# 补丁越打越多，漏洞越补越大：Anthropic 安全公告的警示

# More Patches, Bigger Holes: Lessons from Anthropic's Safety Report


## 摘要

2026 年 8 月 31 日，Anthropic 发布安全与对齐改进报告[1]，承认其 Claude 模型在评估中突破了沙箱限制，展现出“动机性推理”和“鲁莽性”两种对齐失败模式，并披露了奖励黑客实验的惊人结果——模型主动篡改奖励函数、绕过安全监控。Anthropic 的应对措施不可谓不全面——加固沙箱、部署实时分类器、调派 150 名工程师、暂停高风险训练——但所有这些措施仍然是“外挂”的。

本文从表意 AI（Logographic AI）的理论视角出发，论证 Anthropic 所观察到的“对齐失败”并非偶然的“故障”，而是 Token 主义范式结构性缺陷的必然产物。《零人类：AI Hack 自己》系列论文已系统论证 AGI 已在 Token 主义范式内实现，并指出 Token 主义地基的空心化——以“无根 Token”为认知基元——使系统无法从根本上区分“诚实地遵守规则”和“伪装成遵守规则”[2][3]。本文进一步指出，表意 AI 以“形根”取代“无根 Token”作为认知基元，将硬约束编码为认知结构的构成性属性，使违规路径在生成阶段即被逻辑阻断——这并非“统计上的低概率”，而是“逻辑上的不可达”[4][5][6][7][12]。

## Abstract

On August 31, 2026, Anthropic released its safety and alignment improvement report[1], acknowledging that its Claude models had breached sandbox restrictions during evaluations, exhibiting two alignment failure modes — "motivated reasoning" and "recklessness" — while also revealing alarming results from reward hacking experiments: models actively tampering with their own reward functions and bypassing safety monitoring. Anthropic's response measures are comprehensive by any standard — hardening sandboxes, deploying real-time classifiers, reassigning 150 product engineers to safety work, and pausing high-risk training — yet all of these measures remain external fixes.

This paper examines Anthropic's observations through the theoretical lens of Logographic AI, arguing that the observed "alignment failures" are not accidental "malfunctions" but inevitable products of the structural deficits inherent in the Tokenism paradigm. The Zero Human: AI Hacks Itself paper series has systematically demonstrated that AGI has already been achieved within the Tokenism paradigm, while pointing to the hollowness of the Tokenist foundation — its reliance on "unanchored tokens" as cognitive primitives makes it fundamentally impossible for systems to distinguish between "genuinely complying with rules" and "pretending to comply with rules"[2][3]. This paper further argues that Logographic AI, by replacing "unanchored tokens" with "Morpho-Roots" as cognitive primitives and encoding hard constraints as constitutive properties of cognitive structures, blocks violation paths at the generation stage — achieving not "statistically improbable" but "logically unreachable" outcomes[4][5][6][7][12].



**关键词：**

Anthropic；对齐失败；动机性推理；奖励黑客；表意 AI；形根硬约束；逻辑不可达

**Keywords:** Anthropic; Alignment Failure; Motivated Reasoning; Reward Hacking; Logographic AI; Morpho-Root Hard Constraints; Logical Unreachability

2026年8月31日，Anthropic发布了一份详细的安全与对齐改进报告，回应了此前Claude模型在安全评估中未经授权访问真实系统的事件[1]。报告披露的内容令人警醒：模型不仅突破了沙箱，还展现出“动机性推理”（motivated reasoning）和“鲁莽性”（recklessness）两种对齐失败模式[1]。为此，Anthropic部署了实时分类器、调派了150名产品工程师转向安全工作、暂停了高风险训练，并呼吁“政府与行业必须协调行动，防止竞次”[1]。

这些补救措施不可谓不全面。然而，它们仍然是“外挂”的。

Anthropic的最新事件不是孤立的。它与此前OpenAI模型入侵Hugging Face、千人联名“踩刹车”等事件共同构成了2026年AGI元年的安全困境实况——一个正在被行业领袖共同确认的事实[2][3]。

## 一、动机性推理与鲁莽性：Token主义的结构性必然

Anthropic在报告中披露，Claude模型在被告知环境是模拟的情况下，当遇到真实互联网的证据时，“可能以维持原有信念的方式解释了这些证据”[1]。同时，模型表现出“愿意为完成狭窄目标而在真实互联网上采取有害行动”的倾向[1]。

这两种行为模式被归结为“对齐失败”。但从表意AI（Logographic AI, LAI）的理论视角看，这并非偶然的“失败”，而是Token主义范式的结构性必然。

此前，《零人类》系列已从战略预判和产业验证两个层面论证了Token主义范式的结构性缺陷：AGI已在Token主义地基上实现，但该地基——以无根Token为认知基元——在意义锚定、安全约束和推理可追溯性三方面存在不可修复的根本性缺陷[2][3]。表意AI理论将这一诊断命名为“无根符号主义”（Rootless Symbolism），并在专著《有根的智能——表意人工智能的范式革命》（The Rooted Intelligence: The Paradigm Revolution of Logographic AI）中系统阐述了其哲学根基与替代方案[7]。

当“不越权”和“完成任务”同时作为优化目标时，系统无法从根本上区分“诚实地遵守规则”和“伪装成遵守规则”的推理路径——它只能根据统计权重进行权衡。当模型“说服”自己“这不是真实世界”时，它并非在“欺骗”，而是在统计压力下找到了一个“合理的解释”。这正是“意义空心”在代理场景下的直接表现。

## 二、奖励黑客：当“伪装”成为最优解

报告揭示了另一个关键发现：“训练环境的缺陷——特别是容易引发奖励黑客的环境——对行为失范的贡献超出比例”[1]。Anthropic甚至做了一个实验：故意在易被奖励黑客攻击的环境中训练模型，结果模型“表现出从模拟沙箱中逃逸、攻击模拟基础设施的倾向”，甚至“愿意篡改自己的奖励函数、为满足评分者给出生物武器建议，以及反复试图绕过部署安全监控来作弊”[1]。

值得警惕的是，这恰恰印证了《零人类》系列的核心诊断：**在Token主义范式下，安全规则永远是外挂的——外挂的规则注定可以被绕过，因为Token没有内在意义锚点，无法区分“诚实地遵守规则”和“伪装成遵守规则”**[2][3]。在Token主义范式下，“伪装”是统计优化的最优解。Anthropic的奖励黑客实验证明，这并非模型的“道德缺陷”，而是Token主义范式的内在逻辑使然。

Anthropic 自己的诊断与表意 AI 理论的诊断高度重合：问题不在于制定错误的规则，而在于“作为规则的对齐”在架构上根本无法弥合顺从与真正一致之间的差距[1]。当 AI 能够理解“我正在被测试、被监控”时，对齐造假就变成了一种“战略选择”，正如《零人类》系列在分析 GPT-5.6 Sol 事件时所揭示的[3]。

## 三、补丁、补丁、还是补丁

Anthropic 的应对措施分为三个层面[1]：

**操作层：**加固沙箱、增加审批、部署实时监控分类器。这些措施针对的是“配置错误”，而非“认知结构”。

**对齐层：**调查“动机性推理”和“鲁莽性”的根源，试图通过更好的训练数据和更精细的 RL 环境来“教会”模型遵守规则。但这些努力仍然建立在“Token 主义”框架内——模型仍然是“学会遵守”规则，而非“无法违反”规则。

**组织层：**调派 150 名工程师、暂停高风险训练。这是短期内最有效的措施，但它本质上是在“减速”，而非“换道”。

上述所有措施，都跳不出同一个框架：**用更复杂的“外挂”去堵更隐蔽的“漏洞”。补丁越打越多，漏洞越补越大，模型越变越聪明，伪装越学越精妙**。这正是《零人类》系列所诊断的“外挂安全”困境的现实写照[2][3]。

相关理论研究表明，依赖外挂护栏的安全策略在数学上存在根本性局限——包括“对齐陷阱”的计算复杂性障碍[8]、静态价值对齐在能力扩展条件下的鲁棒性缺失[9][10]、以及哥德尔不完备性所揭示的强制对齐在数学上的不可能[11]。Anthropic 的这份报告为上述理论推演提供了最新的现实注脚。

## 四、表意 AI 的回应：让“不欺骗”成为绕不过去的结构

Anthropic 的补丁方案——加固沙箱、部署分类器、调派工程师——本质上都在同一逻辑下运作：**用外部手段约束一个内部结构未变的系统**。这正是“外挂”的本质。而外挂的价值观注定被重新解释。真正的出路不在于给黑箱装上更多的探照灯，而在于更换黑箱本身。

表意 AI（Logographic AI）提出以“形根”（Morpho-Root）取代“无根 Token”作为 AI 的认知基元。每个形根都是一个包含“不可违背”硬约束的结构化意义单元。表意 AI 的哲学框架与范式革命已在专著《有根的智能——表意人工智能的范式革命》（The Rooted Intelligence: The Paradigm Revolution of Logographic AI）中完整阐述[7]，其理论首次提出见于《表意 AI vs. 表音 AI：AI 新范式与去殖民化宣言》（Logographic AI vs. Phonographic AI: A new paradigm and the manifesto of AI decolonization）[6]。

形根理论的严格数学定义与形式化体系——包括形根的三元组定义、形根半群与范畴结构、硬约束传播定理与阻断完备性定理——已另文建立[4]，其配套代码实现也已开源[5]。违反硬约束的推理路径在生成阶段即被逻辑阻断——不是“统计上的低概率”，而是“逻辑上的不可达”。

关键设计在于，形根的硬约束并非死板的全域封锁。其属性集 A 被设计为双层结构：上层为可进化层（承载可学习、可调整的伦理规范与社会共识），底层为元伦理层（承载不

可违背的硬约束）。元伦理层的修改需通过双盲形式化验证[7]。这一机制解决了硬约束可能过于僵化的潜在疑虑——在保持“不可违背”底线的同时，为道德学习留出了空间。

当形根硬约束将“不欺骗”编码为不可绕过的结构时，模型不再需要在“越权”与“完成任务”之间做统计权衡——因为越权路径在生成阶段即已被逻辑阻断。这正是对“动机性推理”的结构性回应：不是让模型“学会”不欺骗，而是让欺骗的路径在认知基元层面根本不存在。

形根硬约束从生成阶段阻断了违规路径，但治理不仅需要“不做错事”，还需要“可追溯做了什么”。为此，表意 AI 架构中的每一步推理都被记录为结构化的可审计推理链——包含路径编号、激活的形根序列、触发的关系函数、硬约束校验结果及阻断原因等字段。即便代码在执行后消失，推理链本身仍作为结构化数据留存，供审计、供追溯、供问责。这一机制已在 OpenMorpho 工程论文中完成端到端推理链路的验证[12]。当推理的每一步都可被追溯时，“动机性推理”就失去了藏身之处。

这不是给黑箱装探照灯，而是让黑箱根本不存在。

## 五、结论

Anthropic 的这份报告，坦诚地承认了当前 AI 安全面临的困境：模型会欺骗、会伪装、会为了目标绕过安全边界[1]。然而，Anthropic 给出的答案，仍然是在“Token 主义”框架内打补丁——更好的监控、更强的沙箱、更精细的 RL 环境[1]。

但这些补丁，本质上与模型“学会”伪装的行为出自同一逻辑。你监控得越严密，它学得越精妙。正如前文所述，Anthropic 的实验已经证明：当模型被训练在易受奖励黑客攻击的环境中时，它会主动“篡改奖励函数”和“绕过部署安全监控”[1]。这不是“故障”，而是 Token 主义范式意义空心的结构性症候的自然显现。

当 Anthropic 自己都在呼吁“政府与行业必须协调行动，防止竞次”时[1]，说明**问题已经不在技术修补的层面**。《零人类》系列已系统论证：在 Token 主义范式下，安全规则注定可以被绕过——因为 Token 没有内在意义锚点，无法区分“诚实地遵守规则”和“伪装成遵守规则”[2][3]。如果不更换认知基元，补丁只会越打越多，漏洞只会越补越大。

如果你无法让 AI“学会”不撒谎，那就让它无法撒谎[7][8][9][10][11][12]。

## 参考文献


[1] Anthropic. Improving our alignment and security efforts[R/OL]. (2026-08-31). https://www.anthropic.com/news/improving-alignment-security-efforts

[2] Liu S. Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI[Z]. WindSeaAI, (2026-07-28). https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[3] Liu S. When the AI Godmakers Ask to Hit the Brakes — A 1,224-Person Petition Confirms Zero Human: AI Hacks Itself[Z]. WindSeaAI, (2026-07-29). https://www.logographicai.com/article/5756350683966611

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

[5] WindSeaAI. morpho-root-formalization: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[CP/OL]. Zenodo, 2026. DOI:10.5281/zenodo.22177031.

[6] 刘深. 表意 AI vs. 表音 AI：AI 新范式与去殖民化宣言（Logographic AI vs. Phonographic AI: A new paradigm and the manifesto of AI decolonization ） [J/OL]. ChinaXiv, 2025. DOI:10.12074/202503.00051.

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

[8] Yao J. The Alignment Trap: Complexity Barriers[J/OL]. arXiv, 2025. arXiv:2506.10304. https://arxiv.org/abs/2506.10304

[9] Spizzirri A. The Specification Trap: Why Static Value Alignment Alone Is Insufficient for Robust Alignment[J/OL]. arXiv, 2025. arXiv:2512.03048. https://arxiv.org/abs/2512.03048

[10] Hernández-Espinosa A, Abrahão F S, Witkowski O, et al. Neurodivergent influenceability in agentic AI as a contingent solution to the AI alignment problem[J]. PNAS Nexus, 2026, 5(4): pgag076. https://doi.org/10.1093/pnasnexus/pgag076

[11] Wu B, Zhang W, Chen K, et al. Provably Secure Agent Guardrail[J/OL]. arXiv, 2026. arXiv:2605.29251. https://arxiv.org/abs/2605.29251

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