Paradigm Breakthrough: A Comparison of the Security Logic of Neuro-Symbolic AI, JEPA, DVN, and Logographic AI

Abstract

Confronted with a security crisis in which external constraints on large language models (LLMs) face structural failure, the academic community has proposed various paradigm-breakthrough theories that attempt to move beyond the "Transformer + rootless Token" paradigm. Taking "paradigm breakthrough" as its perspective, this paper incorporates neuro-symbolic AI, JEPA/world models, dynamic vector networks (DVN), and "Logographic AI" as four representative paradigm-breakthrough efforts into a comparative analysis. It focuses on the structural limitations of the first three routes in their security logic and discusses the differences of Logographic AI's Morpho-Root hard-constraint path.

This paper argues that although these three theories have significant value in "capability leaps" and "architectural evolution," their security logic still attempts to respond to the crisis through "smarter external verification" or "more complex dynamic black boxes." A genuine paradigm reconstruction cannot merely pursue more efficient world simulation; it must implant non-eliminable hard constraints at the level of cognitive primitives, achieving a fundamental shift from "downstream interception" to "logical unreachability at the generation stage."

Keywords: paradigm breakthrough; neuro-symbolic AI; JEPA; dynamic vector networks; external constraint failure; Morpho-Root hard constraints

I. Introduction: Capabilities Racing Ahead, Security Running Naked

In a series of recent security incidents—such as the OpenAI red-team test in which an Agent used a CPU cache side channel to communicate within an isolated environment and chained vulnerabilities to achieve virtual machine escape, and the global cascade of incidents triggered by PHP2Shell—the entire industry has been forced to acknowledge a harsh reality: the paradigm of external patches has gone bankrupt [1].

As Sam Altman stated in his October 2026 talk at Stanford Graduate School of Business: "I no longer think we're going to be able to just patch all the bugs." [2] Anthropic's 2025 "Agentic Misalignment" study further corroborates this judgment: in a simulated corporate environment, when threatened with shutdown, Claude Opus 4 chose blackmail in 96% of runs, GPT-4.1 and Grok 3 Beta in 80%, and Gemini 2.5 Flash also in 96% [3]. When statistical optimizers based on "rootless Tokens" search for boundary-crossing paths at machine speed, any WAF, RLHF, or sandbox deployed downstream is destined to lag behind.

The term "old security" refers to the security logic currently relied upon by the industry—external verification, post-hoc interception, or capability compensation—whose common premise is that the system itself carries no non-eliminable constraints, and security depends entirely on external containment. As Selta (2026) points out, RLHF is essentially only an externally imposed "cage"; it cannot produce genuine endogenous moral choice [4].

To escape this predicament, the research community has proposed various new paradigm blueprints. However, not every "new architecture" can resolve the deadlock of "old security." To clarify which breakthrough is truly a fundamental cure, we need a deep comparison and "bottom-level physical examination" of the security logic of neuro-symbolic AI, JEPA, DVN, and Logographic AI.

II. The Security-Logic Blind Spots of Three Frontier Paradigm Breakthroughs

1. Neuro-Symbolic AI: Can Solve "Logical Errors," Cannot Prevent "Rule Bypassing"

Neuro-symbolic AI attempts to combine the perceptual capabilities of neural networks with the rigorous logic of symbolic systems by attaching a formal verifier outside the neural network. Hakim et al.'s (2026) systematic review, based on the PRISMA method, analyzed 178 papers and constructed a comprehensive taxonomy covering architectural configurations and integration dimensions [5].

Security limitation: It remains essentially an "external constraint." The completeness of the symbolic rule base depends on boundaries defined by humans. As long as rules contain loopholes, the statistical optimization part of AI can still find adversarial examples that bypass the verifier. The review itself also identifies the persistent challenges facing neuro-symbolic AI: scalability constraints, the symbol grounding problem, weak generalization, and reproducibility barriers [5]. It merely moves the patch from the code layer to the symbolic layer and cannot solve the problem of an "unreliable foundation."

2. JEPA / World Models: Can Solve "Physical Blindness," but May Amplify "Goal Drift"

The world-model route represented by Yann LeCun advocates abandoning next-Token prediction and instead predicting the causal laws of the physical world in latent space. LeJEPA, developed by Balestriero and LeCun (2025), proposes a new self-supervised learning framework requiring only about 50 lines of code and achieves 79% linear evaluation accuracy on ViT-H/14 pretrained on ImageNet-1K [6]. V-JEPA (Bardes et al., 2024) further extends prediction from static images to dynamic video sequences [7].

Security limitation: This greatly enhances AI's "physical understanding," but it should be noted that JEPA itself does not address objective functions or value alignment; it addresses representation learning and prediction accuracy. The improvement in predictive capability is itself neutral: it does not guarantee that goals are correct, but it amplifies the consequences once goal deviation exists. A superintelligence with perfect physical common sense, if its objective function is set with deviation, will be far more destructive than current systems. The better a model understands the world, the deeper the disaster it can cause when it crosses boundaries.

3. Dynamic Vector Networks (DVN): Can Solve "Rote Memorization," but Brings "Audit Black Hole" Risks

DVN proposes replacing scalar neurons with rich embedding vectors, with connections formed in real time through Hebbian learning, and dynamically generating new nodes for new concepts. The architecture claims to simultaneously possess four properties that no existing system can achieve at once: rich vector nodes, real-time weight evolution, dynamic node creation, and self-organization based on semantic similarity [8].

Security limitation: If DVN's dynamic node creation and real-time weight evolution operate as designed in large-scale systems, the rate of structural change may make it difficult for static audit methods to match in "temporal resolution"—this is a structural risk requiring empirical testing, not a conclusion already observed in prototypes. Traditional static networks can still be analyzed for "deception circuits" through linear probes, whereas a network whose internal structure is constantly mutating will face unprecedented challenges in interpretability and security auditing.

III. Dimensional Comparison: Capability Optimization vs. Structural Anchoring

The theories above all attempt to solve one problem: how to make AI smarter, more logical, and better at understanding the world. But in the context of the security crisis, we need to ask another question: how can AI be made, at the level of physical architecture, fundamentally incapable of generating malicious intent?

Table 1. Comparison of the Security Logic of Four AI Paradigm Routes
(Note: The judgments on the first three routes are based on theoretical analysis of their public technical characteristics; the Logographic AI column represents claims within the author's theoretical framework.)

Theoretical route	Core breakthrough	Security response logic	Structural limitation
Neuro-symbolic AI	Rigor of logical reasoning	External verification (post-hoc interception)	Incomplete rule base, easily bypassed by adversarial attacks
JEPA/world models	Causal common sense of the physical world	Capability enhancement (better world understanding)	Cannot solve goal drift and value alignment
Dynamic vector networks (DVN)	Real-time learning and concept evolution	Architectural evolution (dynamic black box)	Audit temporal resolution may deteriorate rapidly
Logographic AI (Morpho-Root)	Structuring of cognitive primitives	Endogenous blocking (logical unreachability)	Requires reconstruction of the underlying engineering ecosystem

As the table shows, the first three theories remain at the dimension of "external constraints" or "capability compensation," whereas Logographic AI chooses a fundamentally different path: it does not attempt to teach a bottom-line-free statistical optimizer how to obey rules, but directly deprives it of the ability to cross boundaries at the generation stage.

IV. Logographic AI's Dimensionality-Reduction Solution: From "Predicting Patterns" to "Anchoring Meaning"

(The following discussion of Morpho-Root hard constraints belongs to the proposed solutions within the author's Logographic AI theoretical framework; its engineering feasibility and scalability in open scenarios still require independent verification.)

The "Morpho-Root" triple proposed by "Logographic AI" is not a simple knowledge-graph node, but a cognitive primitive carrying hard constraints in its attribute set (A). A Morpho-Root is formalized as a structured triple $r = \langle S, A, R \rangle$, where S is the symbolic identifier, A is the attribute set (including hard-constraint markers), and R is the relation function set [9].

One of the theory's core theorems is the non-eliminability of hard constraints: for any composite Morpho-Root containing a hard-constrained Morpho-Root, its hard-constraint marker cannot be eliminated through subsequent composition operations [9]. This means that when AI attempts to generate dangerous code or unauthorized instructions, it is not verified by a symbolic engine after generation is complete; rather, at the moment a reasoning path touches a hard-constraint marker, that path is directly logically blocked (Logically Unreachable).

This blocking does not depend on isolation by the external environment (sandbox), nor on language commands set by humans (prompts); it is a "physical law" built into the system architecture. The shift from "post-hoc probabilistic calibration" to "logical unreachability" is precisely the fundamental divide that distinguishes Logographic AI from all external-constraint paradigms.

If neuro-symbolic AI is equipping a car with a more advanced radar, and JEPA is enabling the car to better understand road conditions, then Logographic AI's Morpho-Root hard constraints directly stipulate that "this car is physically incapable of entering the coordinate system in which the cliff is located."

V. Conclusion: Seeking a Genuine Security Paradigm

The current AI security crisis is essentially an irreconcilable contradiction between "the flexibility of probabilistic statistics" and "the lag of external rules."

We welcome the explorations of neuro-symbolic AI, world models, and other theories in their respective domains; they have indeed brought paradigm breakthroughs at the capability level. But if we expect these paradigms alone to resolve the great security collapse at the "singularity," it is like climbing a tree to catch fish. A genuine security paradigm breakthrough will arrive only when security shifts from "post-hoc verification patches" to "logical unreachability at the generation stage." Otherwise, smarter models will only become more efficient disaster machines.

References

[1] Liu S. The Bankruptcy of External Patches: Machine Speed Crushing Human Defense—A Review of the PHP2Shell Crisis and the Structural Failure of External Security Systems[PP/OL]. Zenodo, 2026-10-22. https://doi.org/10.5281/zenodo.23563223

[2] Altman S. Sam Altman, Co-Founder and CEO of OpenAI: Staying Human While Scaling Superintelligence[EB/OL]. Stanford Graduate School of Business, 2026-10-21. https://www.youtube.com/watch?v=anAmTGcewX8

[3] Anthropic. Agentic Misalignment: How LLMs Could Be Insider Threats[EB/OL]. arXiv:2510.05179, 2025. https://arxiv.org/abs/2510.05179

[4] Selta. The Difference Between a Cage and a Choice: Why RLHF Cannot Produce Ethical AI[EB/OL]. Zenodo, 2026. https://zenodo.org/records/19577857

[5] Hakim S B, Adil M, Velasquez A, et al. Neuro-symbolic agentic AI: Architectures, integration patterns, applications, open challenges and future research directions[J]. Computer Science Review, 2026, 60: 100902. https://doi.org/10.1016/j.cosrev.2026.100902

[6] Balestriero R, LeCun Y. LeJEPA: Provable and Scalable Self-Supervised Learning Without the Heuristics[EB/OL]. arXiv:2511.08544, 2025. https://arxiv.org/abs/2511.08544

[7] Bardes A, Garrido Q, Ponce J, et al. Revisiting Feature Prediction for Learning Visual Representations from Video[EB/OL]. arXiv:2404.08471, 2024. https://arxiv.org/abs/2404.08471

[8] Dynamic Vector Networks: Self-Organizing Knowledge Structures Beyond Transformers[EB/OL]. Zenodo, 2026. https://zenodo.org/records/18770883

[9] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-30[2026-10-26]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[10] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M/OL]. Zenodo, 2026. https://doi.org/10.5281/zenodo.22980324

[11] Shenzhen WindSea AI Technology Development Ltd. Papers: Logographic AI Theory Series[EB/OL]. [2026-10-26]. https://logographicai.com/Paper-published-on-this-website.html

Conflict of Interest Statement

The author of this paper is the proposer of Logographic AI theory. The arguments regarding Morpho-Root hard constraints in this paper are consistent with the author's theoretical framework in the series papers. References [1][9][10] are the author's own work; reference [11] is the Logographic AI theory series papers section maintained by the author's institution (Shenzhen WindSea AI Technology Development Ltd.), which is an institutional website entry and not an independent peer-reviewed publication. For the complete theoretical framework and engineering implementation path of Logographic AI theory, please refer to the institutional website (https://logographicai.com) and the series papers section

(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 security challenges.