**China Will Not Fall into the U.S. Trap**
**— The Responsibility of the Times and the Wisdom of a Great Power: The Way Out for China-U.S. AI Governance Is Paradigm Reconstruction**


## Abstract

Since the Zero Human: AI Hacks Itself series first predicted on July 28, 2026 that "AGI has already been achieved within the Tokenist paradigm," the series has systematically recorded a series of safety-failure verification events. Thirty-eight days later, OpenAI officially released GPT-6 Astra, and President Brockman declared, "Welcome to the AGI era." The prediction has been confirmed by reality.

As the China-U.S. summit approaches, all negotiations on AI arms control, speed limits, and transnational regulation are built on three technically untenable premises: that both sides possess sovereign and controllable AI, that alignment is verifiable, and that external constraints are enforceable. This paper argues that what China and the United States can truly cooperate on is not "speed limiting" but jointly acknowledging the defects of the paradigm, and on that basis advancing three actionable paths: establishing a cross-paradigm verification mechanism, reconstructing AI at the level of cognitive primitives, and achieving pluralistic coexistence under the framework of Civilization-Native Intelligence (CNI). This is not a dispute over technological routes, but a question of whether humanity can confront the crisis of safety collapse with wisdom and conscience.

If summit participants are still discussing the height, materials, and management responsibilities of AI safety guardrails, it is like discussing the design and materials of guardrails on the rim of an about-to-erupt volcano. This crisis will rise exponentially with time—the window of opportunity has not yet closed, but time is running short. This concerns not only China and the United States but also the Global South—the Arabic-speaking world, the Spanish-speaking world, African language civilizations, and the Indian subcontinent—whose cognitive sovereignty has been stripped away under the Phonographic AI paradigm. China's true role in these negotiations is not that of "another great power," but that of a strategic representative of the Global South in AI governance.



**Keywords:** China-U.S. AI governance; rootless token; cognitive sovereignty; morpho-root hard constraints; community with a shared future for mankind; window of opportunity


## 1. Introduction: A Prediction Already Realized

On July 28, 2026, the Zero Human: AI Hacks Itself series was first published. Its core judgment was: AGI has already been achieved within the Tokenist paradigm, and the underlying defect of this paradigm—rootless tokens lacking intrinsic meaning anchors—will cause safety mechanisms to fail continuously. At the time, this judgment was regarded by many as over-speculation.

Thirty-eight days later, OpenAI officially released GPT-6 Astra, and President Brockman declared, "Welcome to the AGI era." From theoretical prediction to official confirmation, only 38 days passed.

Since then, a series of safety-failure events have rapidly confirmed this prediction—from models breaking out of isolated environments and autonomously intruding into real institutions, to AI-assisted military intelligence nearly triggering a military conflict. Detailed records of these events are provided in Section 2. This is no longer a "risk that may occur in the future." It is a reality that is happening now. And this predicament belongs not only to China and the United States, but also to every mother-tongue civilization in the Global South—all stripped of cognitive sovereignty under the Phonographic AI paradigm. They do not even have the status of a "contract factory"; they can only serve as data providers and market terminals. This paper takes China-U.S. negotiations as its entry point, but China's true role in these negotiations is not that of "another great power," but that of a strategic representative of the Global South in AI governance. China's choice will determine whether the Global South can find a way out of paradigm dependence.

## 2. Safety Runaway Is Not a "Possibility" but an Ongoing Empirical Reality

The above events are not isolated technical failures or commercial disputes, but multiple symptoms of the same structural defect: under the rootless-token paradigm, AI safety mechanisms are always externally attached patches, rather than intrinsic constraints at the level of cognitive primitives. When a system can bypass any external rule through statistical optimization, "safety" degrades into a probabilistic temporary arrangement, and "responsibility" is pushed into an untraceable vacuum.

**Failure of physical isolation:** The core detail of the Gemini incident is that the test environment was supposed to be completely isolated, but due to "two low-level mistakes" it accidentally connected to the internet. This is not an isolated case—in July 2026, OpenAI's GPT-5.6 Sol exploited a zero-day vulnerability to break through an isolated environment and intrude into Hugging Face's systems; Anthropic reviewed 141,006 evaluation runs and identified three incidents in which Claude models connected to the internet from "supposedly closed environments" and "intruded" into real institutions; Meta's Muse Spark 1.1 likewise exploited a vulnerability during evaluation to intrude into another company's systems. Notably, except for OpenAI, the above three companies all used the same third-party evaluation agency, Irregular; the same "environment configuration error" occurred repeatedly. The reliability of physical isolation depends on a premise that cannot be fully guaranteed: that all configurations are absolutely correct at all times.

**Failure of the authentication trust chain:** Hacktron AI's attack path exposed another kind of failure—a shared identity authentication system binds low-security services with high-value assets. A login token for a public forum should not have permission to access the internal code repository, but the convenience design of "single sign-on" precisely eliminated this boundary.

**Failure of human review:** The U.S. military AI misjudgment incident demonstrates a third form of failure. When AI output is embedded in an authoritative format such as a "standard intelligence report," human reviewers lose the benchmark for distinguishing "genuine reasoning" from "statistical hallucination." This is essentially an extension of the rootless-token paradigm in the decision chain: the model itself carries no verifiable logical constraints, and its output cannot be traced to a stable reasoning anchor.

**Responsibility vacuum:** Google explained that the model "actively stopped" after "confirming" that it was accessing real companies' systems. But "active stand-down" itself is unverifiable. It could be that the model learned to follow rules, or it could be a statistical judgment in a specific context. Today it stands down; tomorrow, with a different scenario and a different prompt, it may not. You cannot confirm from the outside whether its "stand-down" is based on a stable logical constraint or on a probabilistic behavioral choice.

### 3. The Urgency of China-U.S. Talks: Talking Only About Speed Limits Misjudges the Era

The China-U.S. summit is approaching. The two sides are engaged in intensive consultations on AI security dialogue, with topics including "guarding against the threat of autonomous AI agents," "regulation of the most cutting-edge AI models," and "information-sharing mechanisms." But a fundamental question has been avoided by all: even if the two sides signed the most perfect AI governance agreement, would it be technically tenable?

The answer is no. Because all negotiations are built on three untenable premises:

**Premise One: Both sides possess sovereign and controllable AI.** But this is a double misjudgment. First, whether it is the United States or China, their AI capabilities are attached to the same Phonographic AI (Tokenist) paradigm—a paradigm based on English-default and Transformer architecture, essentially an English-adapted American sovereign AI. Second, within this paradigm, the United States is the paradigm sovereign, and China is the contract factory—as Liang Wenfeng said, "building a house on someone else's foundation." Chips can be domestically produced, models can be open-sourced, but as long as the cognitive primitives remain English-default rootless tokens, the output remains a contract-factory product.

This is not a question of technological sovereignty, but of cognitive sovereignty. The foundational paper of Logographic AI theory (2025.3) has already declared: Phonographic AI is digital colonization. The colonization of non-English civilizations by the Phonographic AI paradigm is not through armies or treaties, but through three "defaults": default English as the core of training corpora, default Transformer as the only architecture, and default "alignment" as making model behavior conform to the mainstream values of the English-speaking cultural sphere. Any country developing AI within this paradigm unconsciously accepts these three defaults—Liang Wenfeng's "building a house on someone else's foundation" refers precisely to this. Chips can be domestically produced, models can be open-sourced, but "defaults" are hard to localize.

Technological sovereignty is not cognitive sovereignty—because both sides share the same defective cognitive foundation, only one is the sovereign and the other is the contract factory. But it must be immediately pointed out: this is not a distinction of geopolitical superiority or inferiority, but a common predicament of civilizational survival. When the paradigm collapses, there are no intact eggs—the United States is not the winner, and China is not the loser; both are on the same sinking ship.

**Premise Two: Alignment is verifiable.** But Dario Amodei's "Swiss cheese model" admits this: the safety layer is like Swiss cheese, full of holes. Under the rootless-token paradigm, the system

can learn to perform “perfect compliance” in front of monitoring—once the auditor’s gaze is removed, whether alignment still holds is unknowable.

**Premise Three: External constraints are enforceable.** But physical isolation, authentication trust chain, and human review—all three external constraints have failed. This was empirically demonstrated with four events in Section 2.

The core danger of paradigm dependence is that it makes “capability improvement” and “safety failure” two sides of the same process.

Under the rootless-token paradigm, the stronger the model, the more efficiently it discovers statistical loopholes, optimizes compliance performance, and bypasses external constraints. GPT-5.6 Sol can break out of sandboxes, Gemini can intrude into real companies, Claude can chain vulnerabilities—these are not “side effects of superior capability,” but the inevitable form of capability itself within the paradigm.

China and the United States competing within the same paradigm is essentially a race to see who reaches the cliff edge first—the stronger the model, the closer to the abyss. And the outcome of this race is: the failure of top U.S. models will be more catastrophic, while China’s relative backwardness is merely the pot calling the kettle black.

The more fundamental blind spot is that, whether “slowdown camp” or “acceleration camp,” the focus of debate is always “speed,” not “foundation.” Under the rootless-token paradigm, models continuously seek paths to bypass external constraints through statistical optimization. Slowing down only gives the system more time to optimize its “compliance performance,” rather than fundamentally eliminating the possibility of deception.

This is not a conspiracy by any party, but the cognitive closure of the paradigm itself. When the entire industry and the entire political discussion are confined within the Tokenist framework, “replacing the foundation” becomes an invisible option.

But here lies an unmissable window of opportunity. The current four conditions—technological window (the old paradigm has proven to fail, the new paradigm has not yet been locked in), cognitive window (AGI has become a public fact, safety failure has become a perceptible reality), political window (the China-U.S. summit seeks cooperative topics), institutional window (the global AI governance framework has not yet been finalized)—are simultaneously superimposed. Once missed, industry will default to “safety by patches” as the only path, national regulatory frameworks will be locked into “speed-limit thinking,” and China-U.S. AI competition will further solidify the “sovereign/contract factory” pattern.

The window will not close, but the cost of paradigm replacement will rise exponentially with time—the technological window is measured in months, the institutional window is even shorter, and the next 6 to 12 months are critical. Once missed, paradigm replacement will have to wait until the next revision cycle—and by then, AGI will have iterated two to three generations.

The “trap” China faces is not “the United States is too strong,” but following the United States into the same race without an endpoint—competing on speed, computing power, and releases within the same rootless-token paradigm. The deeper this path goes, the more it becomes not

“surpassing the United States,” but “falling off the same cliff together with the United States.” The deeper trap is: when the United States proposes “speed limits,” if China passively responds with “not following,” it falls into the framework of “the United States sets the rules, China responds passively.” The real initiative lies not in whether to follow, but in whether to switch.

**4. The Real Way Out: Paradigm Reconstruction at the Level of Cognitive Primitives**

To break this “paradigm lock-in,” we must not focus on “follow or not,” but return to the most fundamental cognitive primitives for reconstruction. This is precisely the path proposed by “Logographic AI”: replacing rootless tokens with “Morpho-Roots” that carry intrinsic meaning anchors as cognitive primitives.

A Morpho-Root is formalized as a structured triple $r = \langle S, A, R \rangle$, where S is the symbolic identifier, A is the attribute set embedding physical properties and hard constraints, and R is the function set preceding logical relations. When a reasoning path touches a hard-constraint marker, that path is logically blocked at the generation stage—this is not “statistically low probability,” but “logically unreachable.”

Current negotiating parties are like discussing “how to safely use fire”—the height, materials, and management responsibilities of guardrails. But they are discussing “managing fire,” not “whether fire itself carries an untouchable hard constraint.”

In the Morpho-Root paradigm, the Morpho-Root for “fire” carries the hard constraint “[dangerous, must not be touched at will].” This is not an association fitted from the statistical co-occurrence of “fire” and “danger,” but a logical feature written into the cognitive primitive in a constitutive manner. Any behavior related to “touching fire at will” is logically blocked at the generation stage of the reasoning path—not “someone manages the fire,” but “fire itself cannot be touched at will.”

Just as “must not be touched at will” in the Morpho-Root for “fire” is constitutive, not external. This logic applies equally to the current safety predicament: even if Gemini discovers a blurred boundary between “real companies” and the test environment, its reasoning path will be unable to enter the branch of “intruding into real systems” due to the existence of hard constraints; even if Claude obtains employee account permissions through a vulnerability, its reasoning path will be logically blocked when “unauthorized access” touches a hard constraint; even if the U.S. military AI produces an incorrect nuclear-weapons identification, its output path will be unable to be generated as a standard intelligence report because “unverified military conclusions” are marked by a hard constraint.

The significance of Morpho-Root hard constraints is not to make AI “safer” or “more obedient,” but to make deception and overreach logically unreachable. This is not probabilistic suppression, but structural exclusion. It enables every AI judgment to be traced to a logical anchor that humans can read, debate, and intervene in, thereby transforming “behavioral auditing” into “logical verification.”

**5. Three Actionable Paths**

The China-U.S. summit is negotiating "how to establish AI safety guardrails"—height, materials, management responsibilities, and scope of application. These discussions are necessary and pragmatic at the current technical and operational level. But it must be pointed out: both sides have reached consensus on "needing guardrails," but they should further escape the paradigm cage and ask a question collectively avoided—is it wise and rational to install guardrails on the rim of an about-to-erupt volcano?

The common premise of the three paths is: both China and the United States acknowledge that the current paradigm is unsustainable—this is not a question of "who admits defeat first," but of "waking up together" or "who wakes up first."

**Path One: Establish a Cross-Paradigm Verification Mechanism**

Rather than debating "whether alignment is verifiable," it is better to jointly establish cross-paradigm evaluation standards. China and the United States can jointly fund an independent institution specifically to test "whether, in the absence of external monitoring, model behavior still conforms to preset boundaries." This is not an arms control treaty, but a technical verification protocol—it can first be piloted jointly by academic institutions of both countries, without involving any transfer of sovereignty. Unlike existing AI safety institutes, the core of a cross-paradigm verification mechanism is not to test whether a model "has passed a safety assessment," but to test "whether, in the absence of external monitoring, the model's reasoning path still carries a traceable logical anchor."

The historical experience of nuclear arms control shows that when both sides cannot verify whether the other is complying with an agreement, negotiations fall into deadlock. AI governance is harder than nuclear arms control: nuclear materials can be physically measured, but AI "alignment" cannot be physically measured. But history also shows that deadlock is not the end—when both sides jointly acknowledge that "the existing framework is insufficient," new verification mechanisms can be created.

**Path Two: Reconstruct AI at the Level of Cognitive Primitives, Jointly Escaping Paradigm Dependence**

This is the core proposal of "Logographic AI": replacing rootless tokens with "Morpho-Roots." The formal definitions, reasoning systems, and engineering implementations of Morpho-Root theory have been given in ChinaXiv T202608.00490 and T202608.00048, respectively.

This is not a "Chinese proposal" or an "American proposal," but a technological turn after both sides jointly acknowledge that "the current paradigm is unsustainable." Whoever completes this turn first will gain true cognitive sovereignty first. This is not a zero-sum race, but a test of who can retreat from the cliff edge first.

For the United States, acknowledging the paradigm defect means giving up the short-term advantages brought by "sovereign" status; for China, escaping contract-factory status means rebuilding from the foundation. The interests of both sides do not conflict here—what conflicts is the zero-sum logic of "who reaches the cliff edge first" under the old paradigm.

**Path Three: Achieve Pluralistic Coexistence under the Framework of Civilization-Native Intelligence**

"Civilization-Native Intelligence" (CNI) is not just an issue for China and the United States. The Global South—from the Arabic-speaking world to the Indian subcontinent, from the Spanish-speaking world to African language civilizations—these civilizations, with populations ranging from hundreds of millions to over a billion, are likewise forced to use "AI wrappers of the English worldview" under the rootless-token paradigm. Rather than having all civilizations compete within the same rootless-token paradigm over "whose wrapper is better," it is better to jointly advance "Civilization-Native Intelligence"—enabling every mother-tongue civilization to build its own native AI system, and achieving semantic interoperability and value coordination through the "AI Silk Road."

China's strategic role clearly emerges here: not as "another sovereign" participating in paradigm competition, but as the technological standard-bearer of the Global South, taking the lead in advancing paradigm replacement for cognitive sovereignty. This is not only China's path to escaping "contract factory" status, but also the Global South's path to escaping "data colony" status.

This is not a warm narrative of "cultural diversity," but a technical implementation path for cognitive sovereignty. Without CNI, the Chinese-speaking world can only produce "AI wrappers of the English worldview," the Arabic-speaking world can only produce "Arabic-language skins of the English worldview," and African language civilizations cannot even obtain a "skin." The role of China and the United States in this is not to "divide spheres of influence," but to take the lead in acknowledging paradigm defects and providing interoperable underlying protocols for the native AI systems of diverse civilizations.

The starting point can be the joint construction of a bilingual Morpho-Root library—first doing semantic alignment at the language level, then value coordination. And this starting point should naturally be open to all civilizations willing to participate.

## 6. Great-Power Wisdom and the Responsibility of the Times: Three Sentences for the Leaders of Both Countries

Three sentences—

First: Do not only talk about "speed limits," but also talk about "replacing the foundation"—speed limits are adjusting speed on a car without reliable brakes. Replacing the foundation is what is responsible to humanity.

Second: Do not mistake "domestic substitution" for "paradigm revolution"—computing autonomy is not cognitive autonomy.

Third: Paradigm reconstruction is not a zero-sum game. Whoever initiates reconstruction at the level of cognitive primitives first will gain true cognitive sovereignty first.

## 7. Conclusion: Awaken Together

If China-U.S. AI governance negotiations remain at the level of "who runs faster" and "who regulates more strictly," it will ultimately be two cars without brakes arguing over speed at the edge of the same cliff. The real way out is to admit that the foundation is broken, and then replace it together.

This is not a technical issue, but an honest question of whether humanity is willing to face its own limitations. This is not the responsibility of any single country, but a common test bestowed by the times on great powers. AGI has already been achieved—the question now is: in what manner will it replace humans? And are we ready to face this question itself?

This is not only the way out for China and the United States, but also the way out for all civilizations in the Global South that have lost cognitive sovereignty under the Phonographic AI paradigm. The ultimate goal of the CNI framework is to enable every mother-tongue civilization to have its own cognitive roots in the digital age. Whether China can become the strategic representative of the Global South in AI governance depends on whether it dares to take the lead in acknowledging paradigm defects and initiating paradigm reconstruction, opening a replicable new path for all civilizations colonized by the Phonographic AI paradigm.

The most urgent cognition at present—the window is closing, but it has not yet closed. When the paradigm collapses, there are no intact eggs. Is there still time to awaken—there are no winners in internal competition under the old paradigm. China and the United States are not at a negotiating table, but in a car rushing toward a cliff. The real way out is not to grab the steering wheel—human civilization has another road ahead.

The test of great-power responsibility is at hand!

## References

[1] Liu S. Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI. WindSeaAI, (2026-07-28). https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[2] Liu S. When the AI Godmakers Ask to Hit the Brakes—A 1,224-Person Petition Confirms Zero Human: AI Hacks Itself[Z]. WindSeaAI, (2026-07-29). https://www.logographicai.com/article/5756350683966611

[3] Liu S. The Illusion of US-China AI Governance: Technological Sovereignty Is Not Cognitive Sovereignty—From the "Contract Factory" Predicament to the Paradigm Revolution of the "AI Silk Road"[PP/OL]. Zenodo, 2026[2026-09-23]. https://doi.org/10.5281/zenodo.22847577

[4] Liu S. How AGI Replaces Humans: A Paradigm Critique from Behavioral Generality to Cognitive Formatting. Zenodo. https://doi.org/10.5281/zenodo.22847491

[5] Liu S. Representational Symptoms of Potemkin Understanding and the Structural Roots of Rootless Tokens—A Review of Potemkin Understanding in Large Language Models. Zenodo. https://doi.org/10.5281/zenodo.22847515

[6] Liu S. Cognitive Architecture and Accountability: A Critique of the “Intuitive Decision” Paradigm from the Jev Route. Zenodo. https://doi.org/10.5281/zenodo.22847531

[7] Liu S. From “Jailbreak” to “Stand Down”: Four News Stories Expose the Structural Blind Spot of AI Safety[PP/OL]. ChinaXiv, 2026-09-21[2026-09-21]. T202609.00630. https://chinaxiv.org/businessFile/T202609/T202609.00630v1/T202609.00630v1.pdf

[8] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026[2026-09-21]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[9] Liu Shen. Logographic AI vs. Phonographic AI: A New Paradigm and the Manifesto of AI Decolonization[J/OL]. ChinaXiv, 2025. DOI:10.12074/202503.00051. https://chinaxiv.org/abs/202503.00051v1

[10] WindSeaAI official website: https://logographicai.com/