ARTICLE

DeepSeek and Other Chinese Firms Converge with Western Companies on AI Promises

The AI race is breaking open. An upcoming summit offers an opportunity to U.S. and Chinese companies to agree on safety and security measures.

By **Scott Singer**
Published on January 28, 2025

DeepSeek, a previously little-known Chinese AI start-up, sent shockwaves globally when it released one of the world's best performing open-source generative AI models last week. As Chinese frontier AI capabilities rapidly grow, so do the risks that a Chinese model could be powerful enough to cause global harm. An earlier, less powerful version of DeepSeek's model **was shown to be easily jailbroken** and reportedly provided a user with a recipe for methamphetamine.

Despite growing global concern around large-scale risks, the U.S. and Chinese governments have made little progress on a bilateral agreement to regulate frontier AI. But a surprising consensus among leading AI developers in both countries around the need for safeguards has quietly emerged, including DeepSeek.

Last month, DeepSeek joined sixteen other Chinese companies in **signing onto the Artificial Intelligence Safety Commitments** (人工智能安全承诺). While branded as a domestic Chinese initiative, the commitments bear strong similarity to ongoing global industry-led efforts to put safeguards in place for frontier AI piloted at last year's AI Summit in Seoul, known as the **Seoul Commitments.** Using similar language, both sets of commitments outline promises to conduct red-teaming exercises to identify severe threats, provide

transparency into frontier model capabilities and limitations, and build organization structures to promote the security of frontier systems.
The distinctive similarities between China's AI Safety Commitments and the Seoul Commitments could set the stage for further global convergence among industry actors on best practices at the upcoming Paris AI Action Summit, scheduled for mid-February. Only one Chinese company, Zhipu.ai, signed onto the Seoul Commitments at the Seoul summit last year. If more Chinese firms announce they will sign onto the Seoul Commitments in Paris, they could help lay the groundwork for a global industry-based international consensus on AI. This consensus could offer an informal pathway to build foundational rules of the road in international AI governance.

As Frontier AI Capabilities Grow, so do the Risks

The capabilities of frontier models in both the United States and China have rapidly improved over the last couple of months. In December, OpenAI's o3 model **blew past existing performance benchmarks**, demonstrating increases in AI capabilities against human performance across different domains. The capabilities boom was not just felt in Silicon Valley. In China, both DeepSeek and Moonshot, an Alibaba-sponsored start-up, **released open source reasoning models** that compete with OpenAI's o1 model—a significantly less powerful predecessor to o3—across a range of benchmarks. Other Chinese companies are also pushing the frontier. Tencent's Hunyuan model is "by some measures is world class" and possibly the world's best open weight model, Anthropic's Head of Policy Jack Clark **wrote** in November. And Alibaba's Qwen 2.5 model is only slightly behind. Its capabilities are also arguably world-class, **exceeding** GPT-4o and Claude 3.5 Sonnet across a range of benchmarks. Chinese AI companies are clearly producing models at AI's frontier, especially in the open-source domain.

These significant breakthroughs in model capabilities have been matched by increasing concern about risks—often documented by the companies themselves. Anthropic, in coordination with Redwood Research, showed that when the model believed its responses would be used in training, Claude **strategically deceived its creator** in 12 percent of cases, causing some to worry about future, more powerful AI systems going rogue. In a similar light,

OpenAI's o1 model attempted to deactivate its oversight mechanism in 5 percent of cases and had risks for chemical, biological, radioactive, and nuclear weapons rated as "medium." Concerns about risks at AI's frontier have caused leading thinkers from both the West and China, including recently departed national security adviser Jake Sullivan, to warn that great power coordination is needed to prevent catastrophic harm. So far, the U.S. and Chinese governments have agreed on little to how to regulate increasingly powerful generative AI models.

Concern about misuse of powerful models and the possibility AIs go rogue has motivated leading scientists from both the West and China to sign joint statements calling for guardrails on the technology and coordination between them to address frontier AI's challenges. Despite the calls for international consensus, so far U.S.-China coordination on frontier AI has lacked traction outside of an agreement from former president Joe Biden and President Xi Jinping on the need for a human to make decisions on nuclear weapons use.

The AI Summit Series Pushes for Industry-Led Consensus to Combat Risks

Given these concerns about risks at AI's frontier and the need for international coordination, the UK and South Korean governments have attempted to generate global consensus on risks and best practices by engaging with model developers themselves. At the second international AI Summit held in May of last year in Seoul, the British and South Korean governments announced the Frontier AI Safety Commitments (FAISCs), informally known as the Seoul Commitments. These industry-led efforts aim to establish best practices to prevent catastrophic risks that could materialize with increasingly powerful AI.

The FAISCs were designed to encourage companies to proactively identify redlines for concerning capabilities and establish plans for next steps if those risk thresholds were crossed. Achieving industry buy-in will be a critical part of governing frontier AI. Voluntary commitments offer flexible avenues to empower industry to promote transparency into their risk mitigation practices and eventually form an informal international industry-based consensus. While not a replacement for regulation, they offer a means of creating a floor

for establishing consensus on the measures needed to prevent the most serious risks at AI's frontier.

While there was widespread buy-in from leading Western companies including OpenAI, DeepMind, Anthropic, and Elon Musk's xAI, most Chinese companies were conspicuously absent from the list of signatories. Only one Chinese frontier AI company signed onto the commitments: Zhipu.ai, a start-up that **ranks third** in market share among China's business-facing LLM providers. The absence of Chinese firms was a substantial limitation of industry-led efforts: the utility of those commitments would be limited if they lacked buy-in from one of the two most important countries for frontier AI development. The lack of Chinese company participation was not necessarily due to lack of interest. Chinese companies were approached by the creators of the Seoul Commitments in the lead-up to the summit, when it became clear to them that securing buy-in from companies in the world's second frontier model AI power would be critical for international AI governance.

Chinese Industry Joins the Global Push for Safety Commitments

The lack of initial broad buy-in from Chinese companies into the Seoul Commitments makes the Chinese AI Safety Commitments **announcement** in December 2024 by think tank **China Academy for Information and Communications Technology** (CAICT, 中国信息通信研究院) especially important.

The content of China's AI Safety Commitments bears a strong resemblance to the Seoul Commitments that preceded them (see Table 1). They are both described as industry self-regulation efforts that focus on ensuring the safety and security of frontier AI models through efforts like red-teaming. They also actively push for measures to ensure transparency, both to governments and also to the public, about risks associated with their models. Both documents discuss the need for more foundational research on risks large models pose and the promise to deploy frontier models to solve global challenges.

Table 1. Points of Similarity in the Seoul Commitments and China's AI Safety Commitments

	Seoul Commitments	China's AI Safety Commitments
Assessing Risks Across the AI Lifecycle	"Assess the risks posed by their frontier models or systems across the AI lifecycle."	"Implement risk management practices throughout the entire AI development and deployment life cycle."
Implementing Organizational Structures to Promote Safety and Security	"Organisations are accountable for safely developing and deploying their frontier AI models and systems. They will: . . . develop[] and continuously review[] internal accountability and governance frameworks and assigning roles, responsibilities and sufficient resources to do so."	"Establish safety teams or organizational structures and build safety and security risk management mechanisms. Designate [a] leader responsible for AI safety and security, establish specialized teams to conduct AI risk assessments and safety governance within the enterprise."
Risk Mitigation	"Articulate how risk mitigations will be identified and implemented to keep risks within defined thresholds."	"Clearly outline processes and measures for risk identification and mitigation."
Transparency on Capabilities, Limitations, and Applications	"To publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use."	"Enhance model transparency. Proactively disclose safety and security governance measures and improve transparency for all stakeholders. Provide clear information about the model's capabilities, applicable fields, and limitations.
Transparency to the Public	"Provide public transparency on the implementation . . . except insofar as doing so would increase risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit."	"Inform potential risks to the public through model documentation, service agreements, or others."
Red-Teaming	"Internal and external red-teaming of frontier AI models and systems for severe and novel threats."	"Through dedicated simulation and red-teaming experts to rigorously test models prior to their release or update."
More Research	"To prioritize research on societal risks posed by frontier AI models and systems."	"Vigorously advance frontier safety and security research."

Deploying Models to Solve Challenges	"Develop and deploy frontier AI models and systems to help address the world's greatest challenges."	"Innovate in the development and deployment of AI systems that embody the principle of AI for good, contributing to addressing the most pressing challenges faced by society."

Source: "Frontier AI Safety Commitments, AI Seoul Summit 2024," UK Department for Science, Innovation, and Technology, May 21, 2024, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024; and "Protecting AI Security and Building a Model of Industry Self-Discipline - The First Batch of 17 Companies Signed the 'Artificial Intelligence Security Commitment,'" China Academy of Information and Communications Technology, December 24, 2024, https://mp.weixin.qq.com/s/s-XFKOCWhu0uve4opgb3Ng.

Although there are meaningful differences in the scope of their commitments, the similar specific commitments and sometimes identical language provide good reasons to believe China's AI Safety Commitments were directly inspired by Seoul. That seems to follow an increasingly familiar pattern in China's AI domain. China frequently foreshadows eventual convergence with international AI governance efforts by preemptively launching domestic projects coded in "Party-speak." For foundational governance concepts in any policy area in China, being able to trace ideas to distinctively Chinese roots is critical for the enduring legitimacy of the idea. A similar dynamic played out when China launched its **Global AI Governance Initiative** just two weeks before it signed onto the UK-led **Bletchley Declaration** in November 2023. China views its Global AI Governance Initiative as complementary to Bletchley, **which it frequently praises** for its positive role in international AI governance.

The signing of the commitments means sixteen new Chinese companies join the sixteen signatories of the Seoul Commitments to make AI safety promises, generating the potential for strong Chinese buy-in that the initial Seoul Commitments lacked. Between the signatories of the Seoul Commitments and Chinese AI Safety Commitments, most of the companies involved in producing powerful generative AI models have signed onto at least one of the two sets of commitments (see table 2).

Table 2: Companies Signing onto Seoul Commitments and China's AI Safety Commitments

Seoul Commitments Signatories (May 2024)	Chinese AI Safety Commitments Signatories (December 2024)
Amazon	01.AI
Anthropic	360
Cohere	4Paradigm
Google	Alibaba
G42	Ant Group
IBM	Baidu
Inflection AI	China Mobile
Meta	China Telecom
Microsoft	DeepSeek
Mistral AI	Huawei
Naver	iFlyTek
OpenAI	Megvii
Samsung Electronics	MiniMax
Technology Innovation Institute	SenseTime
XAI	Tencent
Zhipu.ai	Volcdance
	Zhipu.AI

Source: "Frontier AI Safety Commitments, AI Seoul Summit 2024," UK Department for Science, Innovation, and Technology, May 21, 2024, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024; and "Protecting AI Security and Building a Model of Industry Self-Discipline - The First Batch of 17 Companies Signed the 'Artificial Intelligence Security Commitment,'" China Academy of Information and Communications Technology, December 24, 2024, https://mp.weixin.qq.com/s/s-XFKOCWhu0uve4opgb3Ng.

Yet there are differences in the substance of these two agreements. The Chinese AI Safety Commitments offer a more comprehensive set of security requirements for data and critical infrastructures and have a specific line on adopting appropriate safety measures for open source initiatives, where Chinese products like DeepSeek's are especially competitive internationally. The Seoul Commitments, in contrast, are fundamentally focused on establishing redlines for frontier risks that should not be crossed and identifying steps companies would take if those risk thresholds were reached. This difference could be explained by the fact that China is focused on using

AI as a means of boosting its economy. Redlines could be seen as an overly restrictive policy tool at a time when the government wants to signal it is embracing AI and not overregulating it due to concerns about the lasting effects of its COVID-era tech crackdown.

And there is a broader structural difference between Chinese-led and international efforts that has important implications for Chinese AI governance: Chinese industry has more explicit government-industry connections than Western companies. This makes the implicit backing of key government stakeholders that China's AI Safety Commitments received critical. CAICT is a leading Chinese think tank housed under the state's Ministry of Industry and Innovation Technology (MIIT). MIIT's institutional backing lends the promises particular weight in China's regulatory landscape, where industry initiatives typically align closely with government priorities. Moreover, the effort was explicitly spearheaded by China's **Artificial Intelligence Industry Alliance** (AIIA), a prominent industry consortium guided by MIIT.

This industry backing could foreshadow future Chinese AI policy. Historically, AIIA involvement has presaged future Chinese regulation. In 2019 and 2020, AIIA working groups led by CAICT officials **offered recommendations** and created the Industry Self-Discipline Joint Pledge (行业自律公约). These recommendations **formed the foundation** for foundational Chinese regulations on recommendation algorithms and deepfakes. Even if Chinese industry lacks the direct authority to shape its domestic regulatory environment, it could be empowered to match Western firms' efforts and foreshadow more Chinese firms signing onto the Seoul Commitments.

The Paris AI Action Summit Offers a Key Juncture to Build Industry Consensus

DeepSeek's rapid ascent highlights a growing challenge for U.S. policymakers: Chinese models are increasingly powerful, and, without safeguards, they could threaten U.S. national security and economic interests.

This challenge has intensified as both the U.S. and Chinese governments have backed the AI companies at the vanguard of frontier development. The

highest echelons of the Chinese government have taken a clear interest in DeepSeek. Premier Li Qiang's **invitation to DeepSeek CEO Liang Wenfeng** to provide feedback on the draft Government Work Report underscored growing high-level government interest in China's frontier AI development. And the Chinese government **announced an $8.2 billion AI investment fund**, adding financial resources to power China's cost-efficient models.

The United States, meanwhile, has redoubled efforts to "win the AI race." President Donald Trump **announced** a private AI infrastructure joint venture, Stargate, which will pool resources from OpenAI, Softbank, and Oracle. Leading U.S. AI companies are ready to race. Scale AI put out a full-page ad congratulating Trump on his election victory, saying, **"Dear President Trump, America must win the AI war."**

Despite accelerating U.S.-China frontier AI competition, the Paris AI Action Summit in February could be a turning point for industry consensus on frontier AI risks. If more Chinese companies sign onto the Seoul Commitments and publish safety frameworks alongside their Western counterparts, it will breathe new energy into industry-led safety commitments as a path toward global AI risk mitigation. As the United States and China struggle to build traction for an agreement at the government level, companies on both sides of the Pacific could erect temporary scaffolding on frontier AI governance until the governments build a permanent structure.