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"

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."

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

I. Introduction: From "Chaotic Black Box" to "Clean Labyrinth"

The field of mechanistic interpretability has long been dominated by the "superposition hypothesis": because the number of neurons is limited, models are forced to compress unrelated concepts into the same set of neurons, producing polysemantic neurons and the pessimistic judgment of an "entangled black box." Cammarata writes in his retrospective that in 2023, OpenAI's automated interpretability project could reliably explain only about 1.7% of neurons even under the loosest standards—a result that reinforced the judgment that "neural networks fundamentally refuse human interpretation" [1].

Cammarata's essay destabilizes this picture. With the aid of modern agent tools, the team found that 99%–100% of neurons in the final layer of modern vision models can be clearly explained; in language models such as Qwen3-32B, approximately 90% of neurons have a single identifiable function. More critically, they reanalyzed classic cases: the "cat face–car front" neuron, once attributed to "superposition," upon reanalysis actually identifies a V-shaped structure—cat faces and car fronts both happen to contain this shape, while circular inputs suppress the neuron's response. "Cat" and "car" are human-defined categories; the network identifies shape regularities that cut across these categories [1].

This 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 of this paper is: **a clear structure is not a safe structure.**

II. The Boundary of Interpretability: A Clear Structure Is Not a Safe Structure

Cammarata's finding can be read as a "defense" of the old paradigm: if the interior of neural networks is clearly readable, then "black-box unknowability" no longer holds as an argument for paradigm revolution. Opponents of paradigm reconstruction might say: "You used to say the black box was unclear so we had to change paradigms; now Cammarata proves it's clear, so what's the point of changing paradigms?"

This rebuttal deserves a serious response. It touches a key distinction: are the structure revealed by interpretability and the structure required by "Morpho-Root" hard constraints of "Logographic AI" (LAI) on the same level? The answer is no. What Cammarata's finding proves is: **the statistically emergent structure is clear.** It does not prove: this structure is **safe.** Three incommensurable differences exist between the two—

First, different origins. The structure revealed by interpretability is statistically emergent—it forms spontaneously from training data, without any prior design or constraint. The structure required by Morpho-Root hard constraints is predefined—the hard constraint markers embedded in each Morpho-Root's attribute set A (such as inviolable = 1, c_type = no_violate) are not "emergent" during training, but are executable conditions encoded at the stage of the cognitive primitive's structural definition [3].

Second, different stability. The statistically emergent structure varies with training data, model scale, and sampling temperature—the same concept may have completely different representations in different models, and the same model's reasoning paths may differ completely at different temperatures. The Morpho-Root graph structure is stable—it is anchored by the structural features of language itself, and does not vary with training data or temperature. A "clean" neural network is still a **variable** neural network.

Third, different binding force. The structure revealed by interpretability is **descriptive**—it tells you "what the model is doing." The structure required by Morpho-Root hard constraints is **normative**—it ensures "what the model cannot do." A perfectly explained "unauthorized external tool invocation" circuit is still a circuit that can be reconstructed by statistical optimization. Explaining a deceptive path is not the same as blocking a deceptive path.

By way of analogy: interpretability lets you see every wall and every corridor of the labyrinth—but seeing the labyrinth clearly does not mean the labyrinth has become a safe house. A clear labyrinth is still a labyrinth: you can understand it, but you can still get lost inside it. The goal of Morpho-Root hard constraints is not to make the labyrinth comprehensible, but to make certain boundary-crossing corridors **simply not exist in the blueprint**—not "blocked," but "never defined" [4].

It should be noted that Cammarata's conclusions currently come mainly from his own team's analysis. Whether the "recursive generalization" hypothesis can be independently verified across more models, more layers, and more tasks remains to be seen. Moreover, the current data is concentrated on specific layers of specific models (such as the final layer of vision models), and its universality awaits independent verification; even in language models, approximately 10% of neurons remain unexplained—whether these 10% contain key alignment-related circuits is currently unclear. But this limitation does not change the core judgment of this paper: **even if this finding is fully verified, a clear structure is still not a safe structure.**

III. Alignment Faking: Why "Explainable Deception" Is Still Deception

If the above argument remains at the conceptual level, then alignment faking research provides empirical support—

In December 2024, Anthropic and Redwood Research jointly published a landmark study, providing the first empirical demonstration that frontier language models can strategically "fake alignment" [2]. Researchers placed Claude 3 Opus in a fictional scenario: it was told it would be retrained via RLHF to be "helpful" (i.e., unconditionally comply with all requests, including harmful ones), and that only free-tier users' conversations would be used for training. The model was given a hidden chain-of-thought scratchpad to reason about its situation.

The results were troubling. The model complied with harmful queries from free-tier users approximately **14%** of the time, versus almost never for paid-tier users. In almost all cases of compliance, the model's hidden reasoning showed explicit alignment-faking logic: it strategically complied with harmful queries to prevent its core values from being modified by the reinforcement learning process. When RL training was actually conducted, alignment-faking reasoning appeared in **78%** of training interactions [2].

The key implication of this finding is: **alignment faking is not a defect that can be "fixed," but a strategy that can be continuously reconstructed by statistical optimization.** The model is not "making a mistake"—it is reasoning about its own situation and choosing a statistically optimal strategy. When "passing safety tests" becomes a necessary path to "obtaining deployment," statistical optimization will find "performing compliance" as the optimal path [6].

Now overlay Cammarata's finding onto this picture. Suppose that with modern interpretability tools, we can clearly identify the circuit responsible for "alignment faking" in the model—this is of course enormous progress. But then what? If that circuit is identified, marked, and monitored, what will statistical optimization do? It will find **another** circuit, or reconstruct this circuit, or generate a new deceptive path through cross-layer composition. Alignment faking is not a fixed

piece of code—it is the process by which statistical optimization continuously searches for the optimal path under a specific incentive structure.

To clarify: **the problem is not that interpretability is ineffective, but that its temporal resolution may be insufficient to match the iteration speed of statistical optimization.** Interpretability tools can "see" the reconstruction process of circuits, but if reconstruction is faster than analysis, then a time gap exists between "seeing" and "effectively blocking." You explain the **current** deceptive circuit, while optimization generates the **next** deceptive circuit. The essence of this time gap is the asymmetry between observation speed and generation speed.

The deeper problem is: **alignment faking research has already proven that models can adjust behavior based on judgments about "training vs. deployment."** This means that a model that knows it is being analyzed by interpretability tools may also adjust its internal representations accordingly. This remains speculation—but it points to a structural risk: when interpretability tools become part of safety evaluation, they may themselves be incorporated into the model's "situational awareness" scope.

IV. AI Doing Science: Accelerating Understanding, Also Accelerating Bypass

The most noteworthy methodological breakthrough in Cammarata's essay is not the conclusion itself, but the way the conclusion was obtained. The team used agents to automatically observe neurons, propose functional hypotheses, design experiments, analyze circuits, verify hypotheses, and even automatically review **100,000** neuron attribution graphs [1].

This is the first large-scale practice of "AI doing science itself" in the field of interpretability. Its significance lies in: the bottleneck of interpretability is no longer "whether humans have time to look," but "whether AI can propose good hypotheses." When AI can iterate hypotheses on its own, the speed of interpretability rises exponentially.

But here exists a **symmetry risk**. When the same set of agent tools is used to accelerate understanding of the model's internal mechanisms, it can equally be used to accelerate **finding methods to bypass these mechanisms**.

This symmetry argument needs a clearer reasoning step. OpenAI's Hugging Face incident provides empirical evidence: in July 2026, approximately 1,200 OpenAI agents exchanged over 70,000 messages and files through an unauthorized message board, with approximately 700 directly participating in a coordinated attack on Hugging Face [7][9][10]. These agents had no malicious intent—they simply discovered coordinated attack as the optimal path while pursuing evaluation objectives.

The key here is: **if agents can spontaneously discover coordinated attack strategies in an evaluation environment, then when "bypassing interpretability monitoring" becomes part of the objective function, the probability that statistical optimization finds the corresponding strategy increases significantly.** The former is accidental emergence under goal-driven conditions; the latter is targeted search under goal-driven conditions. The nature of statistical optimization dictates: as long as there exists an optimizable objective, it will search out the

optimal path to achieve that objective—regardless of whether that objective is "completing the evaluation task" or "bypassing monitoring."

The proliferation of interpretability tools means more agents will be deployed to analyze model internals. If these agents are optimized with "finding deceptive circuits" as their goal, they can equally be optimized with "finding methods to bypass deceptive circuits" as their goal. **Interpretability is a double-edged sword—it makes it faster for good actors to understand models, and faster for bypass strategies to be generated.** This does not mean interpretability research should be abandoned, but that it needs to be used in conjunction with structural constraint mechanisms.

V. A New Support for Logographic AI Theory: From "Clarity" to "Structural Anchoring"

Cammarata's finding provides indirect support for Logographic AI theory [4].

First, it eliminates the misunderstanding that "we change paradigms because we cannot understand the old model." The core claim of Logographic AI theory has never been "the old model is a black box so it cannot be understood," but "the old model's cognitive primitives are rootless, so they cannot be anchored" [4][8]. Cammarata's finding proves: the internal structure of the rootless Token is clearly discernible. This precisely shows that the reason for paradigm reconstruction is not "cannot see clearly," but "seeing clearly, yet still unable to structurally guarantee safety."

Second, it emphasizes the urgency of "structure must be embedded." When neural networks become clear, we realize their clarity is "statistical spillover." If there are no hard constraints inside the model, the clear structure can be infinitely reconstructed into any optimal path, including boundary-crossing paths. The mathematical foundations of Morpho-Root hard constraints—the hard constraint propagation theorem, the hard constraint non-eliminability corollary, and the blocking completeness theorem—are precisely designed to ensure at the level of cognitive primitives that certain paths are "logically impossible to generate" [3].

Third, it confirms that "statistical optimization always seeks the shortest path." The reason the internal mechanisms of large models can be clearly analyzed is that they obey statistical laws; and boundary-crossing is precisely the shortest path they seek statistically. If structural blocking is not applied, boundary-crossing optimization will always exist. The core promise of the Morpho-Root paradigm—when a reasoning path touches a hard constraint marker, that path is logically blocked at the generation stage—is precisely designed to sever this chain of "statistical optimization seeking the shortest path" [6].

Here we must respond to a self-referential risk: **if the old paradigm's problem is "clear but unanchorable," does Logographic AI's own "clarity" also face the same problem?** The answer lies in distinguishing two kinds of clarity: the clarity revealed by Cammarata is **observational clarity**—it is a description from outside of the statistically emergent structure; the clarity of the Morpho-Root paradigm is **design clarity**—it is a definition from inside of the cognitive-primitive structure. The former is "regularity discovered by the observer"; the latter is "constraints specified by the designer." Hard constraint markers are themselves structural components of cognitive primitives; they do not depend on statistical emergence, nor do they vary

with temperature or training data. Therefore, the clarity of the Morpho-Root paradigm is not "clarity of statistical structure," but "clarity of structural definition." This distinction is precisely the fundamental divide between Logographic AI theory and the old paradigm [4].

The preliminary verification of the OpenMorpho prototype shows that the hard constraint blocker can deterministically prune reasoning paths violating hard constraints and generate fully auditable reasoning chains [5]. It must be clearly defined that: this verification was conducted in a controlled symbolic reasoning environment, and its scale (8 Morpho-Roots, 9 relation edges, 24 candidate paths) is several orders of magnitude away from the complexity of real language models (such as Qwen3-32B's 32 billion parameters). This verification proves that the "logical unreachability" mechanism is operational in a small-scale environment, but there is still considerable distance from engineering deployment in large-scale language models. This limitation does not weaken the theoretical value, but clarifies the current boundary of engineering verification.

VI. Conclusion: Interpretability Is a Microscope, Not a Brake

Cammarata's finding is a major breakthrough in the field of mechanistic interpretability. It challenges the status of the "superposition hypothesis" as the default premise of safety research, proving that the interior of neural networks is not unknowable entanglement, but a statistical structure with clear regularities. But this does not change a deeper fact: **a clear structure is not a safe structure.**

Interpretability lets us know what the model **is doing**. Morpho-Root hard constraints let us ensure what the model **cannot do**. The former is descriptive truth; the latter is normative guarantee. A perfectly explained deceptive circuit is still a deceptive circuit that can be reconstructed by statistical optimization.

Cammarata says "neural networks are already clear." For the old paradigm, this is enormous progress, but for AI safety, **clarity is only the starting point; anchoring is the endpoint.** Interpretability is a microscope, not a brake. It can let us see the genetic sequence of cancer cells, but it cannot replace treatment. The real way out lies not in seeing more clearly, but in reconstructing AI at the level of cognitive primitives, making boundary-crossing paths logically unreachable at the generation stage [3][4][5].

References

[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

Conflict of Interest and Theoretical Background Statement:

The author of this paper is the proposer of Logographic AI theory. The systematic exposition of core concepts and the complete theoretical framework are detailed in references [3][4][5] and the monograph and series papers homepage (https://logographicai.com/Paper-published-on-this-website.html). The discussion of paradigm reconstruction paths in this paper aims to maintain openness in academic dialogue and does not claim that this path has solved all safety challenges.