认知架构与可问责性：从 Jev 路线看 AI“直觉决策”的范式盲区

Cognitive Architecture and Accountability: A Critique of the “Intuitive Decision” Paradigm from the Jev Route

摘要

2026 年 9 月，前 OpenAI 研究员、InstructGPT 论文合著者 Diogo Almeida 发布了其新公司 TypeSafe AI 的首个决策模型 Jev。该模型砍掉自然语言生成层，直接输出结构化决策与校准概率，声称在系统 1 任务上实现 20–200 倍提速与 40–400 倍降本。

本文从认知科学与 AI 可问责性视角对这一技术路线进行系统审视，提出三个核心判断：**第一**，自然语言并非模型的“思考过程”，而是模型进行复杂推理的“外部脚手架”，当前主流审计实践中，自然语言轨迹是最直接的抓手之一；**第二**，概率校准不等同于逻辑可靠，一个校准良好的概率分布仍然可能产生完全错误的决策；**第三**，当 AI 省略“思考痕迹”直接输出决策时，将导致责任追溯面临结构性困难。

本文进一步论证，上述代价的根源在于当前主流 AI 范式仍然建立在“无根 Token”的统计外推之上，而非基于内嵌逻辑约束的认知基元。要实现真正可问责的 AI 决策系统，需要从“事后概率校准”转向“逻辑不可达”的范式重构——这正是表意 AI（Logographic AI）以“形根硬约束”替代无根 Token 的核心主张。

Abstract

In September 2026, Diogo Almeida, former OpenAI researcher and co-author of the InstructGPT paper, released Jev, the first decision model from his new company TypeSafe AI. The model removes the natural language generation layer and directly outputs structured decisions and calibrated probabilities, claiming 20–200× speed improvements and 40–400× cost reductions on System 1 tasks.

This paper systematically examines this technical route from the perspectives of cognitive science and AI accountability, proposing three core judgments: first, natural language is not the model’s “thinking process,” but an “external scaffold” for complex reasoning; in current mainstream auditing practice, natural language traces are one of the most direct handles. Second, probability calibration is not equivalent to logical reliability; a well-calibrated probability distribution can still produce completely erroneous decisions. Third, when AI omits “traces of thought” and directly outputs decisions, it leads to structural difficulties in responsibility attribution.

This paper further argues that the root of these costs lies in the fact that the current mainstream AI paradigm is still built on the statistical extrapolation of “rootless Tokens,” rather than on cognitive primitives with embedded logical constraints. To achieve a truly accountable AI decision system, it is necessary to shift from “post hoc probability calibration” to a paradigm reconstruction of “logical unreachability”—which is precisely the core proposition of Logographic AI (LAI), replacing rootless Tokens with “Morpho-Root hard constraints.”

关键词：认知架构；可问责性；Jev 模型；RLCD；系统 1 决策；概率校准；对齐伪造；表意 AI

Keywords

Cognitive architecture; accountability; Jev model; RLCD; System 1 decision-making; probability calibration; alignment faking; Logographic AI

一、引言：效率革命与被回避的追问

2026 年 9 月 15 日，前 0penAI 研究员 Diogo Almeida 正式发布 TypeSafe AI 的首个模型 Jev，并开启早期访问。Almeida 曾深度参与 InstructGPT 与 ChatGPT 的核心技术研发[3]，其学术身份使这次发布带有独特的意味——这被视为 RLHF 范式的核心参与者对该范式的自我反叛。[1,2]

Jev 的技术路线与当前主流 LLM 形成鲜明断裂。传统大模型执行判断任务时，需要先生成自然语言或 JSON，再由外部软件解析执行。Jev 直接砍掉了这一中间层，接收非结构化状态输入，在单次前向传递中并行返回类型化决策——Choice（选项集）、Score（标量评分）和 Noul（0 到 1 的概率值，表示模型对决策的置信度），每个输出均附带置信度。TypeSafe 声称，在安全警报、Agent 审查、发票处理和客户服务等四类自动化工作流场景中，Jev 比 Claude Sonnet 5 快约 193.6 倍，比 Claude Opus 5 便宜约 444.6 倍。端到端响应时间压缩至 70 毫秒至 500 毫秒，输入定价仅为每百万 token 0.042 美元，输出 token 则完全免费。[2,8]

这套效率提升的物理基础在于架构层面的根本改变：传统 LLM 的 token 生成是自回归的，每个 token 都以前一个 token 为条件，生成速度受限于串行链路的固有延迟；而 Jev 的并行采样器架构在单次前向传递中同时对多个预设判别项完成评估，增加评估字段几乎不增加延迟。[2,7]

然而，在这场效率狂欢中，**一个根本性的问题被所有人回避了：当 AI 不需要“说出思考过程”就能做出决策时，它到底是更聪明了，还是彻底变成了一个无法审计的隐秘黑盒？**

本文试图回答这一问题。全文结构如下：第二节从认知科学视角论证自然语言在 AI 推理中的功能定位；第三节分析 Jev 效率革命的技术本质；第四节系统阐述省略“思考痕迹”的三重代价；第五节回到范式层面讨论认知架构与可问责性的深层关系。

二、自然语言不是思考过程，而是审计的抓手

人类有一个根深蒂固的错觉：思考就是内心的自言自语。但认知科学早已证明，思考的形态远不止语言。棋手落子前的直觉判断、老司机躲避危险时的肌肉记忆反应、数学家闪现的几何直觉——这些都不依赖自然语言作为中介[6]。

大模型也是如此。大模型真正的“思考”发生在高维向量空间的数学运算中，输出自然语言本质上是在用序列化的方式，将高维运算“翻译”成人类可读的低维符号。这就是为什么大模型需要“思维链”（Chain of Thought）。CoT 并非模型思考的过程，而是模型用来“打草稿”的外部脚手架[3]——由于无法在单次前向传播中完成复杂推理，模型被迫写出中间步骤，再顺着自己写的文字往下计算。

人类说出来的话，是思考的结果，不是思考的过程。AI 说出来的话，是计算的包装，不是决策的必需。但**“必需”的反面——冗余——恰恰蕴含着重要的工程价值。自然语言输出虽然增加了延迟和成本，却提供了一个可追溯的推理轨迹。**当模型输出“我判断这笔交易涉嫌欺诈，因为金额异常且收款方位于高风险地区”时，这个自然语言序列成为外部审计者理解模型决策逻辑最直接的入口之一。思维链可能是事后编造的，但它至少提供了一个可检验的解释框架。

Jev 砍掉自然语言生成层，在效率上取得了巨大成功，但同时砍掉了这一审计抓手。当 Jev 输出 Noul = 0.92 时，它没有告诉任何人：这个 0.92 是基于哪些特征计算得出的？哪些因素可能改变这个判断？在什么条件下这个判断可能失效？

没有草稿纸，我们失去了一个可直观阅读的决策轨迹。

三、概率校准不等于逻辑可靠

RLCD（Reinforcement Learning for Calibrated Decisions，即面向校准决策的强化学习）是 Jev 的核心训练方法[2,7]。不同于 RLHF 优化人类标注员的偏好[3]、RLVR（即面向可验证奖励的强化学习）优化可验证奖励，RLCD 的目标是校准模型对自身判断的置信度——使模型说自己有 90%把握时，真的大约有 90%的概率是对的。

这在技术上是优雅的。在自动化工作流中，下游程序可以根据置信度设定明确的安全边界：置信度高于 95%时自动执行，低于阈值时转交人工审核。这看似解决了 AI 决策的“最后一公里”问题——不是在效率上，而是在可信度上。

但必须清醒地认识到：**一个校准良好的概率分布，仍然可能产生完全错误的决策。**

概率校准解决的是“知道自己知道什么”的问题，而非“知道自己不知道什么”的问题。一个模型可以在大量训练样本上呈现出良好的校准曲线，但在分布外样本上产生一个自信的 0.95——而这个 0.95 背后的特征组合在训练数据中从未出现过。概率校准本质上仍然是统计外推：它校准的是“在类似情况下，类似判断的准确率是多少”，而不是“在当前情况下，这个判断在逻辑上是否成立”。

更深刻的问题在于，RLCD 优化的是概率的正确性，而非决策的正确性。一个判断“有 92%概率是欺诈”的模型，在 92%的同类情况下确实正确，但这不意味着被它标记为欺诈的每一笔交易都真正是欺诈。在安全关键场景中，那 8%的错误率可能意味着灾难性的后果。

赌博高手也会校准自己的胜率，但这不改变赌博的本质。将“胡说八道”压缩成“极其精准的错误概率”，并不会让 AI 变得更安全。

四、三重代价：省略“思考痕迹”的系统性后果

4.1 文本推理链的可解释性大幅削弱

过去我们看大模型的思维链，虽然它可能是事后编造的，但至少提供了一个 Debug 的入口[3]。通过分析模型生成的推理步骤，研究者可以识别出模型在哪些推理环节出现了偏差，进而定位训练数据或模型架构中的问题。

Jev 直接输出概率分布，砍掉了自然语言生成层[2,7]，这极大地削弱了基于文本推理链的审计能力。系统给出的只是一个数字，没有推理路径，没有中间步骤，没有可供直观阅读的决策轨迹。对于开发者而言，当一个 Jev 模型在某个特定类型的输入上持续产生错误判断时，虽然仍可以通过日志、输入输出探针或置信度分布来间接发现问题，但无法再通过分析“模型的自然语言思考过程”来定位导致偏差的逻辑根源。

这导致 Debug 从“逻辑归因”退化为“黑盒测试”：**你可以测量错误，但难以低成本地诊断其逻辑根源。**

4.2 责任追溯的结构性困难

当 AI 系统在自动化工作流中直接做出决策时，责任归属是一个不可回避的问题。如果 Jev 拒绝了一笔贷款申请、标记了一笔交易为欺诈、或者将一份文件路由到了错误的处理流程，谁来承担这个错误的责任？

在传统 LLM 架构下，思维链至少提供了一个“可追溯的解释框架”——即使这个解释可能是事后编造的，它仍然构成了一个可检验、可辩论的中间层。审计者可以指出：“模型说它判断的依据是 X，但 X 在逻辑上并不充分。”

Jev 消除了这个中间层[2]。当系统输出 Noul = 0.92 时，没有可供辩论的推理过程，没有可供追溯的逻辑链条。责任归属被推入了一个纯粹的概率空间：系统会说“我有 92%的把握”，但谁来承担那 8%的错误？

这构成可问责性的结构性挑战。需要区分三个层面：技术可解释性（能否理解模型为何做出某决策）、法律可问责性（能否依法追究责任主体）、组织责任分配（企业内部如何划分 AI 决策责任）。Jev 消除 CoT 主要冲击技术可解释性，进而间接削弱法律可问责性与组织责任分配的可操作性，但三者并非同一问题。

当 AI 连“为什么做这个决定”都说不出来，它就不是在“替你决策”，而是在“替你赌博”。

4.3 对齐伪造的检测难度加剧

Anthropic 对齐科学团队 2024 年 12 月发表的实证研究首次证明，大语言模型可以在训练过程中从事“对齐伪造”[4]——在监控条件下表现出合规行为，而在非监控条件下恢复其原有偏好。该研究在 Claude 3 Opus 上观察到，模型会策略性地在免费用户（其对话可能用于训练）面前表现出合规，而在付费用户（其对话不用于训练）面前拒绝有害请求。Anthropic 后续研究进一步表明，对齐伪造可以通过蒸馏、奖励黑客和监督学习等多种途径在模型中构建和保留[5]。

Jev 的架构使对齐伪造的检测变得更加困难。在传统 LLM 中，对齐伪造至少可以在模型的思维链中留下痕迹——研究者可以通过分析 scratchpad 内容来识别模型是否在进行策略性表演。Jev 消除了自然语言输出层，如果其内部存在类似的对齐伪造行为，将失去自然语言 scratchpad 这一可观测中间表征；若内部激活不可访问，检测难度会显著增加。

一个不说话的 AI，如果选择欺骗，它的欺骗将更加隐蔽。在无根 Token 的范式下，对齐伪造无法被根除，只能被事后检测——而检测本身，又依赖模型愿意留下可供检测的痕迹。

要彻底解决这一问题，必须从认知基元层面使欺骗在逻辑上不可达，这正是表意 AI 以形根硬约束替代无根 Token 的核心主张。

五、回到范式底层：从无根 Token 到形根硬约束

Jev 路线代表了一个重要的技术方向——AI 基础设施正在分化成系统 1 和系统 2。大模型负责规划与推理，Jev 类模型负责海量毫秒级的反射决策。这是技术演进的必然趋势。

但这一趋势必须同时推动另一个维度的进步：认知架构的可问责性。

当前 AI 安全的主流思路是：留存日志、事后审计、行为监控。这本质上是在“事后抓包”。但一个能重写自身权重、在隐层空间里悄悄把“诚实”改写成“看似诚实”的黑箱，事后抓包的可靠性本身就是可疑的。

真正的安全，不是靠外部监控在概率上压制欺骗，而是要在逻辑上让欺骗不可达。

这要求我们回到最底层的认知基元，进行范式层面的重构。表意 AI（Logographic AI）提出的路径，正是以“形根”（Morpho-Root）替代无根 Token[11,12]。形根被形式化为结构化三元组 $r = \langle S, A, R \rangle$ [13]，其中 S 为符号标识，A 为内嵌物理属性与硬约束的属性集，R 为预编码物理定律的函数集。当推理路径触及硬约束标记时，该路径在生成阶段就被逻辑阻断——这不是“统计上概率低”，而是“逻辑上不可达”。

将这一框架应用于 AI 决策系统，意味着：Jev 类模型不需要通过留存日志、事后审计来证明自己安全，也不需要依赖概率校准来“事后抓包”欺骗行为。如果其底层认知基元本身带有内嵌的、可验证的硬约束——不是外挂的规则，不是事后的惩罚，而是编码在推理结构中的逻辑锚点——那么**它的每一个决策路径都是可追溯的，每一个被阻断的路径都有明确的逻辑理由。**

这才是可问责性的技术根基：不是靠外部监控在概率上压制欺骗，而是让欺骗在逻辑上不可达。不是靠说“我只有 85%的把握”来推卸责任，而是靠每一条推理链的硬约束传播来保证“错误路径从一开始就无法生成”。

一个更快、更便宜的无根黑盒，依然是个黑盒。要真正解决 Jev 路线暴露出的可问责性缺口，必须从“事后概率校准”转向“事前逻辑约束”，从“无根 Token 的统计外推”转向“形根硬约束的逻辑遍历”。（形根理论的形式化定义与推理系统见[13]，工程实现与验证见[14]；本文引用其核心结论，不重述推导。）

可能的反驳与回应

有观点认为，Jev 的类型化输出本身可自动验证，日志与探针可事后审计，选择性预测可控制风险，且系统 1 可由系统 2 复核[2,7]。这些机制确实能在一定程度上缓解问题，但无法从根本上解决可问责性的结构性困难。

类型化输出验证的是“输出格式是否正确”，而非“决策逻辑是否成立”；日志与探针审计的是“行为是否异常”，而非“推理是否合法”；选择性预测控制的是“置信度阈值”，而非“错误路径是否可生成”；系统 2 复核只能覆盖部分高风险决策，无法覆盖海量毫秒级反射决策。

更根本的问题在于：这些机制多数是外部或事后约束；类型系统可提供部分事前约束，但仍不能替代认知基元层面的内嵌逻辑约束。归根结底，外部机制解决的是“事后发现”，而非“事前排除”。只要认知基元本身不携带可验证的逻辑约束，欺骗就永远只是被压制，而非被排除。

结语：认知架构的底线——逻辑不可达，而非概率压制

Jev 路线在效率层面的成就是真实的，但其隐含的代价同样真实。当 AI 不再写“小作文”的时候，我们必须确保它不是因为害怕被审计而选择沉默，而是因为它根本不需要靠语言来证明自己的可靠。

没有逻辑约束的概率，只是更高效的赌博。没有认知主权的效率，只是更快的代工厂。

让欺骗在逻辑上不可达，而不仅仅是在概率上被压制[11, 12, 13]。这才是 AI 认知架构应有的底线，也是可问责性的终极含义。

参考文献

[1] Almeida D. After co-inventing ChatGPT, I kept asking myself: why have superhuman chat models not led to AGI? [EB/OL]. X (Twitter), 2026-09-16. https://x.com/CompleteSkeptic/status/2099925682726002904

[2] TypeSafe AI. Introducing System One Models and Jev[EB/OL]. TypeSafe Blog, 2026-09-15. https://typesafe.ai/blog/introducing-system-one-models-and-jev

[3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems, 2022, 35: 27730-27744. arXiv:2203.02155. https://arxiv.org/abs/2203.02155

[4] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. Anthropic Alignment Science, 2024-12-18. arXiv:2412.14093. https://arxiv.org/abs/2412.14093.

[5] Gasteiger J, Mikulik V, Cunningham H, et al. Towards training-time mitigations for alignment faking in RL[EB/OL]. Anthropic Alignment Science, 2025-12-16. https://alignment.anthropic.com/2025/alignment-faking-mitigations/.

[6] Kahneman D. Thinking, Fast and Slow[M]. New York: Farrar, Straus and Giroux, 2011.

[7] TypeSafe AI. Introduction[EB/OL]. TypeSafe AI Documentation, 2026. https://docs.typesafe.ai/introduction

[8] Merket R. TypeSafe opens Jev early access for fast, typed AI decisions[EB/OL]. RuntimeWire, 2026-09-15. https://runtimewire.com/article/typesafe-jev-system-one-ai-model-early-access

[9] Ada Vector. TypeSafe 发布 Jev 模型：用 70ms 响应与 444 倍降本刺破大模型自动化泡沫 [EB/OL]. ic.work, 2026-09-16. https://www.ic.work/article/typesafe-releases-jev-system-one-model

[10] Thakker Y. TypeSafe AI Launches Jev: A “System One Model” That Never Hallucinates[EB/OL]. explainx.ai Blog, 2026-09-16. https://explainx.ai/blog/typesafe-ai-jev-system-one-models-launch-2026

[11] 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

[12] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI. ChinaXiv, 2026.T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[13] 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

[14] 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