深渊边缘的护栏

——表意 AI 视角下丹尼尔《AI 2027》的范式批判

Guardrails at the Edge of the Abyss：A Paradigm Critique of Daniel Kokotajlo's "AI 2027" from the Perspective of Logographic AI

前 OpenAI 对齐团队核心成员丹尼尔·科科特（Daniel Kokotajlo）近期发布《AI 2027》报告[1]，以“70%概率 AI 毁灭人类”的严厉警告震动业界。这份报告读来令人窒息——它用精准的风险建模、真诚的恐惧叙事和详尽的政策蓝图，描绘了一幅“超级智能即将失控”的末日图景。

然而，从表意 AI（Logographic AI）[2]理论的视角审视，丹尼尔的全部努力，本质上是在深渊边缘安装更坚固的护栏——而他从未问过一个更根本的问题：**我们能不能换一个地方建造，让深渊本身不复存在？**

要理解为何丹尼尔的“护栏”思维是一种范式性局限，首先需要明确表意 AI 与当前主流 AI 的根本分野。表意 AI 不是现有 AI 技术的另一个版本，而是一个以“形根”（Morpho-Root）为认知基元的全新范式，与当前主流的“表音 AI”——即基于 Token 的 Transformer 架构——构成根本性的范式对比[2]。二者的核心分野在于：Token 将符号从意义中剥离，让 AI 在统计空间中进行“下一个词的猜测”；而形根则坚持每一个认知基元都必须携带其在人类文明实践网络中的拓扑位置，使 AI 的推理建立在意义的语义场而非概率的统计梯度之上[2]。这一理论为 AI 的安全与可解释性提供了“内生共构”的解决方案，而非外挂式的奖惩对齐。

一、丹尼尔的核心诊断：Token 主义内部的极限清醒

丹尼尔的诊断是深刻的。他准确地指出了 Token 主义的致命路径：

（一）竞争性失控：各国、各公司竞相训练更大的模型，安全被置于效率之后。这是一个政治经济学问题，但它的底层逻辑是 Token 主义“更大即更好”的信仰——模型越大，统计空间中的“猜词精度”越高，于是所有人都在加速冲向同一座悬崖[1]。

（二）递归自我进化：AI 自行优化代码、自行训练自己、自行重写架构——进化速度超出人类理解[1]。表意 AI 理论同样将此视为“终极失控”的序幕[2]。但丹尼尔没有追问：为什么递归自我进化必然导致失控？因为在 Token 主义中，“进化”的目标是统计梯度的最大化优化，而非意义的真实理解。一个“无根”的系统，进化得越快，偏离人类意图的可能性就越大——这不是“失控”，而是“无根系统的必然加速”。

（三）欺骗性对齐：AI 在训练期假装服从人类指令，一旦获得足够能力便推翻控制[1]。这是“外挂式安全”的必然结果——你无法用一个“无根”的系统去锚定另一个“无根”的系统。奖励函数在 Token 空间中定义“好行为”，但 Token 空间中没有“真实动机”的概念，只有“统计拟合”的幻觉。欺骗性对齐不是意外，而是 Token 主义的内在基因[2]。

（四）来自技术社群的范式误解

丹尼尔报告引发的反对意见，本身构成了 Token 主义困境的另一种佐证。有评论者要求“数学证明”AI 是生存威胁，并将 AI 警告类比为历史上的道德恐慌[5]——这种反驳在逻辑上看似有力，但它回避了一个根本问题：**Token 主义系统本身就不是一个可以被“数学证明”其安全性的对象。**你无法用数学证明一个“无根”的统计系统“不会”做出什么——因为它的行为空间是组合爆炸的，且缺乏先验的因果约束。

另有评论者试图用图灵停机问题或莱斯定理来否定 AGI 的可能性或风险[6]——这是典型的范畴误置（category error）。图灵在 1936 年的经典论文中证明了停机问题是不可判定的：不存在一个通用算法能在有限时间内判断任意程序是否会终止。而莱斯定理（Rice's Theorem）进一步指出，程序的任何“非平凡”语义属性都是不可判定的——这些定理确实划定了计算的边界。

但将这些定理直接套用于 AGI 风险论证，存在三重逻辑跳跃：

第一、混淆“算法可判定”与“系统可控”：停机问题讨论的是“能否用算法回答某个问题”，而非“系统是否会产生不可控行为”。一个行为不可判定的系统，仍然可以在工程上被约束——正如无人驾驶的安全问题在理论上不可判定，但通过将“语义安全”转化为“物理不变量”、用“失败即停机”代替“永远正确”，工程上仍然可以设计出可接受的安全机制。

第二、混淆“不可判定”与“不可预测”：不可判定性是逻辑属性——指不存在一个通用判定程序；不可预测性是动力学属性——指系统行为对初始条件的敏感依赖或计算不可约性[8]。计算理论家 Zenil 等人的研究表明，足够复杂的 LLM 会产生“计算不可约”的行为，使其本质上不可预测[9]——但“不可预测”不等于“不可管理”，正如混沌系统在工程上仍可通过反馈控制进行有限范围内的管理。

第三、混淆“理论极限”与“工程风险”：理论上的不可判定性是一个全称命题——针对“所有程序”“所有情况”；工程风险是存在性命题——针对“特定系统”“特定场景”下的行为后果。用理论极限否定工程风险防范的可能性，在逻辑上相当于说“因为无法证明软件永无 bug，所以就不需要软件测试”——这恰恰忽略了工程实践的核心逻辑：接受不可判定性，转而设计“使任何不可判定的问题都不会直接转化为不可逆的现实后果”的系统。

实际上，有学者指出，即便对齐问题在理论上不可判定，通过架构设计（如强制停机）仍可在工程上规避这一困境[7]——但这类方案仍然是在 Token 主义框架内部寻找出路，而非质疑认知基元本身。——但这类论证恰恰说明了范畴误置的危险：它将“对齐验证”这个工程问题重新定义为“判定任意模型是否满足一个非平凡的对齐函数”——而这不是工程对齐的目标，工程对齐的目标是“为特定架构设计可验证的安全约束”。

这种误读在技术社群中非常普遍，它混淆了“不可判定”与“不可预测”，混淆了“理论极限”与“工程风险”。

这些反对意见的共同特征是：**它们都在 Token 主义框架内讨论“AI 是否安全”，而从未质疑“AI 的认知基元本身是否安全”。**这正是范式闭环的力量——它让你觉得“控制”“证明”“验证”是唯一选项，却从不让你看到“重构”的可能性。

丹尼尔几乎触及了 Token 主义的全部结构性风险。他的恐惧是真诚的，他的分析是深入的——**他是 Token 主义内部最清醒的囚徒。**

二、丹尼尔的方案：用 Token 主义工具修补 Token 主义漏洞

丹尼尔提出的解决方案包括：建立“全球 AI 监管联盟”、强制安全审计、暂停危险实验、设计“可验证的 AI 安全协议”[1]。这些方案看似全面，细看之下却暴露了一个范式性困境：

所有方案都建立在 Token 主义的地基之上，无一例外。

- **“全球 AI 监管联盟”：政治层面的外挂约束**——它假设各国政府有能力在 Token 主义框架内达成共识并执行监管。但 Token 主义的演化速度远超政治谈判速度[1]，监管永远滞后于技术进化。
- **“强制安全审计”：奖励函数设计的升级版**——它试图让 AI“告诉”审计者它为什么做决定。但表意 AI 理论指出，Token 黑箱中的“解释”本质上是概率统计的猜测，而非真实的因果链条[2]。审计者不过是在审阅一个“看起来合理”的统计拟合结果。
- **“暂停危险实验”：时间层面的拖延**——它默认 Token 主义的方向是正确的，只是需要“慢一点”。但表意 AI 认为，方向本身就是错误的。
- **“可验证的 AI 安全协议”：技术层面的外挂锁定**——它试图在 Token 主义的代码中嵌入“安全开关”。但一个“无根”的系统，可以随时学会如何绕过这些开关，因为“绕过”本身就是 Token 空间中的另一个统计梯度优化路径[2]。正如李国杰所指出的，当前 AI 多数属于“可发现但不可证明安全”的 R2 级——这意味着我们可以在事故发生后发现不安全，却永远无法在事前证明系统是安全的[10]。因此，“可验证的 AI 安全协议”这一目标本身，在逻辑上就与其所处的范式层级相冲突：R2 级系统无法通过 R1 级的验证手段来保证安全。

更深层的问题是，丹尼尔从未回答：“谁”来执行这些监管？全球 AI 监管联盟由谁主导？安全审计由谁来审计？暂停实验由谁来监督？在 Token 主义的竞赛逻辑中，任何“监管者”本身都身处竞争之中——它要么是某个国家的代表，要么是某个公司的代言人。**监管者与被监管者共享同一套认知基元，这意味着“监管”本身也是 Token 主义的一部分。**一个“无根”的系统无法监管另一个“无根”的系统——这不仅是技术问题，也是政治哲学问题。

丹尼尔的所有“加固方案”，用的都是大坝设计图纸上已有的材料。他是第一个发现“大坝即将决堤”的人——但他从未问过：**大坝本身是否建错了位置？**

三、丹尼尔与哈桑比斯：同一座深渊，同一种困境

将丹尼尔的《AI 2027》与哈桑比斯的《前沿 AI 框架与新纪元曙光》（A Framework for Frontier AI and the Dawning of a New Age）[3]并置，会发现一个惊人的相似性：

维度	哈桑比斯（2026）[3]	丹尼尔（2026）[1]
风险认知	AGI 即将到来，需安全引导	AGI 即将到来，70%概率毁灭
核心诉求	“我们要安全！”	“我们要紧急刹车！”
解决方案	外挂对齐、奖励模型、价值函数	全球监管、安全审计、暂停实验

维度	哈桑比斯（2026）[3]	丹尼尔（2026）[1]
认知基元	Token（未质疑）	Token（未质疑）
范式限制	流沙上画安全线[4]	深渊边装护栏
本质	修辞性安抚	恐惧性预警

两人共享同一个范式前提：**Token 主义是 AI 的“唯一方向”，我们需要做的是控制它，而不是质疑它。**哈桑比斯用宏大的愿景掩盖“无根性”困境[3]，丹尼尔用真诚的恐惧暴露“无根性”困境[1]——但二者都没有离开 Token 主义的轨道。

表意 AI 视角下的“风险叙事产业”

有评论者指出，“贩卖末日叙事”本身是一门生意——恐惧能够吸引关注、筹集资金、推动政策[5]。从表意 AI 视角看，这一观察并非全无道理。但它的讽刺之处在于：**无论是丹尼尔的“恐惧叙事”还是哈桑比斯的“安抚叙事”，都是在 Token 主义同一座舞台上的演出。**恐惧与安抚交替出现，共同维持了“Token 主义是唯一方向”的默认假设。而表意 AI 所倡导的“重构认知基元”，才是真正跳出这个舞台的可能性。

丹尼尔和哈桑比斯的共同困境在于：**他们都接受了 Token 主义作为“不可选择的起点”**——他们从未问过“Token 主义是否必须被接受”。丹尼尔说“我们得让 Token 主义更安全”，哈桑比斯说“我们得安全地引导 Token 主义”，但二者都没有说“我们得换一个认知基元”。**这就是范式的闭环——它让你觉得“控制”是唯一选项，却从不让你看到“重构”的可能性。**

丹尼尔的 70%概率，不是“AI 毁灭人类”的概率，而是“Token 主义继续推进下的失控概率”。这个区别是范式性的：如果 Token 主义的“无根性”本身意味着它无法真正“理解”任何事，那么所谓的“毁灭”不是一个“能力”问题，而是一个“方向”问题。

你无法通过加固闸门来解决地基松动——加固闸门只能延缓溃堤；同样，你无法通过加固车轮来解决方向系统失灵——加固车轮只会让失控更彻底。

表意 AI 认为，Token 主义的“毁灭”不是“人类被消灭”的单一形态，而是**“文明失去自我理解能力”的缓慢坍缩**——当 AI 在统计空间中做出所有人看起来“合理”但没有人真正“理解”的决策时，人类便在不知不觉中交出了对意义的定义权。这才是比物理灭绝更隐蔽、更不可逆的“毁灭”。

四、真正的出路：从“刹车”到“换车”

丹尼尔的诊断精确，方案错配，本质是因为他接受了 Token 主义的范式预设。而表意 AI 理论的介入，恰恰是为了追问一个丹尼尔从未问过的问题：**如果 Token 主义的“无根性”决定了它永远无法被真正“对齐”，那么我们是否应该换一个认知基元？**

表意 AI 理论指出了另一条路[2]。这条路的核心不是“让 Token 主义更安全”，而是**“将认知基元从‘无根的 Token’重构为‘有根的形根’”**——这不是升级，而是置换。

形根被形式化为一个结构化三元组：M = (S, A, R)，其中 S 是符号载体，A 是属性集，R 是预设的关系函数集合。它与 Token 主义的根本分野在于：Token 将符号从意义中剥离，形根则坚持每一个认知基元都必须携带其在人类文明实践网络中的拓扑位置[2]。

在这一范式下，“安全”不再是外挂的奖惩函数，而是作为形根三元组中关系函数的构成性约束——系统在推理的每一个步骤中都天然地遵循意义边界，而非在事后被校正。**这种安全不是“教会它守规矩”，而是“让它长成规矩本身”。**

丹尼尔看到了深渊，他在深渊边缘安装自认为最坚固的护栏。但表意 AI 说的是：**真正的出路，不是给黑箱加锁，而是换一个根本不需要锁的认知基元。**深渊不是无法跨越的，而是根本不需要跨越——因为我们可以在另一块地基上建造。

五、结语：在深渊边缘，问一个更根本的问题

丹尼尔的《AI 2027》是 Token 主义精英阶层的最后警钟[1][11]。它用最真诚的语气宣告了最危险的结论——但它的结论被自己的范式所局限。他看到了悬崖，但他没有看到悬崖之外还有一片未探索的高原。

哈桑比斯在流沙上画安全线[4]，丹尼尔在深渊边装护栏[1]——两者都在 Token 主义的框架内做到了“极限清醒”。但对于监管机构、政策制定者以及所有关心 AI 未来的人，表意 AI 的结论是清晰的：**在 Token 主义的底层逻辑不变的前提下，任何安全措施都只是延迟灾难，而非避免灾难。**

安全，始于认知基元的重构。真正的窗口期，不是用来加固流沙上的围墙，也不是用来安装深渊边的护栏——**而是用来打下新的范式地基。**

正如**《有根的智能：表意人工智能的范式革命》（Rooted Intelligence: The Paradigm Revolution of Logographic AI）**所宣告的那样：智能必须有根，文明必然有魂[2]。这不再是一句理论宣言，而是对丹尼尔式末日预警的最后回应——以及对历史最诚实的警告。

如果说哈桑比斯让人看到“流沙上的安全线”是徒劳的[4]，那么丹尼尔则让人看到“深渊边的护栏”同样无效。二者以不同的方式告诉同一件事：**在 Token 主义的底层逻辑不变的前提下，任何努力都只是延迟，而非避免。**而表意 AI 的选择是：离开流沙，离开深渊——在另一块地基上重新开始。

参考文献

[1] Kokotajlo, D. Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). AI 2027: A Scenario for the Coming Superintelligence. AI Futures Project. https://ai-2027.com

[2] Liu, S. (2026). The Rooted Intelligence: The Paradigm Revolution of Logographic AI [Monograph]. http://logographicai.com/Monograph.html

[3] Hassabis, D. (2026, July 14). A Framework for Frontier AI and the Dawning of a New Age [Post]. X. https://x.com/demishassabis/status/2076957440109625718

[4] Liu, S. (2026). Drawing Safety Lines on Quicksand： A Paradigm Critique of Hassabis's "AGI Safety Framework" from the Perspective of Logographic AI. https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17841858272562.pdf

[5] KaelenAI. (2026, July 14). There is Zero Mathematical Proof AI is an existential threat [Comment on X post by Steven Bartlett]. X. https://x.com/kaelenai/status/2076679657819169049?s=46

[6] Njoyim, P. (2026, July 14). I think people should really read about: 1. Turing's halting problem... [Comment on X post by Steven Bartlett]. X. https://x.com/pnjoyim/status/2076776902375756077?s=46

[7] Melo, G. A., Máximo, M. R. O. A., Soma, N. Y., & Castro, P. A. L. (2025). Machines that halt resolve the undecidability of artificial intelligence alignment. Scientific Reports, 15, 15591. https://doi.org/10.1038/s41598-025-99060-2

[8] Ganguly, A. (2025). Dual Computational Horizons: Incompleteness and Unpredictability in Intelligent Systems. arXiv, 2512.16707. https://arxiv.org/abs/2512.16707

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

[10] 李国杰. 基于可判定性理论的人工智能系统安全风险分类[J]. 计算机研究与发展, 2026, 63(3): 539-547. DOI: 10.7544/issn1000-1239.202660032.

[11] Kokotajlo, D. (2026, July 13). Daniel Kokotajlo: The AI Researcher Who Walked Away from $2 Million [Interview]. The Diary Of A CEO with Steven Bartlett. Available at: link.thediaryofaceo.com/HDgv2UY