare against China. He also points to algorithmic bias, privacy violations, and the difficulty of establishing accountability for automated decisions.

The second sphere concerns cybersecurity, espionage, and information security. According to Chen, artificial intelligence lowers the cost and skill barrier for executing cyberattacks, enabling the discovery of vulnerabilities and the automation of attacks, and even “AI versus AI” confrontations. He further warns against foreign intelligence services using the technology to collect and analyze sensitive data, as well as data leakage to foreign AI products.

The third sphere is technological and military competition. Chen frames technological gaps and dependency as a risk, arguing that countries with an advantage in the field use technology restrictions, standards, and closed systems to limit other countries’ access to advanced technologies. At the same time, he points to the increasing integration of AI in intelligence, decision-making, target identification, operational support, and cognitive influence. He describes a shift of the technology from a supporting role to a “pivotal role” on the battlefield.

Equally interesting is what does *not* feature prominently in Chen’s list of risks. While Framework 3.0 (see below) emphasizes the risk of autonomous systems exceeding their assigned tasks or escaping human control, Chen focuses on how people, states, and hostile actors use AI—and on the risks posed by technological lag. This reflects the concept of “holistic national security,” which views technology as a force multiplier for existing threats and as another arena of competition between states. Accordingly, Chen defines innovation as the “primary driving force” and security as the “bottom-line requirement,” calling for the reinforcement of self-reliant AI capabilities alongside mechanisms for monitoring, early warning, and oversight.

Risk Management: The Artificial Intelligence Safety Governance Framework 3.0

Framework 3.0 was developed under the guidance of the Cyberspace Administration of China (CAC) and organized by the National Cybersecurity Standardization Technical Committee (TC260), in collaboration with research institutes, professional bodies, and technology companies. The institutional response is distributed among several entities: the CAC leads regulation in the areas of services, content, algorithms, and data; TC260 handles standardization and technical tools; and the Ministry of State Security brings a national security perspective. AI providers are responsible for implementing the safety requirements, under the supervision of the relevant regulators.

This is the third version of the framework, which has been updated annually since September 2024. It is not a binding law or regulation, but rather a guiding document built around three tiers: mapping and classifying risks, using technological measures to reduce them, and establishing oversight and governance mechanisms. Among other things, the framework recommends risk assessments, monitoring and early warnings, restricting the permissions of AI agents, human involvement in high-risk operations, audits of tools and plug-ins, and mechanisms for shutting down systems that exceed their

assigned tasks. One of the system's guiding principles is promoting AI development while ensuring it remains "safe and controllable" (安全可控).

The framework operates alongside a binding regulatory system that has evolved in China in recent years. This includes rules for generative AI services, registration and review procedures for services and algorithms, and requirements for labeling AI-generated content. As of late August 2026, over one thousand generative AI services have completed the national-level registration process in China. Framework 3.0 provides a risk roadmap and technical principles to guide the further development of standards and regulation, while legal obligations remain anchored in the existing regulatory system.

[According to an analysis by AI Safety in China](#), an independent research project tracking AI policy and governance in China, the number of risks detailed in the framework grew from 30 in Version 2.0 to 54 in Version 3.0. Particular attention is given to AI agents, which can plan, perform tasks, and operate in external systems, and to embodied AI, in which a digital failure could cause physical harm. The framework also addresses systems that obtain unauthorized permissions or resources, deceive safety evaluators, deviate from their intended goal, or act autonomously in cyberspace. In [commentary published following the framework's presentation](#), this change was described as a shift in large models from "answering questions" to "executing tasks."

The framework also addresses open-weight models—models whose parameters learned during training are publicly available, allowing them to be run and modified independently. The framework warns that once weights have been distributed, the model can be run independently, modified, or even stripped of its safety mechanisms, making it difficult to retroactively restrict copies that have already been released. This warning is particularly noteworthy given the prominent role of Chinese open-weight models in the domestic industry. DeepSeek, for instance, [released the weights for its V4 models](#) and presented this openness as part of its development strategy. This creates a tension between the contribution of open weights to innovation, adoption, and China's technological influence on the one hand, and the difficulty of controlling their use after distribution on the other.

[A comparison of the document's three versions](#) illustrates how China's conception of risk has expanded alongside technological developments. Version 1.0, published in 2024, focused primarily on risks stemming directly from the technology itself and its use. Version 2.0 broadened the discussion to social risks, including impacts on employment and ethics. Version 3.0 further expands the focus to risks arising from the evolving capabilities of advanced systems, such as autonomous operation and AI agents.

The difference between Chen's article and Framework 3.0 demonstrates that the concept of risk operates across multiple dimensions. Externally, Beijing is concerned about falling behind in technological competition, dependence on foreign technologies, and the use of AI by states and hostile actors for cyber operations, espionage, and influence. Within China, there are also concerns about threats to political security and social order. Framework 3.0 adds another layer: the risk posed by the behavior of advanced systems themselves. Artificial intelligence is thus viewed simultaneously as a source of national power and a potential threat.

Development Alongside Security: The Chinese Foreign Ministry's Position

On September 14, Chinese Foreign Ministry Spokesperson Guo Jiakun was asked about calls by senior US AI industry figures to slow the development of advanced AI systems, and particularly regarding

[comments by Anthropic CEO Dario Amodei](#), who linked the risks of advanced AI models to technological competition with China. [Guo responded that China advocates for the "open and inclusive" development of AI](#), and warned against threat-mongering, confrontation, and "malicious competition."

The following day, Guo clarified that China also recognizes the risks themselves. According to Guo, regulation must be adapted and safeguards strengthened to prevent the technology from spiraling out of control. He emphasized that China places "[equal emphasis on development and security](#)," takes both the inherent and derivative risks of AI seriously, and works to ensure that the technology remains "safe, reliable, and controllable." In this context, Guo explicitly referenced Framework 3.0, which had been published the previous day.

The tension is clear: [the Chinese Foreign Ministry opposes using AI safety rhetoric to justify confrontation and tech rivalry](#), whereas Chen frames technological lag, dependency, and access restrictions as part of the national security threat landscape.

The issue also reached the highest political level. During their meeting in Washington on September 24, [Presidents Xi Jinping and Donald Trump discussed](#) artificial intelligence. Xi called for continued dialogue between the two countries on the technology's risks and benefits, to prevent its misuse and ensure that it remains under human control. Following the summit, both sides agreed to promote a dedicated dialogue on the subject and establish a communication channel for handling AI-related incidents. However, just days after the meeting, [Trump made clear](#) that he also sees the domain as a direct strategic competition with China, expressing reservations about cooperation that could grant Beijing access to American tech and expertise. This reflects, at the leadership level, the ongoing friction between managing shared AI risks and driving technological competition between the two powers.

Between Development and Control: Implications for the Future

Chen Yixin's article, Framework 3.0, and the Chinese Foreign Ministry's responses present different perspectives within China's system, reflecting the roles and responsibilities of the bodies behind them. The security apparatus emphasizes threats to the state and regime, as well as to China's ability to compete. Cybersecurity and standardization bodies give increasing weight to the behavior of advanced systems and the possibility of losing control over them. The Foreign Ministry seeks to acknowledge the risks while opposing the use of safety dialogue to justify restrictions on China's technological development. The Chinese concept of *anquan* (安全) makes it possible to bring these dimensions together under a single, broad framework encompassing technological safety, cyber and information security, political stability, and national security.

These developments also have implications for international AI governance. The agreements reached at the Xi–Trump meeting suggest there is some basis for dialogue between China and the United States on loss of control, malicious use, cyber risks, and autonomous systems. However, ongoing technological competition makes it difficult to separate safety from national security. US officials view restricting China's access to advanced chips and capabilities as a mechanism to mitigate risks, whereas Beijing considers technological lag and dependency on foreign entities to be critical risks in their own right. Dialogue between the two countries will therefore need to address both the definition of risks and the disagreement over legitimate methods for managing them. The overlap in identifying risks, such as loss of control and the misuse of advanced systems, may nevertheless provide a basis for focused cooperation.

Editors of the series: Anat Kurz, Rinat Harash, Eldad Shavit and Keri Rosenbluh