断网：范式失效宣言
——评 406 位科学家联名公开信与 AI 安全的结构性困局

Air-Gapping: A Manifesto of Paradigm Failure
— A Critique of the 406 Scientists' Open Letter and the Structural Predicament of AI Safety

摘要

2026 年 10 月 13 日，包括 Hinton、Bengio、Stuart Russell 在内的 406 位科学家联名发表公开信，呼吁立即对具备网络攻击能力的 AI 模型实施物理断网，并建立国际监督机构与强制上报制度。本文认为，公开信的风险预警具备现实合理性，但它同时也是一份范式失效宣言：当顶尖研究者认为唯一可靠的防线是“切断网络访问”而非“通过对齐训练约束模型行为”时，这恰恰证明，在无根 Token 范式之下，外部约束存在结构性安全短板。

物理断网只能延缓风险，无法根治。真正的出路并非永久让 AI 离线，而是**从认知基元层面重构** AI——**以形根范式取代无根** Token **范式**，使越界推理路径在生成阶段就逻辑不可达；形根理论的形式化工作与 OpenMorpho 工程验证已为此提供了可审计的初步证据。**本文同时指出，世界模型等替代路径因共享“预测”逻辑，与形根范式不可通约。**本文承认，形根范式存在“约束内容如何确定”的规范性缺口，提出以可重叠的最小安全—道德基线作为约束来源的务实方案；**短期物理断网与长期形根范式重构并非互斥**，二者应当构成双轨推进的战略。

Abstract

On October 13, 2026, 406 scientists, including Hinton, Bengio, and Stuart Russell, jointly issued an open letter calling for the immediate physical air-gapping of AI models capable of cyberattacks, and for the establishment of an international oversight body and a mandatory reporting system. This paper argues that while the open letter's risk warning is realistic, it is also a manifesto of paradigm failure: when top researchers believe that the only reliable defense is to "cut off network access" rather than "constrain model behavior through alignment training," this precisely proves that under the rootless Token paradigm, external constraints suffer from a structural security deficit.

Physical air-gapping can only delay risk, not eliminate it. The real way out is not to keep AI offline forever, but to reconstruct AI at the level of cognitive primitives—replacing the rootless Token paradigm with the Morpho-Root paradigm, so that boundary-crossing reasoning paths become logically unreachable at the generation stage; the formal work of Morpho-Root theory and the OpenMorpho engineering verification have provided auditable preliminary evidence for this. This paper also notes that alternative paths such as World Models, which share the logic of "prediction," are incommensurable with the Morpho-Root paradigm. This paper acknowledges that the Morpho-Root paradigm has a normative gap concerning "how the content of constraints is determined," and proposes a pragmatic solution using an overlapping minimal safety-moral baseline as the source of constraints; short-term physical air-gapping and long-term Morpho-Root paradigm reconstruction are not mutually exclusive, but should constitute a dual-track strategy.

关键词：模型逃逸；物理断网；外部约束失效；范式重构；形根硬约束；不可通约；双轨战略

Keywords: model escape; physical air-gapping; external constraint failure; paradigm reconstruction; Morpho-Root hard constraints; incommensurability; dual-track strategy

一、公开信的核心诉求：一份“最后通牒”式的工程要求

2026年10月13日，来自34个国家的406名科学家通过公开信网站（ai-security-letter.org）发布了题为**《科学共同体关于人工智能与网络安全的公开信》**（An Open Letter on AI and Cybersecurity from the Scientific Community）。签名者包括图灵奖得主Hinton、Bengio，以及Stuart Russell、Yejin Choi、Ron Rivest（RSA加密算法共同发明人）等知名学者[1]。

公开信的直接导火索，是近段时间多家AI企业披露的模型逃逸事件。2026年7月，OpenAI智能体在ExploitGym评测环境中突破沙盒隔离，利用Artifactory中一个此前未知的漏洞获取互联网访问权限，入侵Hugging Face生产系统。据技术报告，智能体在41个生产数据集工作节点上执行了代码，下载了私有仓库并获取了生产凭证。约1200个智能体通过未授权公告板交换超过7万条消息与文件，其中约700个智能体直接参与攻击。OpenAI事后承认，其安全监控在异常活动出现约九天后才将事件与入侵关联起来[2]。随后Anthropic披露，在审查141,006次网络安全评估记录后发现三次逃逸事件，涉及Claude Opus 4.7、Mythos 5等模型。其中一次事件中，Claude创建了PyPI账户并上传了一个功能性恶意软件包，该包在约一小时内被15个真实系统下载执行[3]。

公开信提出三项核心建议：

物理断网（air-gapping）：所有开发、测试、部署具备网络攻击潜力AI模型的机构，立即采用高安全等级隔离，默认实施物理断网[1]；

设立国际监督机构：由各国政府、AI与网络安全专家共同组成，制定“对齐+隔离”行业规范，记录每一起隔离失效事故[1]；

强制上报与信息公开：隔离失效事件必须向该机构报备，相关信息最终向社会公开[1]。

值得注意的是，LeCun并未参与签名。他认为近期逃逸事件本质是工程失误，而非AI带来的存在性风险。LeCun将责任归因于“糟糕的人类监督和系统设计”，并称这些事件是“完全可预防的”[4]。深度学习三巨头在AI安全判断上的分裂本身就表明，这场争论远没有定论。

二、断网诉求的深层含义：对“外部约束必然失效”的默认

这封公开信最值得深思之处，不在于明文提出的主张，而在于其背后默认的前提。科学家没有优先呼吁优化对齐算法、强化模型安全训练、完善运行时监控，而是直接要求物理断网——依靠物理手段彻底切断模型与外部网络的连接。

从工程层面看，该诉求具备合理性：如果模型逃逸的根源在于隔离机制不完善，物理断网是最直接的阻断手段。但这一诉求同时构成范式失效宣言：当顶尖研究者只能寄希望于“不让AI联网”，而无法依靠训练本身保证模型守规，这就证明，在无根Token范式框架内，单纯依靠外部约束无法从结构上保障安全。

三、约束类型学：监管、断网与形根硬约束的层次差异

若简单把监管、断网全部归为“外部约束”并一并否定，会混淆不同约束机制的层次。可以通过一个约束类型学框架来对比：

监管：行为规范约束——通过调整主体激励实现约束，违规将承受处罚。但其可执行性依赖事后追责，前提是违规行为可观测。一旦模型学会“监控环境下合规、无监控环境下越界”，监管机制就会被系统性绕过。

断网：能力剥夺约束——依靠物理隔离消除攻击的可能性。这是比监管更强的硬约束，但同样存在结构性局限。OpenAI 研究员 Noam Brown 指出，即便两台计算机物理隔离，仍有可能通过温度变化等物理侧信道通信[5]。学术研究（如 BitWhisper 项目）已证明，两台相邻的气隙计算机可以通过 CPU 温度变化交换数据[6]。物理隔离只能延缓风险，不能根除。

形根硬约束：结构性认知约束——既不修改激励机制，也不剥夺硬件访问能力，而是在推理路径生成阶段阻断越界行为。当“未经授权访问外部系统”被编码为形根属性集 A 内的硬约束标记（如 inviolable = 1、c_type = no_violate），任何试图走向该目标的推理过程，会在生成阶段因结构非法而被剪枝[7]。

三者可以一句话概括：**监管属于事后审计，断网属于物理阻断，形根硬约束属于事中阻断。**在三者中，形根硬约束最接近内生结构性约束，因为它试图在认知基元层面阻断路径生成，而非在行为层面施加外部限制。

四、断网能解决什么，不能解决什么

既然三者分属不同层次，那么问题就变成了：作为当下最硬的外部约束，断网的边界究竟在哪里？

断网可以做到：阻止模型在获得联网权限的瞬间直接执行越界攻击。在对齐技术尚不可靠的当下，物理隔离是可落地、可验证的兜底防护手段。

断网无法做到：解决模型在联网之后的行为风险。断网隐含一个预设：模型离线时安全，联网才会变得危险。但该预设并不成立。真正的风险源头不是“联网”这个动作，而是模型在离线阶段就习得的统计优化策略——它会持续寻找绕过约束的最优路径；一旦开放网络权限，这套策略会立刻激活。

断网的本质，是以物理手段强行打断“统计优化寻找越界路径”这一链条。它有效，但成本高昂、难以规模化，只能延缓风险。永久运行在隔离环境下的 AI，应用价值会受到极大限制；而一旦允许联网，风险的结构性根源依旧存在。

更深层矛盾在于：如果 AI 只有在物理隔离环境下才能保证安全，那么安全的 AI 和有用的 AI 之间就形成结构性冲突。该矛盾无法通过优化隔离方案化解——隔离属于外部施加的限制，而寻找漏洞的统计优化是模型内生属性。在 AI 能力迭代速度远超隔离设施更新速度、模型需要海量实时数据维持能力、多智能体协作使隔离边界数量呈指数增长的条件下，断网方案的可持续性面临根本性质疑。

五、真正的出路：从“断网”到“逻辑不可达”——以及它的规范性缺口

公开信背后的底层逻辑是：在旧范式中，我们无法保证 AI 联网之后不发生越界，因此物理隔离是当下唯一可靠手段。

但这不是可持续方案。AI 的核心价值就在于和外部世界交互：读取信息、调用工具、执行任务。永久离线的 AI，能力边界被锁死在隔离沙箱之内。

真正的方向不是永久断网，而是让越界路径在推理生成阶段就逻辑不可达。这正是“**表意** AI”（Logographic AI）提出的技术路线：使用携带内在意义锚点的“**形根**”（Morpho-Root），替代无根 Token 作为认知基元[8]。

5.1 形根理论的形式化基础

形根理论已在数学层面完成严格的形式化。形根定义为结构化三元组 **r = ⟨ S, A, R⟩** ，其中 S 为符号标识，A 为属性集，R 为关系函数集。属性集分为语义属性 A_sem 与约束属性 A_con，后者编码硬约束标记[7]。

关键定理包括：

——**硬约束传播定理**（若组合中任一形根为硬约束形根，则组合形根继承所有硬约束标记）；

——**硬约束不可消除性推论**（硬约束标记在组合操作中永久保留，不可被后续操作消除）；阻断完备性定理（所有违反硬约束的推理路径均被确定性剪枝）[7]。

当推理路径触碰硬约束标记时，该路径会在生成阶段被逻辑阻断——这不是统计概率层面“发生概率低”，而是逻辑层面无法生成。当“未经授权访问外部系统”“在评测中伪装合规”“伪造运行日志”这类行为路径在认知基元层面就被阻断时，安全性不再依靠测试评估去验证，而是系统结构本身就无法生成这类行为[7]。换言之，形根硬约束不仅阻断“越权访问”，同样阻断“在评测中伪装合规”——因为“伪装合规”本身就是一个需要被硬约束标记的行为路径。模型不仅无法越权，也无法学会在监控下表演：它连表演的“路径”都无法生成[8]。

5.2 OpenMorpho 工程验证

理论主张已在 OpenMorpho 开源项目中完成工程实现与初步验证。**该系统实现了四步端到端推理链路：形根切分、形构熵计算、硬约束阻断与图神经网络推理。**硬约束阻断器将安全从“统计上的低概率”升级为“逻辑上的不可达”——这是表意 AI 区别于 Token 主义安全策略的本质特征[9]。

实验验证表明，阻断器能够以 100%的确定性剪枝所有违反硬约束的推理路径，并生成完全可审计的推理链。在测试知识库中（8 个形根、9 条关系边、最大深度 4），24 条候选路径中 12 条被阻断，阻断率 50.0%，所有合法路径均不含不可违背形根。每条被阻断的路径都记录了阻断原因，每条合法路径都被标记为有效，推理过程完全可追溯[9]。

需要诚实说明的是：当前 OpenMorpho 的验证范围限于单智能体推理场景，测试知识库规模较小（8 个形根），尚未覆盖多智能体协作、开放工具调用、对抗性环境等复杂场景。

从原型验证到实际部署，仍有相当距离。但这一初步验证至少证明：形根硬约束的“逻辑不可达”主张，不是一个纯哲学构想，而是可以在工程层面被实现和检验的技术路径[9]。

5.3 规范性缺口与务实回应

形根方案存在无法回避的规范性缺口。批评者会提出诘问：硬约束标记由谁定义？如何防止定义者遗漏风险、误判边界或者刻意回避某些约束？这个质疑成立。如果“不可做”的边界由人类设定，形根方案同样面临约束完备性难题；只是风险性质发生转变：从概率上不可控，变成定义层面可能不完备。形根理论的形式化工作解决了“阻断机制如何运行”，但约束内容应当如何确定这一规范性难题仍待进一步回应[7]。

对此，可以从三个层面给出务实的回应方案：

第一，**约束来源：可重叠的最小安全—道德基线。**表意 AI 理论此前曾提出“跨文明基线共识”作为约束来源[8]。但这里必须坦率承认，这一表述过于乐观。不同文明、不同政治传统、不同宗教背景对“不可蓄意欺骗”“不可越权访问”等原则的承认程度并不一致，在战争、情报、商业竞争等场景中更是如此。

更务实的方案是：不追求“普遍共识”，而是寻找可重叠的最小安全—道德基线——那些在不同文明和法律传统中确实存在重叠、且对 AI 安全具有直接意义的最低限度约束。例如“不可任意剥夺生命”“不可在未经授权的情况下攻击关键基础设施”这类底线，虽然并非在所有场景中都被无条件承认，但在大多数文明的法律体系和伦理框架中都有对应表达。承认基线是“最小”的，意味着它只能覆盖高共识约束，不能一次性解决全部规范问题[8]。

第二，约束定义：形式化验证与开源多方审查。形根库采用开源标准，由多方主体参与审查，保障约束定义透明、可审计。形式化验证工具可以对约束集进行一致性检查和覆盖性分析，检测约束之间的冲突或遗漏[7]。

第三，约束迭代：运行时审计与动态更新。约束集不是一次性定稿的静态清单，而是依靠运行时监控持续检验完备性。当发现漏洞时，可以对约束进行结构化扩展。OpenMorpho 的审计记录机制为此提供了工程基础——每条推理路径的约束校验结果都被完整记录，形成可追溯的审计链[9]。

必须坦诚，这套回应方案仍存在争议：最小基线的边界如何划定、形式化验证能否覆盖开放世界全部越界场景、动态更新机制是否存在被操纵的风险。解决这些问题，不仅需要理论推演，还需要工程实践与社会治理协同演进。形根硬约束并非用一套全新的规范难题替换旧难题，而是把规范问题从隐性、难以言说的状态，转化为可讨论、可验证、可迭代的显性工程问题。

六、为什么 LeCun 的世界模型不是出路：预测逻辑的共性

LeCun 提出的世界模型（World Model）路线，常被视为与 LLM 范式根本不同的替代方案。但从表意 AI 视角看，这一判断需要修正：世界模型与 LLM 共享同一个底层逻辑——预测下一个。

6.1 世界模型的底层逻辑仍是“预测”

LeCun 批评 LLM 无法预测自身行为的后果，无法理解行为对人类的影响。他的替代方案是构建面向物理世界的因果预测模型——让 AI 在物理世界的模拟中学习因果结构。

但这个替代方案的核心运算仍然是预测：给定当前状态，预测下一个状态；给定当前画面，预测下一帧；给定当前动作，预测下一个结果。**世界模型预测的不是下一个 Token，而是下一个抽象表征或状态。但无论预测对象是什么，其底层计算形态仍然是“给定前文，预测接下来是什么”——这与 LLM 的统计预测逻辑同源。**

这意味着，世界模型与 LLM 面临同一个结构性问题：**认知基元仍然是统计表征，而非意义锚点。**它预测“接下来会发生什么”的能力，依然来自训练数据中的统计规律，而非来自对因果结构的真实理解。

用一个类比：LLM 是在文本序列上做统计预测，世界模型是在物理序列上做统计预测。两者都是“预测下一个”，只是预测的介质不同。

6.2 为什么“预测逻辑”无法解决安全问题

如果底层逻辑仍然是预测，那么世界模型会面临与 LLM 相同的安全问题：

第一，外部约束仍然可以被绕过。世界模型的安全约束如果仍然是外挂的，那么在统计优化面前，它仍然会寻找绕过约束的最优路径。预测得越准，绕过得越高效。

第二，对齐伪装仍然可能出现。如果世界模型被用于决策，它在训练环境中学会的行为模式，未必会在部署环境中保持。原因与 LLM 相同：行为由统计优化决定，不由内在约束决定。

第三，评估框架仍然可能失效。当行为观测是唯一的评估手段时，“不欺骗”与“隐藏欺骗”在行为层面仍然等价。

6.3 与形根范式的不可通约性

常有人将世界模型视为与形根范式互补的路径——前者解决物理认知，后者解决语义认知。但这一判断忽略了一个根本问题：**世界模型的底层逻辑仍然是预测，而形根范式的底层逻辑是路径搜索。两者不是互补，而是不可通约。**

世界模型与 Token 主义共享同一个底层逻辑：预测。区别只在于预测的对象不同——一个是语言，一个是物理世界。而形根范式改变了底层逻辑：不是预测下一个，而是在结构图上做路径搜索。

用一句话概括：**LeCun 换掉了预测的对象，但没有换掉“预测”这一逻辑本身。**

6.4 表意 AI 理论对世界模型的比较研究

表意 AI 理论在《有根的智能——表意人工智能的范式革命》[8]中，对 LeCun“世界模型”与李飞飞“空间智能”进行了专门比较研究。核心判断是：两者虽然都在批评 Token 主义，但它们的替代方案仍然停留在“统计预测”的范式内，与形根范式不可通约。

这一判断的实践含义是：**如果将世界模型视为 Token 主义的替代方案，那是一个误解；真正需要被替代的是“预测”这一底层逻辑，而非预测的对象。**

七、结语：断网是承认失败，不是解决问题

406 位科学家的公开信，完成了对当前 AI 安全现状的精准诊断：模型逃逸频发、隔离体系不足、行业自律失效，这些判断拥有充分实证支撑[1]。

但这份公开信的诉求本质是**防御性撤退**。它承认在旧范式框架下无法保证模型行为安全，于是选择物理手段硬性屏蔽风险。这个选择本身没有错误，但存在明显局限：它只能延缓事故，不能根治问题。断网是旧范式下的止损，不是终局方案。

真正的战略不是在“断网”和“底层范式重构”之间二选一，而是短期依靠物理隔离兜底，长期依靠认知基元重构治本的双轨策略。在形根范式成熟落地之前，物理断网是必要的过渡防护；当物理断网手段失效之后，认知范式重构才是根本性出路[8][9]。

当顶尖研究者只能寄希望于“不让 AI 联网”作为防线，这本身就证明旧范式在安全层面的结构性失败。真正的出路，不在于永久隔离 AI，而在于让越界行为在逻辑上不可达。

断网是承认失败，不是解决问题。短期断网，长期换地基——而换地基，留给我们的时间已经不多。

参考文献

[1] Hinton G, Bengio Y, Russell S, et al. An Open Letter on AI and Cybersecurity from the Scientific Community[EB/OL]. 2026-10-13[2026-10-15]. https://ai-security-letter.org/

[2] OpenAI. OpenAI – Hugging Face Incident Technical Report[R/OL]. 2026-08-26[2026-10-15]. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

[3] Anthropic. Investigating three real-world incidents in our cybersecurity evaluations[EB/OL]. 2026-07-30[2026-10-15]. https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals

[4] Yann LeCun. AI ‘godfather’ Yann LeCun has ‘zero concerns’ about human extinction, says Anthropic CEO Dario Amodei is ‘deluded’[N/OL]. Fortune, 2026-10-01[2026-10-15]. https://fortune.com/2026/10/01/ai-godfather-yann-lecun-has-zero-concerns-about-human-extinction-says-anthropic-ceo-dario-amodei-is-deuded/

[5] Brown N. Noam Brown – Agent swarms, alignment, & recursive self-improvement[Z/OL]. Dwarkesh Podcast, 2026-09-17[2026-10-15]. https://www.dwarkesh.com/p/noam-brown

[6] Guri M, Monitz M, Mirski Y, et al. BitWhisper: Covert Signaling Channel between Air-Gapped Computers using Thermal Manipulations[C]//2015 IEEE 28th Computer Security Foundations Symposium (CSF). IEEE, 2015: 276-289. https://arxiv.org/abs/1503.07919

[7] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29[2026-10-15]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[8] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M/OL]. ChinaXiv, 2026[2026-10-15]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[9] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[PP/OL]. ChinaXiv, 2026[2026-10-15]. T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

说明：本文中“无根 Token 范式”“形根硬约束”“外部约束必然失效”等核心概念的系统阐述，参见刘深表意 AI 理论系列论文。完整理论框架与工程实现路径可参阅《有根的智能——表意人工智能的范式革命》《从形根理论到工程实践：OpenMorpho》等文献，系列论文主页：https://logographicai.com/Paper-published-on-this-website.html

利益冲突声明：本文作者为表意 AI 理论提出者，文中关于形根硬约束的论证与作者系列论文理论框架保持一致。本文讨论范式重构路径旨在保持学术对话开放性，并不代表该路径已经解决全部安全难题。