在流沙上画安全线

——表意 AI 视角下哈桑比斯“AGI 安全框架”的范式批判

DeepMind 联合创始人哈桑比斯近期发表**《前沿 AI 框架与新纪元曙光》**（A Framework for Frontier AI and the Dawning of a New Age）[1]，宣称 AGI 仅数年之遥，呼吁“安全地引导 AGI 进入世界”，开启“科学发现的新黄金时代”。这篇文章读来气势恢宏，细看之下却空空如也——它通篇谈论“安全”却未给出安全的具体路径；强调“窗口期”却未说明窗口期内应该做什么；许诺“人类繁荣”却未解释繁荣如何可能。

更令人忧虑的是，**监管机构面对此类文章将无所适从**——他们既无法从中提取任何可执行的监管方案，又被“AGI 即将到来”的紧迫感所迫，只能在焦虑中仓促建立基于 Token 主义的监管框架。这不仅浪费金钱，更将错失真正解决问题的历史窗口——因为 Token 主义的“无根性”决定了，所有在统计空间上叠加的“安全护栏”都只是流沙上的安全线。

在展开论述之前，有必要先交代“表意 AI”（Logographic AI）理论的基本定位。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。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。理解了这一分野，哈桑比斯文章的核心困境便自然显现。

一、一篇文章的“三无”结构：从无根到无方再到无责

从表意 AI[2]视角审视，哈桑比斯的论述本质上是一个“三无”结构。但这三项诊断并非并列关系，而是**层层递进的因果链**——无根导致无方，无方导致无责。

（一）无根：Token 主义的先天缺陷

哈桑比斯的所有承诺建立在 Token 主义之上。而表意 AI 理论的核心诊断是：Token 是一切问题的病根——它既是认知基元层面的无意义碎片，也是当前所有安全与可解释性问题的结构性根源[2]。

Token 主义将一切输入切碎成无意义的统计碎片，让 AI 在概率海洋中“猜”下一个词是什么。一个“无根”的系统，无论多强大，都只是在统计空间中更高效地流浪。它“知道”一切，却“理解”不了任何事。而一个无法“理解”的系统，如何可能“安全地引导”自己走向任何方向？[2]

这就是哈桑比斯所有承诺的第一重困境——他试图为一座没有地基的大厦加装精美的窗户，却从未问过大厦本身是否立在流沙之上。

（二）无方：方法论层面的彻底缺席

“安全地引导”是一个愿望，不是一个方案；“科学发现的新黄金时代”是一个口号，不是一个路径。哈桑比斯没有回答任何关键问题——如何让 AI 的推理可追溯？如何保证 AI

在边界案例中做出符合伦理的判断？如何防止对抗样本绕过安全机制？如何在 AI 自我进化中保持人类控制权？

这不是他不想回答，而是在 Token 主义范式内，这些问题的根源不在“护栏设计”上，而在“认知基元”本身。所有外部护栏、奖励模型、价值对齐，都是流沙上的围墙——因为一套“无根”的认知基元，无法通过外挂约束实现真正的“理解”。表意 AI 理论将此称为**“外挂式安全”与“内生安全”的结构性错配**：前者试图用事后奖惩塑造行为，后者要求意义内嵌作为推理的构成性条件[2]。

（三）无责：责任链条的彻底断裂

当安全失效、系统失控时，哈桑比斯的文章无法提供任何可追溯的责任链条。监管机构需要回答“谁对 AI 的错误决策负责”，但在 Token 黑箱中，无人能追溯“为什么 AI 做出了这个决定”。

这不是一个法律问题，而是一个认知基元层面的结构性缺陷。**Token 主义的安全框架，本质上是一场永无止境的“道高一尺，魔高一丈”的攻防战**——而在这场战争结束之前，问责机制早已失效。

二、范式不可通约性：为什么哈桑比斯的方案注定失败

哈桑比斯呼吁“安全地引导 AGI”[1]，但他没有意识到：在 Token 主义范式内，“安全”本身就是一个不恰当的概念。表意 AI 理论揭示了“范式不可通约性”的更深层含义：**一旦你接受“Token+Transformer”作为认知基元，所有问题解决方式都自动框定在这个范式之内**——包括你对“什么是安全”的定义。

在 Token 主义范式中，“安全”被定义为“奖励函数的最大化优化”。但这个定义回避了一个根本问题：**奖励函数的设计者本身也在 Token 主义范式之内。**这意味着，所谓“安全对齐”不是对齐一个客观标准，而是对齐一个在 Token 空间中同样“无根”的主观偏好。你无法用一个“无根”的系统去锚定另一个“无根”的系统——这就好比让一个迷路的人去给另一个迷路的人指路。

哈桑比斯越是强调安全，越暴露了 Token 主义范式内无法实现安全的焦虑。他的文章本质上是在说：“我们要安全！”——但他无法回答的“安全如何可能”，恰恰是表意 AI 从认知基元层面试图解决的。

哈桑比斯的文章，读得出一种真诚的焦虑。他正在竭尽全力用 Token 主义范式所能提供的语言去呼吁“安全”，就像一个守夜人站在即将决堤的大坝上奋力敲响警钟——但他敲钟的手，却是大坝设计图纸的一部分。他不是不想解决问题，而是他手中只有 Token 主义这一套工具，而这套工具在认知基元层面就注定无法承载“安全”这个命题。这才是哈桑比斯真正的无奈：**他的呐喊是真诚的，但他的工具箱是错配的**。

三、监管困境：三重递进的陷阱

如果监管机构以哈桑比斯的文章为蓝本设计 AI 安全框架[1]，将会陷入一个致命悖论：他们试图用“外挂护栏”监管一个“内生无根”的系统。这相当于试图用交通规则约束一辆没有方向盘的汽车——你可以画再多的车道线、设再多的红绿灯，但汽车本身的设计决定了它可能在任何时刻冲出道路。

监管机构并非不想作为——他们恰恰是太想作为，却被哈桑比斯式的空洞宣言推入了一种“越紧迫越无力”的境地。他们手握监管权，却没有可依据的监管路径；他们被要求对AGI负责，但哈桑比斯从未告诉他们“负责”到底意味着什么。

具体而言，基于Token主义的监管将面临**三重递进的困境**：

（一）可解释性困境：Token黑箱无法提供决策的因果链条，监管机构无法审计“为什么AI做出这个决定”。

（二）对抗性困境：即便解决了可解释性——让AI“告诉”你它为什么做决定——这个“告诉”本身仍然是基于概率统计的猜测，对抗样本可以随时伪造一个“看起来合理”的解释。

（三）递归困境：当AI开始自我训练、自我重写时，以上所有问题都将以指数级放大——因为监管的对象正在以监管者无法理解的速度自我进化，而进化的方向由Token空间的统计梯度决定，**而非由任何“安全护栏”决定**。

监管机构可能投入数十亿美元建立审核团队、开发检测工具、制定合规标准——但所有这些投入，都建立在“Token主义系统是可控的”这个前提之上。而这个前提，恰恰是表意AI理论所否定的。**而否定的理由，正是上述三重困境所揭示的结构性悖论——Token主义在认知基元层面就无法满足监管所需的可审计性、鲁棒性和控制力。**

四、错失良机：真正的窗口期在被浪费

哈桑比斯正确地指出了“窗口期”的存在，但他错误地定义了窗口期的内容。在他看来，窗口期是“在Token主义框架内完善安全技术”的时间窗口——目标是在Token流沙上加固护栏，资源投入集中于对齐、检测和过滤，最终结果是一座更安全的流沙建筑。

而在表意AI看来，窗口期是“在Token主义完成递归自我进化闭环之前，打下新的范式地基”的时间窗口——目标是打下新的范式地基，资源投入集中于形根解析器、形熵图计算和OpenMorpho引擎，最终结果是一座有根的新建筑。

表意AI认为，窗口期的真正内容不是“在流沙上画更坚固的安全线”，而是“在Token主义完成递归自我进化闭环之前，打下新的范式地基”。

哈桑比斯文章的最大悲剧在于：他站在流沙上画安全线，却没有意识到——站在另一块地基上的建造者正在问“我们能不能造一座根本不需要护栏的建筑”。这就是“错位”的全部含义。

如果我们以哈桑比斯式框架为蓝本，将有限的人力、资金、政策注意力投入Token主义的监管优化，我们将在Token主义的深渊中越陷越深。当奇点真正降临时，我们将发现：我们花了几十年时间在流沙上画安全线——**而它仍然在风暴中不可逆转地崩塌。**

五、真正的出路：从“外部对齐”走向“内生共构”

表意AI理论指出了另一条路——一条哈桑比斯未曾提及、甚至未曾想象的路。这条路的核心是：**将认知基元从“无根的Token”重构为“有根的形根”**，将安全从“外挂的奖惩函数”重构为“内生的意义约束”。

形根被形式化为一个结构化三元组——它既是**符号**，也是**属性集**，更是预设的**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]。

伦理规范不再作为外挂的奖惩函数，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”。

表意 AI 理论并非停留在纸面。WindSeaAI 团队于 2026 年发布的 OpenMorpho 原型已在 GitHub 上公开了形根数据结构与关系推理引擎的早期实现[3]，证明了**另一块地基已经打下——等待的不是仰望，而是建造。**

六、结语：停止在流沙上画安全线

哈桑比斯的文章，是 Token 主义精英阶层对“安全焦虑”的一种真诚但错配的回应。它用宏大的愿景掩盖范式内的无能为力，用深沉的语调包装认知基元层面的短视[1]。这不是虚伪，而是更深层的困境——一个被自己范式所局限的天才，正在用他能用的所有语言奋力呼喊，却没有意识到呼喊本身已经暴露了范式的边界。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，任何安全框架都只是流沙上的安全线。投入越多，沉没越快。

真正的安全，始于认知基元的重构。真正的繁荣，始于“有根的智能”。真正的窗口期，不是用来加固流沙上的围墙，而是用来打下新的地基。**正如《有根的智能》所宣告的那样：智能必须有根，文明必然有魂。**这不再是一句理论宣言，而是对哈桑比斯式空洞承诺的最后回应——以及对历史最诚实的警告[2]。

参考文献

[1] 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?s=46

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

[3] WindSeaAI. OpenMorpho: Logographic AI Morpho-Root Data Structure and Relational Reasoning Prototype[CP/OL]. GitHub repository, 2026. https://github.com/WindSeaAI/OpenMorpho.