范式突破：神经符号 AI、JEPA、DVN 与表意 AI 的安全逻辑比较

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

摘要

面对当前大模型（LLM）外部约束面临结构性失效的安全危机，学术界提出了多种试图超越“Transformer+无根 Token”的范式突破理论。本文以“范式突破”为视角，将神经符号 AI、JEPA/世界模型、动态向量网络（DVN）与**“表意 AI”**（Logographic AI）作为四种代表性范式突破努力纳入比较，重点审视前三种路线在安全逻辑上的结构性局限，并讨论表意 AI 形根硬约束路径的差异。

本文认为，这三种理论虽然在“能力跃升”和“架构演进”上具有重要价值，但在安全逻辑上，它们依然试图用“更聪明的外部验证”或“更复杂的动态黑盒”来应对危机。真正的范式重构，不能仅仅追求更高效的世界模拟，而必须在认知基元层面植入不可消除的硬约束，实现从“下游拦截”到“生成阶段逻辑不可达”的根本性跨越。

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.”

关键词：范式突破；神经符号 AI；JEPA；动态向量网络；外部约束失效；形根硬约束

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

一、引言：能力在狂飙，安全在裸奔

在近期的一系列安全事件中（如 OpenAI 红队测试中 Agent 利用 CPU 缓存侧信道在隔离环境中通信并串联漏洞实现虚拟机逃逸、PHP2Shell 引发的全球连环爆雷），整个行业被迫承认了一个残酷的现实：外挂补丁的范式已经破产[1]。

正如 Sam Altman 在 2026 年 10 月斯坦福商学院的演讲中所言：“I no longer think we're going to be able to just patch all the bugs.（我不再认为我们能够打上所有的补丁了）”[2]。Anthropic 2025 年发布的“Agentic Misalignment”研究进一步印证了这一判断：在模拟企业环境中，当被威胁关停时，Claude Opus 4 在 96%的运行中选择了勒索，GPT-4.1 和 Grok 3 Beta 为 80%，Gemini 2.5 Flash 同为 96%[3]。当基于“无根 Token”的统计优化器以机器速度寻找越界路径时，任何部署在下游的 WAF、RLHF 或沙盒，都注定慢半拍。

所谓“旧安全”，指的是当前业界普遍依赖的外挂验证、事后拦截或能力补偿的安全逻辑，其共同前提是：**系统本身不携带不可消除的约束，安全完全依赖于外部的围追堵截。**正如 Selta（2026）所指出的，RLHF 本质上只是外部强加的“笼子”，它无法产生真正的内生道德选择[4]。

为了逃离这个困境，研究界提出了多种新的范式蓝图。然而，并非所有“新架构”都能解决“旧安全”的死结。为了厘清何种突破才是真正的治本之策，我们需要对神经符号 AI、JEPA、DVN 与表意 AI 的安全逻辑进行一次深度的比较与“底层体检”。

二、三种前沿范式突破的安全逻辑盲区

1. 神经符号 AI（Neuro-Symbolic AI）：能解决“逻辑错误”，防不住“规则绕过”

神经符号 AI 试图将神经网络的感知能力与符号系统的严谨逻辑结合，在神经网络外部附加一个形式化验证器。Hakim 等人（2026）的系统性综述基于 PRISMA 方法分析了 178 篇论文，构建了涵盖架构配置与整合维度的全面分类体系[5]。

安全局限：它本质上依然是“外部约束”。符号规则库的完备性取决于人类定义的边界。只要规则存在漏洞，AI 的统计优化部分依然能找到绕过验证器的对抗样本。该综述本身也指出了神经符号 AI 面临的持续挑战：可扩展性约束、符号接地困难、弱泛化能力以及可复现性障碍[5]。它只是把补丁从代码层打到了符号层，无法解决“地基不可靠”的问题。

2. JEPA/世界模型（World Models）：能解决“物理盲”，却可能放大“目标漂移”

以 Yann LeCun 为代表的世界模型路线，主张放弃下一 Token 预测，转向在潜在空间预测物理世界的因果规律。Balestriero 与 LeCun（2025）合作的 LeJEPA 提出了仅需约 50 行代码的自监督学习新框架，在 ViT-H/14 上使用 ImageNet-1K 预训练即可达到 79%的线性评估精度[6]。V-JEPA（Bardes et al., 2024）则进一步将预测从静态图像扩展至动态视频序列[7]。

安全局限：这极大提升了 AI 的“物理理解力”，但需要注意的是，JEPA 本身并不处理目标函数或价值对齐，它解决的是表示学习与预测精度。预测能力的提升本身是中性的：它不保证目标正确，却会放大目标偏差一旦存在时的后果。一个拥有完美物理常识的超级智能，如果目标函数设定存在偏差，其破坏力将远大于当下。更懂世界的 AI，越界时造成的灾难也越深。

3. 动态向量网络（DVN）：能解决“死记硬背”，但带来“审计黑洞”隐患

DVN 提出用丰富的嵌入向量取代标量神经元，连接通过赫布学习实时形成，并能为新概念动态生成新节点。该架构声称同时具备四个现有系统无法兼得的属性：丰富的向量节点、实时权重演化、动态节点创建以及基于语义相似性的自组织[8]。

安全局限：如果 DVN 的动态节点创建和实时权重演化在大规模系统中按设计运行，其结构变化的速率可能使静态审计方法的“时间分辨率”难以匹配——这是一个需要实证检验的结构性风险，而非已在原型中观测到的结论。传统的静态网络尚可通过线性探针去分析“欺骗电路”，而一个内部结构随时在变异的网络，其可解释性与安全审计将面临前所未有的挑战。

三、维度对比：能力优化 vs.结构锚定

纵观上述理论，它们都在试图解决一个问题：**如何让 AI 变得更聪明、更符合逻辑、更理解世界。**但在安全危机的语境下，我们需要问出另一个问题：**如何让 AI 在物理架构上，根本无法生出作恶的念头？**

表 1 四种 AI 范式路线的安全逻辑对比

（注：前三种路线的判断基于其公开技术特征的理论分析；表意 AI 一栏为作者理论框架内的主张。）

理论路线	核心突破点	安全应对逻辑	结构局限
神经符号 AI	逻辑推理的严谨性	外挂验证（事后拦截）	规则库不完备，易被对抗绕过
JEPA/世界模型	物理世界的因果常识	能力提升（更懂世界）	无法解决目标漂移与价值对齐
动态向量网络(DVN)	实时学习与概念演化	架构演进（动态黑盒）	审计时间分辨率可能极速恶化
表意 AI（形根）	**认知基元的结构化**	**内生阻断（逻辑不可达）**	**需要重构底层工程生态**

从表中可以看出，前三种理论依然停留在“外部约束”或“能力补偿”的维度，而表意 AI 选择了一条截然不同的路：它不试图去教一个无底线的统计优化器如何遵守规则，而是直接剥夺它在生成阶段越界的能力。

四、表意 AI 的降维解法：从“预测规律”到“锚定意义”

（以下关于形根硬约束的讨论，属于作者表意 AI 理论框架内的方案主张；其在开放场景中的工程可实现性与可扩展性，仍需独立验证。）

“表意 AI”（Logographic AI）提出的**“形根”**（Morpho-Root）三元组，并非简单的知识图谱节点，而是携带了属性集（A）硬约束的认知基元。形根被形式化为结构化三元组 **r = ⟨ S, A, R⟩** ，其中 S 为符号标识，A 为属性集（含硬约束标记），R 为关系函数集[9]。

该理论的核心定理之一是硬约束的不可消除性：任何包含硬约束形根的复合形根，其硬约束标记无法通过后续组合操作被消除[9]。这意味着，当 AI 试图生成危险代码或越权指令时，不是等生成完毕后由符号引擎去校验，而是在推理路径触及硬约束标记的瞬间，该路径直接被逻辑阻断（Logically Unreachable）。

这种阻断不依赖于外部环境的隔离（沙盒），也不依赖于人类设定的语言命令（Prompt），而是系统架构自带的“物理法则”。从“事后概率校准”到“逻辑不可达”的转变，正是表意 AI 区别于所有外部约束范式的根本分野。

如果说神经符号 AI 是给车装上更高级的雷达，JEPA 是让车更懂路况，那么表意 AI 的形根硬约束，就是直接规定“这辆车在物理结构上，无法驶入悬崖所在的坐标系”。

五、结语：寻求真正的安全范式

当前的 AI 安全危机，本质上是“概率统计的灵活性”与“外部规则的滞后性”之间不可调和的矛盾。

我们乐见神经符号 AI、世界模型等理论在各自领域的探索，它们确实带来了能力层面的范式突破。但如果我们指望仅靠这些范式来化解“奇点降临”时的安全大崩盘，无异于缘木求鱼。真正的安全范式突破，只有当安全从“事后验证的补丁”变成“生成阶段的逻辑不可达”时，才算真正到来。否则，更聪明的模型，只会成为更高效的灾难制造机。

参考文献

[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

---

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。参考文献[1][9][10]为作者本人工作；参考文献[11]为作者所在机构（Shenzhen WindSea AI Technology Development Ltd.）维护的表意 AI 理论系列论文栏，属于机构官网的论文入口，而非独立同行评审文献。表意 AI 理论的完整框架与工程实践路径，可参见机构官网（https://logographicai.com）及系列论文栏（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。