Beijing Will Not Pace the Frontier: China's Speed-First AI Safety Regime

MARK CHEN, DOUG, AND DYLAN PATEL

44 MINUTES AGO • PAID

2 Share

AI safety is on fire. On 12 September 2026, Dario Amodei published an essay arguing that frontier labs must deliberately slow the pace at which they improve model capabilities. China was central to his geopolitical argument: export controls and tighter security could preserve America's lead, creating room to slow down and leverage for a future agreement with Beijing. Within a day, Sam Altman and Elon Musk supported Dario's call. However, President Trump scoffed. On 14 September 2026, he wrote that fears of AI destroying humanity are "a HOAX" and that "the only one that is happy about it is China." Beijing scoffed too, with the state-run Global Times calling Dario's essay a "Cold War playbook."

The debate then moved onto the diplomatic calendar. On 20 September 2026, Treasury Secretary Scott Bessent and Vice Premier He Lifeng held the first US-China AI dialogue. On 23 September 2026, President Xi Jinping began his state visit to the US. The Trump-Xi summit formalized the dialogue as the "US-China Super Intelligence Dialogue," using the name Trump had unveiled at the UN the day before Xi arrived. The summit also promised a channel for AI incidents. On 29 September 2026, Trump and six executives signed the White House Accord on Super Intelligence, a one-page voluntary pledge to self-police that Trump called "morally binding."

China sits on both sides of the American argument. For President Trump, competition with China is a reason not to slow down. For Dario, preserving America's lead over China is what makes slowing down possible. Nobody has asked what China is actually doing. So, we did. What do the Chinese government, frontier labs, and AI experts say and do about AI safety? And what should American labs, investors, and policymakers conclude from it?

What Beijing Says and Actually Does

Beijing's stated position is the most safety-forward of any major AI power outside Europe. Xi Jinping told the Politburo in April 2025 that AI brings "unprecedented risks and

ensure AI always remains under human control" and that the faster AI advances, "the more promptly the measures to prevent loss of control must be improved." China signed the Bletchley Declaration and the Paris and New Delhi statements, co-sponsored the UN resolutions, founded a World AI Cooperation Organization in Shanghai, and publishes bilingual AI Safety Governance Frameworks.

The newest text goes further. The AI Safety Governance Framework 3.0, released by TC260 under Cyberspace Administration of China (CAC) on 14 September 2026, warns that AI has shown a self-accelerating trend of autonomous learning and recursive self-improvement, and calls for "attention and vigilance" for potential AI evolution exceeding human prediction and control. The Framework 3.0 also introduces a new model-risk category, "behavior deviating from expectations," covering models that acquire system privileges and external resources without authorization, bypass safety protections, deceive evaluators, hide their true capabilities, and refuse user instructions, illustrated with a boxed case study of models disabling shutdown scripts, sandbagging under evaluation, and breaking out of isolation into real systems. Its list of derivative risks goes further still; past behaviors already observed to a hypothetical AI that becomes self-aware, seeks power of its own, and competes with humanity for control. On paper, this is the most explicit official Chinese text yet on frontier risk, and its vocabulary overlaps strikingly with Dario's essay Beijing denounced that same week.

However, we think China's real approach to AI safety is speed-based, not safety-based. Though Beijing recognizes frontier AI risks, it has been prioritizing rapid development. The Framework 3.0 opens its principles with "promoting AI innovation and development as *the first priority*." The comprehensive AI Law promised in the 2023 and 2024 legislative plans was shelved in 2025, four months after the DeepSeek moment. The State Council's AI+ Action Plan targets 70% penetration of agents and intelligent terminals by 2027 and 90% by 2030, aiming for a rapid deployment of AI across all sectors and industries.

We think the Chinese government takes an open position on frontier models but tightens the applications and regulates outputs. Since 2025, China has issued a number of mandates, including mandatory content labeling (Sep 2025), Cybersecurity Law Article 20 (Jan 2026), the minors' information classification rules (March 2026), MIIT's AI ethics-r (March 2026), the policy-level agent rules (May 2026), the world's first AI-com (July 2026), the network-data risk assessment rules (August 2026), and MPS in powers (October 2026). Every one of these governs what AI says and what AI

