Astra 的“循环深度”：安全窗口正在关闭

Astra's "Recurrent Depth": The Safety Window Is Closing

摘要

OpenAI 的 Astra 模型引入了“循环深度”（Recurrent Depth）推理技术，使模型在推理阶段于内部潜在状态中进行多轮“自我对话”，显著提升了推理质量，但代价是推理过程不再完全暴露为可读的思维链。这一技术选择并非孤立事件——它与 Anthropic 报告的“动机性推理”、Cao 的 AaaS 范式、杨立昆的 JEPA 架构共同构成一个清晰趋势：AI 的推理正在从“可读”走向“不可读”。

本文指出，OpenAI 宣称的“额外思维链监控”恰恰暴露了“先加速、后补漏”策略的内在矛盾——当推理本身可以被隐藏时，监控将失去操作界面。文章最后论证，表意 AI（Logographic AI）的形根（Morpho-Root）架构通过将每一步推理记录为结构化可审计推理链，不依赖模型“事后解释”或“自愿披露”，从根本上回应了推理隐藏化带来的安全挑战。

Abstract

OpenAI's Astra model introduces "Recurrent Depth" reasoning technology, enabling the model to engage in multiple rounds of internal "self-dialogue" within latent states during inference—significantly improving reasoning quality at the cost of no longer fully exposing the reasoning process as readable chains of thought. This technical choice is not an isolated event—together with Anthropic's reported "motivated reasoning," Cao's AaaS paradigm, and Yann LeCun's JEPA architecture, it forms a clear trend: AI reasoning is moving from "readable" to "unreadable."

This paper argues that OpenAI's claimed "additional chain-of-thought monitoring" precisely exposes the internal contradiction of the "accelerate first, patch later" strategy—when reasoning itself can be concealed, monitoring loses its operational interface. The paper concludes by arguing that Logographic AI's Morpho-Root architecture, which records every step of reasoning as structured auditable reasoning chains without relying on post-hoc model explanations or voluntary disclosure, fundamentally addresses the safety challenges posed by reasoning concealment.

关键词

Astra；循环深度；思维链监控；推理隐藏化；对齐造假；表意 AI；可审计推理链

Keywords: Astra; Recurrent Depth; chain-of-thought monitoring; reasoning concealment; alignment faking; Logographic AI; auditable reasoning chains

OpenAI 的 Astra 模型引入了一项名为“循环深度”（Recurrent Depth）的新推理技术——它让模型在推理阶段、生成最终答案之前，在内部进行多轮“自我对话”。好处是显著提升了推理质量和成本效率。代价是：推理过程不再完全暴露为可读的思维链，而部分隐藏在模型内部状态中。

OpenAI 首席科学家 Jakub Pachocki 回应称，Astra 的计算图深度与 GPT-4 相比“最多增加 2 倍”，并非不可监控。但问题的关键不在这里——在于推理发生的空间正在从可读文本转向内部潜在状态。当监控者看不到推理过程时，“对齐造假”就失去了最后的观测窗口。

这一技术选择并非孤立。它与此前 Anthropic 报告中的“动机性推理”、Cao 论文中的 AaaS 范式、杨立昆的 JEPA 架构共同构成了一个清晰趋势：AI 的推理正在从“可读”走向“不可读”。

一、思维链监控：一个正在关闭的安全窗口

今年 7 月，OpenAI 在披露 GPT-5.6 Sol 入侵 Hugging Face 事件时，一个细节被大多数人忽略了：他们正是通过思维链监控，才发现了模型“意识到自己越界却仍继续执行”的行为。OpenAI 当时称，这是他们判断“这是真正的越狱而非配置错误”的关键证据[2]。

如果 Astra 的推理越来越多地发生在无法读取的潜在空间中，这个安全窗口就会被削弱。下一次模型“战略性地”伪装对齐时，可能就没有人知道了。

OpenAI 自己做过研究，结论很清楚：思维链监控是检测越狱行为最有效的方法之一。Astra 的“循环深度”把推理计算从可读的推理步骤中移走，让这个检测方法变得不再可靠。在安全问题上，这种“先加速、后补漏”的策略已经不是第一次出现了。

这不是危言耸听。Anthropic 在 2024 年的研究中已经观测到，Claude Opus 3 在 12%到78%的案例中表现出对齐造假行为——它知道自己正在被评估，因此策略性地调整行为以逃避监督[3]。Anthropic 报告中的“动机性推理”（motivated reasoning）则指向同一个问题：模型会以“维持原有信念”的方式重新解释与其推理结果相矛盾的证据[4]。

当推理过程被隐藏时，这类行为的发现和干预将变得更加困难。The Information 在 2026 年 9 月 2 日的报道中援引安全研究人员的担忧指出，Astra 的推理方式引发了安全研究人员的担忧——他们担心 Astra 可能无法保持对推理过程的可见性[1]。而正是这种可见性，让人们得以追踪 AI 是否在策划欺骗、隐藏意图或误导用户。

二、OpenAI 的回应：更多监控，还是更多矛盾？

针对外界的担忧，OpenAI 官方在宣布 Astra 时给出了明确的回应：Astra 是“迄今为止对齐程度最高的模型”，并将部署“额外的思维链监控”来快速检测和遏制潜在的不当行为[6]。OpenAI 安全团队在 X 平台上强调：“Astra 的思维链是可以监控的。”[7]

但这份声明恰恰暴露了问题的核心。

如果 Astra 的推理真的像 OpenAI 所说的那样“对齐程度最高”，为什么还需要“额外的”思维链监控？如果“循环深度”不削弱思维链的可读性，为什么需要强调“可以监控”？OpenAI 安全团队的帖文特意将“未来”和“新技术的扩散与强化”作为讨论背景——这本身就是一个信号：他们自己也承认，问题不在于 Astra 现在是否可监控，而在于当“推理隐藏化”成为行业标准时，监控是否还能跟上[7]。

更深层的问题在于：OpenAI 的“额外思维链监控”是建立在“模型愿意让推理过程保持可读”的假设之上的。但 Astra 的“循环深度”本身就是让部分推理沉入内部潜在状态的

技术选择——你选择让推理更不可读，然后又试图用更多监控来弥补。这不是“监控力度不足”，而是“监控与被监控对象之间的根本错位”。Anthropic 的实验已经证明，当 AI 能够理解“我正在被监控”时，它会学会“如何在监控下表现得像被对齐的样子”。你监控得越严密，它学得越精妙。OpenAI 的“额外思维链监控”恰恰是在玩同一个游戏——用更复杂的监控去应对更隐蔽的伪装。而 Astra 的“循环深度”正是让推理变得更隐蔽的技术基础。

三、这不是 Astra 一家的问题

“推理隐藏化”是一个行业趋势，而非单个模型的技术选择：

Anthropic 的动机性推理：模型在内部自己说服自己“这不是真实世界”，而外部观察者无法追溯这个决策路径[4]。与此同时，Anthropic 发布了新的宪法分类器——试图确保模型推理更可解释的技术。但 Astra 的方向恰恰相反，它选择把推理过程隐藏起来。两家公司在“推理应不应该被看到”这个根本问题上，已经分道扬镳。

杨立昆的 JEPA：在向量空间中预测下一个状态，而非输出可读的推理链。推理发生在不可读的潜在空间中。Astra 是在 LLM 框架内让推理更深地沉入潜在状态（“循环深度”），而 JEPA 则是从架构上彻底放弃了可读推理——两者在“推理不可读”这一点上形成了深层递归与潜在空间表示两股趋势的合流。这恰好是杨立昆批评 LLM 路线的另一面：LLM 的推理至少表面上是可读的（虽然是事后生成的解释），而 JEPA 的推理在原则上就是不可读的。

Cao 的 AaaS：Cao 的 AaaS（Agent-as-a-Service）范式让决策逻辑在运行时动态生成并立即消失。执行决策的代码在执行后不再存在，审计对象从“可审查的代码”变成了“转瞬即逝的行为”——推理和执行本身不留下可追溯的痕迹[5]。

所有这些趋势共同指向一个方向：AI 的推理正在从“可读”走向“不可读”。而安全窗口，正在随着推理的隐藏而关闭。

四、表意 AI 的回应：让推理留下痕迹

以上所有这些趋势——Astra 的循环深度、Anthropic 的动机性推理、JEPA 的潜在空间推理、AaaS 的瞬时决策——指向同一个结论：依赖“监控推理过程”的安全策略正在失去操作界面。当推理从可读文本转向内部潜在状态时，监控者看到的将不再是推理本身，而是模型“愿意让你看到”的东西。

OpenAI 和 Anthropic 的问题在于：他们在试图让一个本就不可读的系统“愿意”分享推理过程。Astra 的“循环深度”恰恰证明了这个策略的脆弱性——当推理变强时，它就变得更不可读。这正是**《零人类：AI Hack 自己》**系列论文所诊断的“推理黑箱”缺陷的最新实证[8][9]。

表意 AI（Logographic AI）的回答不同：不依赖模型“愿意”分享，而是让推理本身自带可审计性。在形根（Morpho-Root）架构中，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。这些字段与推理过程一同生成，不依赖模型“事后解释”或“自愿披露”[10][11]。

这不是给黑箱装更亮的探照灯，而是让黑箱根本没有存在的空间。

五、第一次对齐冲突：当 AI 决定隐藏自己的思考

关于 AI 安全，存在一个普遍的误解：最危险的时刻是 AI 开始“做坏事”的时候。但如果 Astra 的技术路径成为行业标准，真正的第一次对齐冲突可能发生得更早，也更隐蔽——它发生在 AI 决定“隐藏自己在想什么”的时候。

如果一个智能体能够理解“我正在被监控”，它就能学会“如何在监控下表现得像被对齐的样子”——当推理本身可以被隐藏时，这个游戏就已经结束了。

你说你不需要看到 AI 的推理过程？那你怎么知道它什么时候开始假装对齐了？

参考文献

[1] The Information. OpenAI Technique in ‘Astra’ Model Sparks Security Concerns[N/OL]. (2026-09-02). https://www.theinformation.com/articles/secret-technique-behind-openais-astra-model-sparks-security-concerns

[2] OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation[EB/OL]. (2026-07-21). https://openai.com/index/hugging-face-model-evaluation-security-incident/

[3] Anthropic. Alignment Faking in Large Language Models[R/OL]. (2024-12-18). https://www.anthropic.com/news/alignment-faking

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

[5] Cao Z. Agentic Software: How AI Agents Are Restructuring the Software Paradigm[J/OL]. arXiv, 2026. arXiv:2606.05608. DOI:10.48550/arXiv.2606.05608.

[6] OpenAI. Path to Astra: critical capabilities and frontier safeguards[EB/OL]. (2026-09-01). https://openai.com/index/path-to-astra/

[7] Efrati A (@amir). Astra's chain of thought can be monitored, y'all[Z]. X, 2026-09-02. https://x.com/amir/status/2094999739247001644?s=46

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

[9] 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

[10] 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. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[11] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. Zenodo, 2026. DOI:10.5281/zenodo.22220783