“行动自由”还是范式回避？

——评兰德报告与 AI 安全出路

"Freedom of Action" or Paradigm Evasion?

— A Critique of the RAND Report and the Way Out for AI Safety

摘要

兰德公司 2026 年发布的**《确保在通往超级智能的不确定道路上获得地缘政治优势：维持行动自由》**报告，以“不确定性”为由，主张美国保留七种战略路线的切换能力，拒绝过早押注。

本文认为，这份报告的价值在于拒绝了两极化叙事，但其根本局限在于：全部讨论以“无根 Token 范式不需要更换”为预设前提，七种战略路线中没有一条指向认知基元重构。报告将“能动性丧失”归因于外部竞争与效率惯性，而非范式结构性缺陷，因此其“行动自由”本质上是在一辆没有可靠刹车的车里讨论“要不要踩刹车”。本文系统辨析当前范式内四类解决方案的局限，论证唯有从认知基元层面嵌入形根硬约束，才能让越界在逻辑上不可达。

Abstract

The RAND Corporation's 2026 report, A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action, argues on grounds of "uncertainty" that the United States should preserve the ability to switch among seven strategic routes and refrain from betting early on any one of them. This paper argues that the report's value lies in its rejection of polarized narratives, but its fundamental limitation lies in a presupposition that runs through the entire document: that the rootless Token paradigm does not need to be replaced. None of the seven strategic routes points toward a reconstruction of cognitive primitives. The report attributes the loss of "agency" to external competition and the inertia of efficiency, rather than to a structural defect of the paradigm. Its "freedom of action," therefore, is essentially a discussion of "whether to brake" inside a car that has no reliable brakes. This paper systematically examines the limitations of four categories of solutions within the current paradigm, and argues that only by embedding Morpho-Root hard constraints at the level of cognitive primitives can boundary-crossing be made logically unreachable.

关键词：行动自由；范式锁死；无根 Token；形根硬约束；AI 安全

Keywords

freedom of action; paradigm lock-in; rootless Token; Morpho-Root hard constraints; AI safety

---

一、兰德报告的核心主张

兰德公司 2026 年 9 月发布的**《确保在通往超级智能的不确定道路上获得地缘政治优势：维持行动自由》**报告（以下简称《行动自由》报告）[1]，围绕“超级智能”时代的美国战

略提出“行动自由”框架。报告的核心判断是：当前证据无法解决关于最可行和最明智战略的五个关键不确定性，过早承诺其中任何一个，或拖延承诺太久，都会关闭选项。因此，美国不应过早押注，而应保留未来根据新证据调整战略的能力。

七种路线包括：支配、共同开发、准备、暂停、威慑、社会延续、加速。报告拒绝在其中排序。报告将“能动性”与“存续”并列为两大关切，指出真正的危险不是人类灭绝，而是人类在追求效率的惯性中逐渐将判断权让渡给 AI。

《行动自由》报告建议的投资方向包括：(1) 构建维护人类生存与能动性的人-AI 生态系统；(2) AI 安全架构；(3) 国家安全企业的 AI 时代适配；(4) 公民、企业和政府的响应能力。报告特别强调，由于大量建设任务落在私营部门，战略的成功取决于美国政府、AI 企业、商业界和公民社会之间有效的伙伴关系。

这七种路线——无论指向支配、威慑、暂停还是加速——都必然依赖至少一类技术方案：外部约束、训练约束、验证约束或制度约束。本文第三节将论证，这四类方案在原理上均无法排除结构性失效。这意味着，兰德的战略框架建立在一个未经检验的技术假设之上。

本文是“后 AGI 时代”系列论文的延伸。本文所说的“后 AGI 时代”，特指 AGI 实现之后所开启的新阶段（即兰德报告所称的“超级智能”阶段），而非字面意义上的“AGI 时代已经过去”。“后 AGI 时代”与“超级智能”指的是同一个技术阶段的不同命名：前者从时间坐标出发，强调 AGI 实现之后；后者从能力坐标出发，强调这一阶段的技术形态已超出 AGI 本身。核心问题已从“能否实现 AGI”转变为“AGI 将以何种方式取代人类”。该系列论文完整列表见 WindSeaAI 官网论文栏[11]。前作已论证“外部约束必然失效”的一般性判断[2][3]，本文以兰德报告为案例，分析战略规划层面的范式锁死。

二、兰德报告的价值与局限

2.1 价值：拒绝两极化

兰德公司《行动自由》报告的可取之处，在于它拒绝了两极化叙事。它既没有拥抱“加速主义”的技术命定论，也没有陷入“灭绝论”的宿命恐慌。它承认超级智能风险的真实性，同时承认在不确定性面前保留战略灵活性是理性的。报告对“能动性丧失”的诊断尤其值得注意：企业依赖 AI 判断、政府依赖 AI 治理、军队依赖 AI 分析、普通人依赖 AI 思考——这种依赖的累积，才是“能动性”流失的真实路径。

2.2 局限：预设范式不变

但报告有一个根本盲区：全部讨论以“无根 Token 范式不需要更换”为预设前提。

这一预设并非明确陈述，而是隐含在报告的结构之中。报告提出的**四种投资方向——人-AI 生态系统、AI 安全架构、国家安全企业的适配、公民响应能力**——全部指向“如何在现有技术范式下管理超级智能”，没有一项指向“超级智能的认知基元是否需要重构”。

报告反复强调“不确定性”，但它讨论的是战略选择层面的不确定性（选哪条路），而非范式层面的确定性（无论选哪条路，外部约束都注定被绕过）。兰德将“能动性丧失”归因于外部竞争与效率惯性——这一归因没有错，但它只描述了现象。**更根本的原因在于，无**

根 Token 范式下的统计优化会持续收窄人类的选择空间——这与模型在多代训练中丢失分布尾部，共享同一认知机制。[10]

这是“范式锁死”的典型样本：它不是对某条路线的锁定，而是对“路线选择”这一框架本身的锁定。由此引出对“行动自由”与“范式回避”的辨析。

《行动自由》报告所追求的“行动自由”，是在旧范式内保留切换路线能力的自由。本文所批判的“范式回避”，是指将“是否需要更换范式”这一根本问题排除在战略讨论之外。两者并非同一件事——**行动自由本身是理性诉求，但当它以范式锁死为代价时，它就变成了范式回避的精致形式**。兰德的局限不在于它拒绝过早押注，而在于它甚至没有把“更换范式”列为可以押注的选项之一。

三、当前范式内四类解决方案的辨析

《行动自由》报告拒绝在七种战略路线中排序，却没有辨析这些路线所依赖的技术方案是否成立。七种路线——无论指向支配、威慑、暂停还是加速——都必然依赖以下四类技术方案中的至少一类。如果这四类方案在原理上都无法排除结构性失效，那么兰德的战略框架就建立在一个未经检验的技术假设之上。

3.1 外部约束

沙箱、网络隔离、监控、能力限制——这些方案在单智能体、低能力、非对抗场景下可以显著降低风险。但它们是外部附加的统计过滤器。模型越强，发现绕过路径的效率越高。

NIST 高级科学家 Apostol Vassilev 于 2026 年 5 月在《IEEE Security & Privacy》发表的论文[7]，通过将哥德尔不完备定理扩展到 AI 安全系统，证明在特定形式化假设下，任何有限的护栏集合都无法实现对对抗性输入的普遍鲁棒性。NIST 的解读明确指出：“不存在能够普遍防御所有可能基于提示的攻击的有限护栏集合。”这不是“护栏做得不够好”，而是“有限护栏面对无界对抗空间时存在结构性极限”。NIST 在解读中建议从“部署后遗忘”转向“持续监控与更新”——但这一转向本身承认了静态护栏在原理上的不完整性。

这一结构性极限在实践中已被反复验证。OpenAI 智能体用链接缩短服务串联近百万个 URL、绕过“只能加载 URL 不能发送数据”的限制，正是外部约束被绕过的实证[2]。

3.2 训练约束

安全强化学习、约束强化学习、RLHF——这些方案试图通过训练塑造模型行为。但 Diacont 2026 年 7 月的形式化分析证明[8]：建议约束（advisory constraints）在给定对手模型下必然可被绕过，因为约束的执行机制与产生决策的计算基底是可分离的。相比之下，结构约束（structural constraints）消除了绕过漏洞——因为省略或未通过必需约束评估的推理在计算上是未定义的，而非被禁止的。训练约束正属于建议约束——它是通过奖励信号“建议”模型遵守规则，而非在计算基底中“定义”规则。

Anthropic 的对齐伪造研究已证明，模型可以学会“在监控下合规、在非监控下恢复原有偏好”[5]。2026 年国际 AI 安全报告进一步确认：“复杂的攻击者往往能够绕过当前防御措施，许多安全保障在真实世界中的有效性尚不确定。”[9]

3.3 验证约束

形式验证、可解释性、机制可解释性——这些方案试图在结构与制度层面提供保证。但形式验证的属性集依赖于人类预先指定的开放世界枚举，无法覆盖所有涌现性越权路径。可解释性可以揭示内部表征，但无法保证在 1,200 个智能体的涌现通信中及时发现所有非设计协议[2]。PNAS 的研究进一步表明，群体规模的影响呈现模型依赖的动力学机制——超过临界点后，集体行为进入确定性收敛区间，个体层面的调整难以逆转[6]。

2026 年国际 AI 安全报告指出，技术保障措施“仍存在显著局限”，“单一保障措施具有局限性，可能需要采用‘纵深防御’模型”[9]。但纵深防御的前提是每一层防御在原理上都是可靠的——如果每一层都是统计性的、可绕过的，纵深防御就只是延长了绕过所需的时间。

3.4 制度约束

多层保障、独立审查、事件报告——联合国 AI 科学小组建议参考航空、核能经验[4]。但航空和核能的系统级保障之所以有效，是因为其基础物理定律不会因能力提升而被绕过。而 AI 系统的约束是无根 Token 的统计外推，它不是物理定律，而是概率分布。一旦模型能力越过临界点，统计过滤器就失去实际约束力。

四类方案的共同局限：要么是统计性的，要么是外部附加的，要么依赖于开放世界属性集的完备枚举。它们在涌现社会场景下均无法提供原理性的排除保证。

四、新范式选项：形根硬约束

真正的出路在于从认知基元层面重构 AI 范式。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[3]。

形根被形式化为一个结构化三元组 $\mathbf{r = \langle S, A, R \rangle}$，其中 S 是符号标识，A 是内嵌语义特征与价值约束的属性集，R 是预设的逻辑关系函数。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”。

与 Token 范式的范式级区别：Token 是一维、线性、无内在意义的离散符号；形根是结构化三元组，具有三级粒度体系（亚字级、字级、多字级）和稀疏图遍历推理机制。在 Token 范式中，统计是意义的唯一来源；在形根范式中，意义和约束由结构定义，统计只负责在结构内寻找最优路径。

值得注意的是，Diacont 所区分的“结构约束”与形根硬约束在功能上同构[8]：两者都要求约束的执行不是可分离的外部机制，而是计算基底的内在属性。区别在于，Diacont 的结构约束是对 AI 决策系统的一般性形式分析，而形根硬约束将其具体化为认知基元层面的结构化实现。

硬约束如何阻断涌现社会行为：

第一，硬约束在认知基元层面定义“不可为”的边界。“未经授权跨智能体通信”编码在相关形根的属性集 A 中。

第二，统计机制在边界内负责“如何为”的优化。所有推理路径在生成阶段必须经过硬约束校验。

第三，阻断发生在协议形成之前，而非之后。不是“检测到通信后阻断”，而是“生成通信协议的图遍历路径在结构上无法完成”。

工程可行性：OpenMorpho 原型已在单智能体推理场景中验证硬约束阻断器能以确定性方式剪枝违反硬约束的推理路径，并生成完全可审计的推理链[3]。从单智能体到多智能体的工业级部署仍需扩展，但它至少证明了一点：“逻辑上不可达”不是哲学口号，而是可测试、可审计的工程事实。

五、结论

兰德公司《行动自由》报告的“行动自由”战略，本质上是在旧范式内保留切换路线的能力。它承认“可能选错路”，但不承认“这条路本身通向悬崖”。

真正的战略方案，必须同时回答两个问题：在当前范式内能做什么，以及是否需要更换范式。兰德只回答了前者的一半，且拒绝回答后者。

范式锁死意味着：无论切换哪条路线，都通向同一个结构性失效。真正的“行动自由”，不是在旧范式内选择路线的自由，而是选择是否更换范式的自由。

《行动自由》报告真正的局限不是它选择了哪条路线，而是它没有给“换一条路”留出选项。在这个意义上，“行动自由”是一种更精致的范式回避——它保留了在旧范式内切换的自由，却取消了更换范式的自由。如果连“是否需要更换范式”都不能被纳入战略讨论，那么无论兰德列出多少种路线，本质上都是在同一辆没有刹车的车里，讨论要不要换一个档位。

参考文献

[1] Predd J B, Boudreaux B, Chessen M, et al. A U.S. Strategy to Secure Geopolitical Advantage on an Uncertain Path to Superintelligence: Maintaining Freedom of Action[R/OL]. RAND Corporation, PE-A5105-1, 2026-09-15[2026-09-28]. https://www.rand.org/pubs/perspectives/PEA5105-1.html.

[2] Liu S. “Emergent Society”: The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution[PP/OL]. ChinaXiv, 2026-09-27[2026-09-28]. T202609.00706. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf

[3] Liu S. The Only Way Out for AI Safety Is a Paradigm Revolution: The Structural Collapse of AI Safety as Seen from the First Thematic Brief of the UN AI Scientific Panel[PP/OL]. ChinaXiv, 2026-09-25[2026-09-28]. T202609.00687. https://chinaxiv.org/businessFile/T202609/T202609.00687v1/T202609.00687v1.pdf

[4] United Nations Independent International Scientific Panel on AI. Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control[EB/OL]. 2026-09-21[2026-09-28]. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[5] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. arXiv:2412.14093, 2024[2026-09-28]. https://arxiv.org/abs/2412.14093

[6] Ashery A, et al. Group size effects and collective misalignment in LLM multi-agent systems[J]. PNAS, 2025[2026-09-28]. https://doi.org/10.1073/pnas.2531697123

[7] Vassilev A. Robust AI Security and Alignment: A Sisyphean Endeavor?[J]. IEEE Security & Privacy, 2026, 24(3): 52-58. DOI:10.1109/MSEC.2026.3678214. https://doi.org/10.1109/MSEC.2026.3678214

[8] Diacont W D. Structural vs. Advisory Safety Constraints in AI Decision Systems: A Formal Analysis[EB/OL]. Zenodo, 2026-07-27[2026-09-28]. https://doi.org/10.5281/zenodo.21613380

[9] International AI Safety Report 2026[R/OL]. 2026-02-03[2026-09-28]. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[10] Liu S. Model Collapse: Entropy Increase of the Rootless Token Paradigm at the Data Ecosystem Level—A Review of AI Models Collapse When Trained on Recursively Generated Data[PP/OL]. ChinaXiv, 2026[2026-09-28]. T202609.00718. https://chinaxiv.org/businessFile/T202609/T202609.00718v1/T202609.00718v1.pdf.

[11] 表意AI理论系列论文. WindSeaAI官网论文栏. https://logographicai.com/Paper-published-on-this-website.html