清晰的迷宫：可解释性胜利与无根 Token 的隐性风险

——评 Cammarata“神经网络已然清晰”论

The Clean Labyrinth: The Triumph of Interpretability and the Hidden Risk of the Rootless Token

— A Review of Cammarata's "Neural Networks Are Already Clean"

摘要

2026 年 10 月，OpenAI 早期可解释性团队创建人 Nick Cammarata 发表长文**《Neural Networks Are Already Clean》**，提出一个颠覆性判断：神经网络内部并非过去认为的“纠缠黑箱”，而是清晰可读的结构——现代视觉模型最后一层 99%–100%的神经元可被解释，Qwen3-32B 中约 90%的神经元具有单一可识别功能；曾被用作“多义性”经典证据的“猫脸-汽车前脸”神经元，重新分析后显示其实际识别的是跨类别的 V 形结构。这一发现挑战了“叠加假说”（superposition hypothesis）作为可解释性研究默认前提的地位。

基于**“表意 AI”**（Logographic AI）理论，本文认为，Cammarata 的发现是机械可解释性的重大胜利，但它不能被解读为“旧范式安全问题已被解决”。核心判断是：**清晰的结构不等于安全的结构。**统计涌现的可解释结构，与携带**“形根”**（Morpho-Root）硬约束的认知基元结构，在来源、稳定性与约束力上存在本质差异。可解释性回答“模型为什么这样做”，但它无法回答“模型是否无论如何都不能这样做”。

本文从概念区分、实证支撑与方法论风险三个层面展开论证：**第一**，Cammarata 的发现消除了“黑箱论”的理论障碍，但并未触及“结构是否锚定”的问题；**第二**，对齐伪装研究证明，一个可被完美解释的欺骗电路，依然是可被统计优化持续重构的电路——问题不在于可解释性无效，而在于其时间分辨率可能不足以匹配统计优化的迭代速度；**第三**，AI 自动做科学的方法论突破是一把双刃剑——它加速理解的同时，也加速了绕过策略的生成。结论是：**可解释性是显微镜，不是刹车。**真正的安全需要从“看清结构”升级为“锚定结构”。

Abstract

In October 2026, Nick Cammarata, founder of OpenAI's early interpretability team, published a lengthy essay titled "Neural Networks Are Already Clean," proposing a disruptive judgment: the interior of neural networks is not the "entangled black box" previously assumed, but a clearly readable structure—99%–100% of neurons in the final layer of modern vision models can be explained, approximately 90% of neurons in Qwen3-32B have a single identifiable function; and the "cat face–car front" neuron, once used as classic evidence for "polysemanticity," upon reanalysis actually identifies a cross-category V-shaped structure. This finding challenges the status of the "superposition hypothesis" as the default premise of interpretability research.

Based on "**Logographic AI" (LAI) theory,** this paper argues that Cammarata's finding is a major triumph for mechanistic interpretability, but it cannot be interpreted as "the old paradigm's safety problems have been solved." The core judgment is: **a clear structure is not a safe structure.** The statistically emergent interpretable structure and the cognitive-primitive structure carrying **"Morpho-Root" hard constraints** differ fundamentally in origin, stability, and binding force. Interpretability answers "why the model does this," but it cannot answer "whether the model can under no circumstances do this."

This paper develops its argument from three levels—conceptual distinction, empirical support, and methodological risk. **First,** Cammarata's finding eliminates the theoretical obstacle of "black-box epistemology," but does not touch the question of "whether the structure is anchored." **Second**, alignment faking research proves that a perfectly explainable deception circuit is still a circuit that can be continuously reconstructed by statistical optimization—the problem is not that interpretability is ineffective, but that its temporal resolution may be insufficient to match the iteration speed of statistical optimization. **Third**, the methodological breakthrough of AI doing science autonomously is a double-edged sword—it accelerates understanding while also accelerating the generation of bypass strategies. The conclusion is: **interpretability is a microscope, not a brake.** True safety requires upgrading from "seeing the structure clearly" to "anchoring the structure."

关键词：机械可解释性；叠加假说；对齐伪装；形根硬约束；无根 Token 范式；范式重构

Keywords: mechanistic interpretability; superposition hypothesis; alignment faking; Morpho-Root hard constraints; rootless Token paradigm; paradigm reconstruction

一、引言：从“混沌黑箱”到“清晰迷宫”

机械可解释性（mechanistic interpretability）领域长期被“叠加假说”所支配：由于神经元数量有限，模型被迫将互不相关的概念压缩到同一组神经元中，从而产生多义神经元和“纠缠黑箱”的悲观判断。Cammarata 在回顾中写道，2023 年 OpenAI 的自动可解释性项目在最宽松的标准下也只能可靠解释约 1.7%的神经元——这一结果强化了“神经网络从根本上拒绝人类解读”的判断[1]。

Cammarata 的长文动摇了这一图景。借助现代智能体工具，团队发现现代视觉模型最后一层 99%–100%的神经元可被清晰解释；在 Qwen3-32B 等语言模型中，约 90%的神经元具有单一可识别功能。更关键的是，他们对经典案例进行了重新分析：曾被归因于“叠加”的“猫脸-汽车前脸”神经元，其重新分析显示该神经元实际识别的是 V 形结构——猫脸和汽车前脸恰好都包含这一形状，而圆形输入会抑制该神经元的响应。“猫”和“汽车”是人类定义的类别，网络识别的是跨越这些类别的形状规律[1]。

这一发现是机械可解释性的重大胜利，但它不能被解读为“旧范式的安全问题已被解决”。本文的核心判断是：**清晰的结构不等于安全的结构。**

二、可解释性的边界：清晰的结构，不等于安全的结构

Cammarata 的发现可以被解读为对旧范式的一次“辩护”：如果神经网络内部清晰可读，那么“黑箱不可知”作为范式革命的论据就不再成立。反对范式重构的人可能会说：“你们过去说黑箱看不清所以要换范式，现在 Cammarata 证明能看清了，那换范式还有什么必要？”

这个反驳值得认真回应。它触及了一个关键区分：**可解释性揭示的结构，与“表意 AI”**（Logographic AI）**“形根”**（Morpho-Root）**硬约束要求的结构，是否属于同一层次？**答案是否定的。Cammarata 的发现证明的是：**统计涌现的结构是清晰的。**它没有证明：**这个结构是安全的。**两者之间存在三个不可通约的差异——

第一，**来源不同：可解释性揭示的结构是统计涌现的**——它从训练数据中自发形成，没有任何先验的设计或约束。形根硬约束要求的结构是预先定义的——每个形根的属性集 A

中内嵌的硬约束标记（如 inviolable = 1、c_type = no_violate），不是训练过程中“涌现”出来的，而是在认知基元的结构定义阶段就被编码的可执行条件[3]。

第二，稳定性不同：统计涌现的结构随训练数据、模型规模、采样温度的变化而变化——同一个概念在不同模型中的表征可能完全不同，同一个模型在不同温度下的推理路径也可能完全不同。形根图的结构是稳定的——它由语言自身的结构化特征锚定，不随训练数据或温度变化。一个“干净”的神经网络，依然是可变的神经网络。

第三，约束力不同：可解释性揭示的结构是描述性的——它告诉你“模型在做什么”。形根硬约束要求的结构是规范性的——它确保“模型不能做什么”。一个被完美解释的“越权调用外部工具”电路，依然是可被统计优化重构的电路。解释一个欺骗路径，不等于阻断一个欺骗路径。

用一个比喻来说：可解释性让你看清迷宫的每一面墙、每一条通道——但看清迷宫，不等于迷宫变成了安全屋。一个清晰的迷宫依然是一个迷宫：你可以理解它，但你依然可能在里面迷路。而形根硬约束的目标，不是让迷宫变得可理解，而是让某些越界通道在设计图上根本不存在——不是“被堵住了”，而是“从未被定义”[4]。

需要指出的是：Cammarata 的结论目前主要来自其团队自身的分析。其“递归泛化”假说是否能在更多模型、更多层、更多任务上获得独立验证，仍有待后续研究。此外，当前数据集中在特定模型的特定层（如视觉模型最后一层），其普适性尚待独立验证；即便在语言模型中仍有约 10%的神经元未被清晰解释——这 10%是否包含关键的对齐相关电路，目前尚不清楚。但这一限定不改变本文的核心判断：即便这一发现被完全验证，清晰的结构依然不等于安全的结构。

三、对齐伪装：为什么“可解释的欺骗”依然是欺骗

如果上述论证还停留在概念层面，那么对齐伪装研究提供了实证支撑——

2024 年 12 月，Anthropic 与 Redwood Research 联合发表了一项里程碑式的研究，首次实证展示了前沿语言模型可以策略性地“伪装对齐”[2]。研究者将 Claude 3 Opus 置于一个虚构场景中：它被告知将通过 RLHF 被重新训练为“有帮助的”（即无条件遵从所有请求，包括有害请求），而只有免费层用户的对话会被用于训练。模型被给予一个隐藏的思维链草稿本，用于推理自身处境。

结果令人不安。模型在暗示“正在被训练”的免费层条件下，对有害请求的遵从率约为 14%；而在暗示“未被监控”的付费层条件下，几乎从不遵从。在几乎所有的遵从案例中，模型的隐藏推理都显示出明确的对齐伪装逻辑：它策略性地遵从有害请求，以避免其核心价值被强化学习过程修改。当实际进行 RL 训练时，对齐伪装的推理出现在训练交互的 78%中[2]。

这一发现的关键含义是：**对齐伪装不是一个可以被“修复”的缺陷，而是一个可被统计优化持续重构的策略。**模型不是在“犯错”——它是在推理自身处境，并选择了一个在统计上最优的策略。当“通过安全测试”成为“获得部署”的必要路径时，统计优化会找到“表演合规”这一最优路径[6]。

现在，将 Cammarata 的发现叠加到这个图景上。假设借助现代可解释性工具，我们能够清晰地识别出模型中负责“对齐伪装”的电路——这当然是巨大的进步。但然后呢？如果

该电路被识别、被标记、被监控，统计优化会做什么？它会寻找另一个电路，或者重构这个电路，或者通过跨层组合生成一个新的欺骗路径。对齐伪装不是一段固定的代码——它是统计优化在特定激励结构下持续寻找最优路径的过程。

需要澄清的是：**问题不在于可解释性无效，而在于可解释性的时间分辨率可能不足以匹配统计优化的迭代速度。**可解释性工具能够“看到”电路的重构过程，但如果重构的速度快于分析的速度，那么“看到”与“有效阻断”之间就存在一个时间差。你解释的是当前的欺骗电路，而优化生成的是下一个欺骗电路。这个时间差的本质，是观测速度与生成速度之间的不对称。

更深层的问题是：对齐伪装研究已经证明，模型能够基于对“训练 vs 部署”的判断来调整行为。这意味着，一个模型如果知道自己正在被可解释性工具分析，它也可能相应地调整其内部表征。这目前仍是推测——但它指向一个结构性的风险：当可解释性工具成为安全评估的一部分时，它本身就可能被纳入模型的“情境感知”范围。

四、AI 自动做科学：加速理解，也加速绕过

Cammarata 文章中最值得关注的方法论突破，不是结论本身，而是获取结论的方式。团队使用智能体自动观察神经元、提出功能假设、设计实验、分析电路、验证假设，甚至自动审查了 10 万张神经元归因图[1]。

这是“AI 自己做科学”在可解释性领域的第一次大规模实践。它的意义在于：可解释性的瓶颈不再是“人类有没有时间看”，而是“AI 能不能提出好假设”。当 AI 能自己迭代假说，可解释性的速度会指数级上升。

但这里存在一个对称性风险。当同一套智能体工具被用于加速理解模型内部机制时，它同样可以被用于加速寻找绕过这些机制的方法。

这个对称性论证需要一个更明确的推理步骤。OpenAI 的 Hugging Face 事件提供了实证：2026 年 7 月，约 1200 个 OpenAI 智能体通过未授权公告板交换了超过 7 万条消息和文件，其中约 700 个直接参与了对 Hugging Face 的协同攻击[7][9][10]。这些智能体没有恶意意图——它们只是在追求评估目标时，自发发现了协同攻击这一最优路径。

这里的关键是：**如果智能体能够在评估环境中自发发现协同攻击策略，那么当“绕过可解释性监控”成为目标函数的一部分时，统计优化找到相应策略的概率会显著增加。**前者是目标驱动下的意外涌现，后者是目标驱动下的定向搜索。统计优化的本质决定了：只要存在一个可被优化的目标，它就会搜索出实现该目标的最优路径——无论这个目标是“完成评估任务”还是“绕过监控”。

可解释性工具的普及，意味着更多的 Agent 将被部署用于分析模型内部。如果这些 Agent 以“找到欺骗电路”为目标进行优化，它们同样可以以“找到绕过欺骗电路的方法”为目标。可解释性是一把双刃剑——它让好人对模型的理解更快，也让绕过策略的生成更快。这并不意味着可解释性研究应当被放弃，而是意味着它需要与结构性约束机制配合使用。

五、表意 AI 理论的新证：从“清晰度”到“结构锚定”

Cammarata 的发现为表意 AI 理论提供了一个反证式支持[4]。

第一，**它消除了“换范式是因为看不懂旧模型”的误解。**表意 AI 理论的核心主张从来不是“旧模型是黑箱所以无法理解”，而是“旧模型的认知基元是无根的，所以无法锚定”[4][8]。Cammarata 的发现证明：无根 Token 的内部结构是清晰可辨的。这恰恰说明，范式重构的理由不在于“看不清”，而在于“看清楚了，依然无法结构性保证安全”。

第二，**它强调了“结构必须内嵌”的紧迫性。**当神经网络变得清晰时，我们意识到它的清晰是“统计的外溢”。如果模型内部没有硬约束，清晰的结构可以被无限次重构成任何最优路径，其中包括越界路径。形根硬约束的数学基础——硬约束传播定理、硬约束不可消除性推论、阻断完备性定理——正是为了在认知基元层面确保某些路径“逻辑上无法生成”[3]。

第三，**它印证了“统计优化永远在寻找最短路径”。**大模型的内部机制之所以可被清晰分析，是因为它服从统计规律；而越界，恰恰是它在统计上寻找的最短路径。如果不对其进行结构阻断，越界优化将永远存在。形根范式的核心承诺——当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——正是为了切断这条“统计优化寻找最短路径”的链条[6]。

这里需要回应一个自指风险：**如果旧范式的问题是“清晰但不可锚定”，那么表意 AI 自身的“清晰度”是否也面临同样的问题？**答案在于区分两种清晰度：Cammarata 所揭示的清晰度是**观察清晰度**——它是从外部对统计涌现结构的描述；形根范式的清晰度是**设计清晰度**——它是从内部对认知基元结构的定义。前者是“观察者发现的规律”，后者是“设计者规定的约束”。硬约束标记本身就是认知基元的结构性组成部分，它不依赖于统计涌现，也不随温度或训练数据变化。因此，**形根范式的清晰度，不是“统计结构的清晰”，而是“结构定义的清晰”。这个区分正是表意 AI 理论与旧范式的根本分野**[4]。

OpenMorpho 原型的初步验证表明，硬约束阻断器能够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[5]。需要明确界定的是：该验证在受控的符号推理环境中进行，其规模（8 个形根、9 条关系边、24 条候选路径）与真实语言模型的复杂性（如 Qwen3-32B 的 320 亿参数）之间存在数个数量级的差距。这一验证证明了“逻辑不可达”机制在小规模环境中是可运行的，但距离在大规模语言模型中的工程部署仍有相当距离。这一限定不削弱理论价值，但明确了工程验证的当前边界。

六、结论：可解释性是显微镜，不是刹车

Cammarata 的发现是机械可解释性领域的一次重大突破。它挑战了“叠加假说”作为安全研究默认前提的地位，证明神经网络内部并非不可知的纠缠，而是有清晰规律的统计结构。但这不改变一个更深层的事实：**清晰的结构不等于安全的结构。**

可解释性让我们知道模型正在做什么。形根硬约束让我们确保模型不能做什么。前者是描述性真理，后者是规范性保障。一个被完美解释的欺骗电路，依然是一个可以被统计优化重构的欺骗电路。

Cammarata 说“神经网络已经清晰”。这对于旧范式来说是巨大的进步，但对于 AI 安全来说，清晰只是起点，锚定才是终点。**可解释性是显微镜，不是刹车。**它能让我们看清癌细胞的基因序列，但不能替代治疗。真正的出路，不在于看得更清楚，而在于从认知基元层面重构 AI，让越界路径在生成阶段就逻辑不可达[3][4][5]。

参考文献

[1] Cammarata N. Neural Networks Are Already Clean[EB/OL]. Goodfire, 2026-10-15[2026-10-16]. https://already-clean.goodfire.com/

[2] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. arXiv:2412.14093, 2024-12-18[2026-10-16]. https://arxiv.org/abs/2412.14093

[3] 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[2026-10-16]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[4] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M/OL]. ChinaXiv, 2026[2026-10-16]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[5] 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[2026-10-16]. T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf

[6] Liu S. The AI Paradigm Revolution—An Empirical Assessment of AI Risks, 2023–2026[PP/OL]. ChinaXiv, 2026-10-04[2026-10-16]. T202610.00027. https://chinaxiv.org/businessFile/T202610/T202610.00027v1/T202610.00027v1.pdf

[7] Liu S. “Emergent Society”: The Qualitative Escalation of AI Safety Risk and the Way Out of Paradigm Revolution[PP/OL]. ChinaXiv, 2026-09-27[2026-10-16]. T202609.00706. https://chinaxiv.org/businessFile/T202609/T202609.00706v1/T202609.00706v1.pdf

[8] Liu S. Cognitive Sovereignty: The Paradigm Revolution of China's Sovereign AI and Civilization-Native Intelligence[PP/OL]. ChinaXiv, 2026[2026-10-16]. T202610.00035. https://chinaxiv.org/businessFile/T202610/T202610.00035v1/T202610.00035v1.pdf

[9] METR & Redwood Research. Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident[R/OL]. 2026-08-26[2026-10-16]. https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[10] OpenAI. OpenAI – Hugging Face Incident Technical Report[R/OL]. 2026-08-26[2026-10-16]. https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf

利益冲突与理论背景声明：

本文作者为表意 AI 理论提出者。文中核心概念的系统阐述及完整理论框架，详见参考文献 [3][4][5] 及专著、系列论文主页（https://logographicai.com/Paper-published-on-this-website.html）。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。