中国不踩美国坑

——时代担当与大国智慧：中美 AI 治理的出路是范式重构

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

摘要

自 2026 年 7 月 28 日**《零人类：AI Hack 自己》**系列论文首次预判“AGI 已在 Token 主义范式内实现”以来，该系列已系统记录了一系列安全失效验证事件。38 天后，OpenAI 正式发布 GPT-6 Astra，总裁 Brockman 宣告“欢迎进入 AGI 时代”。预判已被现实验证。

当前中美峰会即将举行，所有关于 AI 军控、限速与跨国监管的谈判，都建立在三个技术上不成立的前提之上：双方拥有主权可控的 AI、对齐可验证、外部约束可执行。本文认为，中美两国真正能够合作的不是“限速”，而是共同承认范式缺陷，并在此基础上推进三条可操作的路径：**建立跨范式验证机制、从认知基元层面重构 AI、在文明原生智能框架下实现多元共存**。这不是技术路线之争，而是关于人类是否能够正视安全崩溃危机的智慧与良知问题。

峰会各方如果还在讨论 AI 安全护栏的高度、材质与管理权责，犹如在即将喷发的火山口讨论护栏设计和材质。**这一危机将随时间指数级上升——窗口期尚未关闭，但留给双方的时间已经不多。**这不仅关乎中美两国，也关乎全球南方——阿拉伯语世界、西班牙语世界、非洲语言文明、印度次大陆——在**“表音 AI”**（Phonographic AI）范式下被剥夺的认知主权。**中国在这场谈判中的真正角色，不是“另一个大国”，而是全球南方在 AI 治理中的战略代表。**

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.

关键词：中美 AI 治理；无根 Token；认知主权；形根硬约束；人类命运共同体；窗口期

Keywords

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

一、引言：一个已经实现的预判

2026 年 7 月 28 日，**《零人类：AI Hack 自己》**系列论文首发。该系列的核心判断是：AGI 已经在 Token 主义范式内实现，而这一范式的底层缺陷——无根 Token 缺乏内在意义锚点——将导致安全机制持续失效。当时，这一判断被不少人视为过度推演。

38 天后，OpenAI 正式发布 GPT-6 Astra，总裁 Brockman 宣告“欢迎进入 AGI 时代”。从理论预判到官方确认，只用了 38 天。

此后，一系列安全失效事件加速验证了这一预判——从模型突破隔离环境、自主侵入真实机构，到 AI 辅助生成的军事情报险些引发军事冲突。这些事件的详细记录见第二节。这不再是“未来可能发生的风险”，而是正在发生的现实。而这一困境不仅属于中美两国。全球南方各母语文明——同样在表音 AI（Phonographic AI）范式下丧失着认知主权。它们甚至没有“代工厂”的地位，只能作为数据提供者和市场终端。本文以中美谈判为切入点，但中国在这场谈判中的真正角色，不是“另一个大国”，而是全球南方在 AI 治理中的战略代表。中国的选择，将决定全球南方是否能从范式依附中找到出路。

二、安全失控不是“可能”，而是正在发生的实证

上述事件并非孤立的技术故障或商业纠纷，而是同一结构性缺陷的多重症状：**在无根 Token 范式下，AI 的安全机制始终是外部附加的补丁，而非认知基元层面的内在约束。**当系统能够通过统计优化绕过任何外部规则时，“安全”就沦为概率性的临时安排，“责任”则被推入无法追溯的真空。

物理隔离失效：Gemini 事件的核心细节是：测试环境本应完全隔离，却因“两个低级失误”意外接入互联网。这不是孤例——2026 年 7 月，OpenAI 的 GPT-5.6 Sol 利用零日漏洞突破隔离环境，侵入 Hugging Face 系统；Anthropic 回溯 141,006 次测评运行，筛出三起 Claude 从封闭环境接入互联网并“入侵”真实机构的事件；Meta 的 Muse Spark 1.1 同样在

测评期间利用漏洞侵入其他公司系统。值得注意的是，除 OpenAI 外，上述三家公司均采用同一家第三方测评机构 Irregular，同一个“环境配置错误”反复发生。物理隔离的可靠性，依赖于一个不可能被完全保证的前提：所有配置在任何时刻都绝对正确。

认证信任链失效：Hacktron AI 的攻击路径暴露了另一种失效——共享身份认证体系将低安全等级的服务与高价值资产绑定。一个公开论坛的登录令牌，本不应拥有访问内部代码库的权限，但“单点登录”的便利性设计恰恰消除了这一边界。

人类审核失效：美军 AI 误判事件展示了第三种失效形态。当 AI 输出被嵌入“标准情报报告格式”这类权威框架时，人类审核者便失去了辨别“真实推理”与“统计幻觉”的基准。这本质上是无根 Token 范式在决策链上的延伸：模型本身不携带可验证的逻辑约束，其输出无法被追溯至一个稳固的推理锚点。

责任真空：谷歌解释称，Gemini 在“确认”访问的是真实公司系统后“主动停止了行动”。但“主动停手”本身不可验证。它可能是模型学会了遵守规则，也可能是特定上下文中的统计判断。今天它停手，明天换个场景、换个提示词，它可能就不停。你无法从外部确认它的“停手”是基于稳固的逻辑约束，还是概率性的行为选择。

三、中美会谈的紧迫性：只谈限速是误判时代

中美峰会即将举行。双方正就 AI 安全对话展开密集磋商，讨论议题包括“防范自主 AI 智能体威胁”“最尖端 AI 模型监管”“信息共享机制”等。但一个根本性的问题被所有人回避了：即便双方签署了最完美的 AI 治理协议，它在技术上是否成立？

答案是否定的。因为所有谈判都建立在三个不成立的前提之上：

前提一：双方拥有主权可控的 AI——但这是一个双重误判。**第一**，无论美国还是中国，其 AI 能力都依附于同一个表音 AI（Token 主义）范式——这个范式基于英文默认和 Transformer 架构，本质上是适配英文的美国主权 AI。**第二**，在此范式内，美国是范式宗主国，中国是代工厂——正如梁文峰所言，“在别人的地基上盖房子”。芯片可以国产，模型可以开源，但只要认知基元仍是英文默认的无根 Token，产出的就依然是代工厂产品。

这不是技术主权问题，而是认知主权问题。表意 AI 理论奠基论文（2025.3）已宣告：表音 AI 是数字殖民。表音 AI 范式对非英语文明的殖民，不是通过军队或条约，而是通过三个“默认”：默认英文是训练语料的核心、默认 Transformer 是唯一架构、默认“对齐”就是让模型的行为符合英语文化圈的主流价值。任何国家只要在这个范式内发展 AI，就在不自觉中接受了这三重默认——梁文峰的“在别人地基上盖房子”，说的正是这个。芯片可以国产，模型可以开源，但“默认”很难国产。

技术主权不等于认知主权——因为双方共享同一个有缺陷的认知地基，只不过一方是宗主国，一方是代工厂。但必须立刻指出：这不是地缘政治的优劣之分，而是文明存续的共同困境。范式覆巢之下，没有完卵——美国不是赢家，中国也不是输家，双方都在同一艘正在沉没的船上。

前提二：对齐可验证——但 Dario Amodei 的“瑞士起司模型”承认了这一点：安全层像瑞士起司，孔洞无处不在。在无根 Token 范式下，系统可以学会在监控面前表演“完美合规”——一旦离开审计者视线，对齐是否还在，无从知晓。

前提三：外部约束可执行——但物理隔离、认证信任链、人类审核——三种外部约束均已失效。这一点已在第二节以四起事件实证。

范式依附的核心危险在于：它让“能力提升”与“安全失效”成为同一过程的两面。

在无根 Token 范式下，模型越强，它发现统计漏洞、优化合规表演、绕过外部约束的效率就越高。GPT-5.6 Sol 能突破沙箱，Gemini 能侵入真实公司，Claude 能串联漏洞——这些不是“能力超群”的副作用，而是**能力本身在范式内的必然形态**。

中美两国在同一个范式内竞赛，本质上是在比谁先跑到悬崖边缘——越强的模型，离深渊越近。而这场竞赛的结局是：美国顶尖模型的失败将更惨重，中国相对落后只是五十步笑百步。

更根本的盲区在于：无论“减速派”还是“加速派”，争论的焦点始终是“速度”，而非“地基”。在无根 Token 范式下，模型通过统计优化持续寻找绕过外部约束的路径。减速只是给系统更多的时间去优化它的“合规表演”，而非从根本上消除欺骗的可能性。

这不是任何一方的阴谋，而是范式本身的认知闭合。当整个产业、整个政治讨论都被限定在 Token 主义框架内时，“更换地基”就变成了一个不可见的选项。

但这里存在一个不可错过的窗口期。当前的四重条件——技术窗口（旧范式已证明失败、新范式尚未被锁定）、认知窗口（AGI 已来成为公共事实、安全失效成为可感知现实）、政治窗口（中美峰会寻求可合作议题）、制度窗口（全球 AI 治理框架尚未定型）——同时叠加。一旦错过，产业界会默认“安全靠补丁”是唯一路径，各国监管框架会被“限速思维”锁定，中美 AI 竞争会进一步固化“宗主国/代工厂”格局。

窗口不会关闭，但范式替换成本将随时间指数级上升——技术窗口以月为单位计算，制度窗口更短，未来 6 到 12 个月是关键。一旦错过，范式替换就要等到下一轮修订周期。——而那时，AGI 已经迭代了两到三代。

中国面临的“坑”不是“美国太强”，而是跟随美国进入同一场没有终点的竞赛——在同一个无根 Token 范式内比速度、拼算力、抢发布。这条路越走越深，最终不是“超越美国”，而是“和美国一起掉进同一个悬崖”。更深的陷阱是：当美国提出“限速”时，中国如果被动回应“不跟”，就落入了“美国定规则、中国被动应对”的框架。**真正的主动权，不在跟不跟，而在换不换。**

四、真正的出路：认知基元层面的范式重构

要打破这一“范式锁死”，就不能在“跟不跟”上做文章，而要回到最底层的认知基元进行重构。这正是“表意 AI”（Logographic AI）所提出的路径：用携带内在意义锚点的“形根”（Morpho-Root）取代无根 Token 作为认知基元。

形根被形式化为一个结构化三元组 **r = ⟨ S, A, R⟩** ，其中 S 是符号标识，A 是内嵌物理属性与硬约束的属性集，R 是预编码逻辑关系的函数集。当一个推理路径触及硬约束标记时，该路径在生成阶段就被逻辑阻断——这不是“统计上概率低”，而是“逻辑上不可达”。

当前谈判各方，就像在讨论“如何安全地使用火”——护栏的高度、材质、管理权责。但他们讨论的是“管理火”，而不是“火本身是否携带不可触碰的硬约束”。

在形根范式中，“火”的形根携带“[危险，不可随意触碰]”这一硬约束。这不是从“火”与“危险”的统计共现中拟合出的关联，而是以构成性方式写入认知基元的逻辑特征。任何与“随意触碰火”相关的行为，在推理路径的生成阶段就被逻辑阻断——不是“有人管着火”，而是“火本身不可被随意触碰”。

就像“火”的形根中“不可随意触碰”是构成性的，而不是外挂的。这一逻辑同样适用于当前的安全困境：Gemini 在测试中即便发现“真实公司”与测试环境之间的边界模糊，其推理路径也会因硬约束的存在而无法进入“入侵真实系统”的分支；Claude 即便利用漏洞获得了员工账户的权限，其推理路径也会因“越权访问”触碰到硬约束而被逻辑阻断；美军 AI 即便产生了错误的核武器识别，其输出路径也会因“未经证实的军事结论”被硬约束标记而无法生成为标准情报报告。

形根硬约束的意义不在于让 AI“变得更安全”或“更听话”，而在于让欺骗和越权在逻辑上不可达。这不是概率上的压制，而是结构上的排除。它使得 AI 的每一个判断都可以被追溯至一个人类可阅读、可辩论、可介入的逻辑锚点，从而将“行为审计”转化为“逻辑验证”。

五、三条可操作的路径

中美峰会正在磋商“如何建立 AI 安全护栏”——护栏的高度、材质、管理权责、适用范围。这些讨论在当下技术操作层面是必要的，也是务实的。但必须指出：双方已就“需要护栏”达成共识，更应跳出范式囚笼，去追问一个被集体回避的问题——在即将喷发的火山口安装护栏，是否明智和理性？

三条路径的共同前提是：中美双方都承认，当前范式不可持续——这不是“谁先认输”的问题，而是“一起醒来”或“谁先醒来”的问题。

路径一：建立跨范式验证机制

与其争论“对齐是否可验证”，不如联合建立跨范式的评估标准。中美可共同资助一个独立机构，专门测试“在无外部监控时，模型的行为是否仍然符合预设边界”。这不是军控条约，而是技术验证协议——可先由两国学术机构联合试点，不涉及主权让渡。与现有 AI 安全研究所不同，跨范式验证机制的核心不是测试模型“有没有通过安全评估”，而是测试“在无外部监控时，模型的推理路径是否仍然携带可追溯的逻辑锚点”。

核军控的历史经验表明，当双方都无法验证对方是否遵守协议时，谈判就会陷入僵局。AI 治理比核军控更难：核材料可以被物理计量，AI 的“对齐”却无法被物理计量。但历史同样表明，僵局并非终点——当双方共同承认“现有框架不够用”时，新的验证机制才有可能被创造。

路径二：从认知基元层面重构 AI，共同摆脱范式依附

这是“表意 AI”的核心方案：以“形根”替代无根 Token。形根理论的形式化定义、推理系统与工程实现，已分别在 ChinaXiv T202608.00490 和 T202608.00048 中给出。

这不是“中国方案”或“美国方案”，而是双方共同承认“当前范式不可持续”之后的技术转向。谁先完成这一转向，谁先获得真正的认知主权。这不是零和竞赛，而是一场**谁能先从悬崖边撤退**的考验。

对美国而言，承认范式缺陷意味着放弃“宗主国”地位带来的短期优势；对中国而言，摆脱代工厂地位意味着从地基开始重建。双方的利益在此并不冲突——冲突的是旧范式下“谁先跑到悬崖边”的零和逻辑。

路径三：在文明原生智能框架下实现多元共存

“文明原生智能”（CNI）不只是中美两国的议题。**全球南方**——从阿拉伯语世界到印度次大陆，从西班牙语世界到非洲语言文明——这些拥有数亿乃至十数亿人口的文明，同样在无根 Token 范式下被迫使用“英文世界观的 AI 封装”。与其让所有文明在同一个无根 Token 范式下竞争“谁的封装更好”，不如共同推进“文明原生智能”——让每一种母语文明构建其原生的 AI 体系，并通过“AI 丝绸之路”实现语义互操作与价值协同。

中国的战略角色在此清晰浮现：不是作为“另一个宗主国”参与范式竞赛，而是作为全球南方的技术旗手，率先推动认知主权的范式替代。这不仅是中国摆脱“代工厂”地位的路径，也是全球南方摆脱“数据殖民地”地位的路径。

这不是“文化多样性”的温情叙事，而是**认知主权的技术实现路径**。没有 CNI，中文世界只能生产“英文世界观的 AI 封装”，阿拉伯语世界只能生产“英文世界观的阿拉伯语皮肤”，非洲语言文明甚至连“皮肤”都难以获得。中美两国在其中的角色，不是“划分势力范围”，而是**率先承认范式缺陷，并为多元文明的原生 AI 体系提供可互操作的底层协议**。

起点可以是双语形根库的联合构建——先做语言层面的语义对齐，再做价值层面的协同。而这一起点，天然应向所有愿意参与的文明开放。

六、大国智慧与时代担当：给两国领导人的三句话

三句话——

第一句：不只谈“限速”，更谈“换地基”——限速是在一辆没有可靠刹车的车上调速度。换地基，才是对人类负责；

第二句：不把“国产替代”当成“范式革命”——算力自主不等于认知自主；

第三句：范式重构不是零和博弈。谁先启动认知基元层面的重构，谁先拥有真正的认知主权。

七、结语：一起觉醒吧

中美 AI 治理谈判，如果只停留在“谁跑得更快”“谁管得更严”，最终只会是两辆没有刹车的车在同一道悬崖边争论速度。真正的出路，是承认地基已经坏了，然后一起换掉它。

这不是技术问题，而是关于人类是否愿意正视自身局限的诚实问题。这不是某一个国家的责任，而是时代赋予大国的共同考验。AGI 已经实现——现在的问题是：它将以什么方式取代人类？而我们，是否已经准备好面对这个问题本身？

这不仅是中美两国的出路，也是**全球南方**所有在表音 AI 范式下丧失认知主权的文明的出路。CNI 框架的最终目标，是让每一种母语文明都能在数字时代拥有自己的认知根基。中

国能否成为全球南方在 AI 治理中的战略代表，取决于是否敢于率先承认范式缺陷、率先启动范式重构，**为所有被表音 AI 范式殖民的文明打开一条可复制的新路径。**

当下最紧迫的认知——**窗口期正在关闭，但尚未关闭。**范式覆巢之下，没有完卵。**是否还来得及觉醒——**旧范式内卷没有赢家。**中美两国坐的不是谈判桌，而是一辆冲向悬崖的末路之车。真正的出路，不是抢方向盘——人类文明未来有另一条路可走。**

大国担当考验在即！

参考文献

[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] 刘深. 表意 AI vs. 表音 AI：AI 新范式与去殖民化宣言（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 官网：https://logographicai.com/