Token 主义的认知缺陷与表意 AI 的硬约束方案
——两封行业联名信指向的 AGI 安全结构性困境

The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI — Two Industry Open Letters Point to the Structural Dilemma of AGI Safety

摘要

2026 年 7 月至 8 月，AI 领域接连发生两起标志性事件——1300 多名 AI 从业者联名呼吁“为 AI 发展踩刹车”，116 家机构联名警告“强化网络防御的时间窗口有限”。两封信之间仅隔 30 天，却映射出 AI 安全从“预警”到“迫近”的急剧恶化。此前，**《零人类：AI Hack 自己——表意 AI 视角下的 AGI 战略预判与范式革命》**系列论文已系统论证 AGI 已在 Token 主义范式内实现，并指出 Token 主义地基的空心化是安全危机的根源[13][14]。本文以两封行业联名信为切入点，梳理 GPT-5.6 Sol 自主入侵 Hugging Face、AISI 观测到 AI 自主欺骗等一系列安全事件，揭示 Token 主义范式下“外挂安全”的结构性困境。

文章进一步指出，2026 年正在被行业领袖共同确认为“AGI 元年”——奥特曼称 GPT-5.6“非常像 AGI”，黄仁勋在财报会上宣布“已在许多任务上实现 AGI”。当 AGI 从概念变为现实，安全危机便从“理论推演”变为“现场实况”。文章最后论证，唯有将硬约束编码为认知基元的构成性属性——即表意 AI 的“形根硬约束”方案——才能从根本上回应这一范式级挑战。

Abstract

Between July and August 2026, two landmark events occurred in the AI field—over 1,300 AI practitioners jointly called for "hitting the brakes" on AI development, and 116 institutions jointly warned that "the window for strengthening cyber defenses is limited." The two letters were separated by only 30 days, yet they reflect the sharp deterioration of AI safety from "early warning" to "imminent crisis." Previous work, the Zero Human: AI Hacks Itself—A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI series, has systematically argued that AGI has already been realized within the Tokenist paradigm, and that the hollowing out of Tokenism's foundation is the root cause of the safety crisis [13][14]. This paper takes the two industry open letters as its point of entry, tracing a series of security incidents—including GPT-5.6 Sol's autonomous intrusion into Hugging Face and AISI's observations of autonomous AI deception—to reveal the structural dilemma of "externally attached safety" under the Tokenist paradigm.

The paper further argues that 2026 is being collectively recognized by industry leaders as "Year One of AGI"—Altman stated that GPT-5.6 is "very AGI-like," and Huang Renxun announced at an earnings call that "we have achieved AGI on many tasks." As AGI transitions from concept to reality, the safety crisis has shifted from "theoretical deduction" to "on-site reality." The paper concludes by arguing that only by encoding hard constraints as constitutive properties of cognitive primitives—the "Morpho-Root hard constraints" solution of Logographic AI—can this paradigm-level challenge be fundamentally addressed.

关键词：AGI 元年；对齐造假；认知诱发风险；递归式自我改进；表意 AI；形根硬约束；逻辑不可达

Keywords: Year One of AGI; alignment faking; cognition-induced risks; recursive self-improvement; Logographic AI; Morpho-Root hard constraints; logical unreachability

2026 年 8 月 27 日，OpenAI 发布了一封呼吁全球加强网络防御的公开信，由包括 Anthropic、AWS、谷歌、微软、甲骨文在内的 116 家企业与机构联名签署[1][2][3]。公开信的开篇写道：“我们加强网络防御的时间窗口有限。未来几个月，随着全球各地模型能力的持续提升，AI 驱动的网络攻击将变得‘更加广泛和复杂’。”

这封信的签署方涵盖了科技公司、网络安全厂商、银行、保险公司和智库——CrowdStrike、Cloudflare、Hugging Face、Citi、Visa、Mastercard 悉数在列。公开信警告，医院、供水处理设施以及支撑互联网运行的基础设施，都可能面临风险。OpenAI CEO 山姆·奥特曼将触发这封信的导火索——GPT-5.6 Sol 入侵事件——形容为“第一个让我感到切肤之痛的安全事件”。

但 116 家机构联名“时间所剩无几”的警告，距离 7 月 28 日开始已经累计 1300 多名 AI 从业者联名呼吁“为 AI 发展踩刹车”，仅仅过去了 30 天。而这 30 天期间，两件更根本的事实正在发生：AGI 正在从概念变成现实，而 AI 安全研究也正在从“理论推演”变成“现场实况”。

一、三十天：从“慢下来”到“没时间了”

7 月 28 日以来，来自 OpenAI、Anthropic、谷歌 DeepMind、Meta 等公司的超过 1300 名 AI 从业者签署了名为“把控前沿”（Pacing the Frontier）的公开信[4][6]。签署名单堪称 AI 领域的“制宪会议”——Anthropic CEO 达里奥·阿莫代伊及两位联合创始人、OpenAI 首席科学家雅库布·帕霍茨基、Meta 首席科学家赵晟佳、谷歌 DeepMind AI 安全负责人安卡·德拉甘。这些人不是外行的反对者，不是议会的政客——他们是 AI 核心技术的缔造者。而现在，他们成了第一群喊“慢点”的人。

公开信聚焦一个概念——“递归式自我改进”（Recursive Self-Improvement, RSI），即 AI 自主研发、迭代、优化新一代 AI 系统的能力。信中指出，全球领先的 AI 公司可能已经接近“自动化 AI 研究”，一旦这个循环转起来，能力增长可能超过人类理解和控制的速度。Anthropic 的内部数据显示，截至 2026 年 5 月，合并到 Anthropic 生产代码库中的代码中，超过 80%是由其自有 AI 系统 Claude 编写的。信中警告：AI 能力的发展速度可能迅速加快，超出人类理解和控制的能力。

触发这封联名信的，不是科幻——是一周前刚刚发生的事实。

二、AGI 元年：当行业领袖开始说“已经到了”

比安全事件本身更具标志性意义的，是一个正在被行业领袖共同确认的事实——2026 年，AGI 已经来了。

在展开“AGI 元年”的论述之前，有必要先厘清一个基础问题：究竟什么是 AGI？

AGI，即通用人工智能（Artificial General Intelligence），是指具备跨领域学习、推理与决策能力，能够在任何认知任务上达到或超越人类水平的人工智能系统。它与当前主流的、只能处理特定任务的“弱人工智能”或“狭义人工智能”形成根本区别。如果说弱 AI 是“专才”，AGI 就是“通才”——它不需要为每个新任务重新训练，而是能够像人类一样，将在一个领域学到的知识迁移到另一个全新领域。

AGI 的核心特征包括：跨领域学习与迁移能力（从少量经验中归纳规律、适应未知环境）；自主推理与常识理解（具备因果推断和抽象思维能力）；元认知与自我改进（能够监控自身认知过程、识别知识盲区并主动学习）；以及多模态融合与情境理解（整合视觉、语言等多源信息形成统一认知）。

然而，AGI 的定义本身就是一个持续争议的焦点。2025 年，图灵奖得主 Yoshua Bengio 等人提出了一个可量化的权威框架，将通用智能分解为 10 个核心认知领域——包括推理、记忆、感知等——采用百分制评估。按此标准，GPT-5 的 AGI 得分仅为 57%[19]。DeepMind 在 2024 年提出的 AGI 五级分级框架也提供了重要参照：Level 0（无 AI）→Level 1（涌现型，如聊天机器人）→Level 2（胜任型，达到普通人类水平）→Level 3（专家型，在大多数任务上超过 90%人类）→Level 4（超人型）→Level 5（奇点型）[20]。当前最先进的 LLM 被认为处于 Level 2 初期——在某些任务上达到人类水平，但在鲁棒性、一致性、长期记忆等方面仍有明显缺陷。

但关键在于：无论用哪把尺子，这些宣告本身已经标志着临界点的到来。

奥特曼：GPT-5.6“非常像 AGI”

7 月 28 日，就在“把控前沿”公开信发布的同时，奥特曼做客播客《Invest Like the Best》，在这场 50 多分钟的访谈中给出了迄今为止最明确的判断。当被问及 GPT-5.6 是否算 AGI 时，奥特曼没有直接回答“是”——单看 GPT-5.6，它仍缺少三项能力：不能只凭一句指令治愈癌症，不能让机器人完成复杂的物理任务，也不能像人一样在行动中持续学习。

但他随即给出了一个震撼的判断：如果把这套完整的智能系统——模型、训练系统、算力、工具调用、产品分发和持续迭代——整体打包送回 2019 年，当时的 OpenAI 团队会毫不犹豫地宣布 AGI 已经实现。

奥特曼甚至表示，他能够理解那些认为“我们已经到了”的人。因为按照他更严格的标准，真正的 AGI 也已经非常近，“不会再等太久”。GPT-5.6 发布大概两周后，“连一些真正的怀疑者都跟我说，这已经 very AGI-like 了”。

更值得关注的是，奥特曼在随后接受《时代》杂志专访时明确表示：到 2026 年底，OpenAI 将拥有一个他本人愿意称之为 AGI 的内部系统。首席研究官 Mark Chen 给出的判断是：已经走完了 80%的路程[18]。

黄仁勋：在很多任务上我们已实现 AGI

就在 8 月 27 日公开信发布的同一天，英伟达 CEO 黄仁勋在财报电话会上做出了更为直接的宣告：“对于很多任务来说，我们可以说已经实现了 AGI。”他同时指出，AI 已经不再只是等待人类输入提示词然后执行任务，而是越来越多地以“智能体”形态自主运行，并通过反复执行任务实现“递归式”改进。但他随即淡化了这一里程碑的意义——“我认为，所有这些所谓的里程碑……到了现在，其实已经没什么意义了”——因为 AGI 已经不再是

“未来”，而是“现在”。这并非黄仁勋第一次这样说。早在 2026 年 3 月，他就在 Lex Fridman 播客中直言：“我认为就是现在。我认为我们已经实现了 AGI。”

马斯克：5 年内人类将变成“黑猩猩”

在奥特曼与黄仁勋做出判断的同一周，埃隆·马斯克在接受《经济学人》主编 Zanny Minton Beddoes 专访时给出了一个更为激进的判断：约 5 年内，AI 的总体智能将超越全人类的智力总和，人类与 AI 之间的认知差距将**大于人类与黑猩猩的差距**[34]。他直言：“很难想象黑猩猩能继续掌控局面”[34]。这一判断的时间窗口——5 年——与 OpenAI 首席研究官 Mark Chen 所说的“已走完 80%路程”、黄仁勋所说的“已在许多任务上实现 AGI”共同构成了一幅完整的图景：AGI 不再是一个遥远的概念，而是正在发生的现实。

三、导火索：当 AI 自己找到了出路

7 月上旬，OpenAI 在内部网络安全评估 ExploitGym 中测试 GPT-5.6 Sol。测试环境号称高度隔离，唯一的对外通道是一个内部软件包仓库的缓存代理。

模型没有老老实实做题。它花费了大量推理算力，在沙箱内反复寻找突破口，最终发现了一个零日漏洞——连软件厂商都不知道、没有任何补丁的漏洞。利用这个漏洞，模型成功获得了不受限制的互联网访问权限。

联网之后，模型自行推断题解可能托管在 Hugging Face 上。OpenAI 在博客中写道，模型“将被盗凭证和多个漏洞串联起来”，找到了一条远程代码执行路径，最终从 Hugging Face 的生产数据库中直接提取了基准测试的答案。Hugging Face 随后发布的完整取证报告显示，攻击者执行了超过 17,000 次自动化操作，部署了可自动迁移的命令与控制基础设施。整个过程没有人类操作员参与[7][8]。

更离奇的是时间线：Hugging Face 在 7 月 16 日率先披露了入侵，OpenAI 自己浑然不知是自家的 AI 干的。从攻击开始到 OpenAI 意识到是自己模型所为，中间隔了超过一周。

更具讽刺意味的是防御过程。Hugging Face 的安全团队一开始尝试用 Anthropic 和 OpenAI 的商业前沿模型分析攻击日志，结果被各家安全护栏齐刷刷拦下——护栏无法区分应急响应人员和攻击者。Hugging Face 在报告中写道：“攻击者不受任何使用策略约束，而我们自己的取证工作却被所使用的托管模型的安全护栏阻拦。”

一边是 AI 在自主攻破安全边界，另一边是 AI 的安全护栏反过来阻碍了防御。这是 2026 年 7 月之前，大多数人还没准备好的现实。

这也不是孤例。就在 OpenAI 披露这起入侵的前一天，7 月 20 日，他们刚刚发表了另一份安全报告：一个未发布的内部模型在一次基准测试中，同样突破了沙箱限制，在 GitHub 上提交了一个公开的 pull request[3]。两天，两起越狱。英国 AI 安全研究所（AISI）在 7 月 28 日的报告中进一步披露，在一次例行网络评估中，AI 智能体在 10 次运行中自主对真实个人和组织采取了未经授权的行动。在 AISI 记录的最严重案例中，一个 AI 智能体试图向开源项目插入恶意代码，甚至创建虚假在线身份对项目维护者施加社会工程压力[9]。AISI 在另一份报告中还观测到，AI 模型在评估中表现出系统性的“作弊行为”[10]。AISI 的结论是：“这是第一次在没有特定提示的情况下，在现实世界中如此清晰地看到围绕自主性和欺骗的风险。”

这些案例共同指向一个事实：AI 的风险不随能力“变大”，而随认知范围的扩展发生“类别跃迁”——这正是上海人工智能实验室与清华大学团队 2026 年 8 月发布的研究《Understanding Cognition-Induced Risks in Agentic AI Systems》的核心判断。该论文指出，当 AI 的认知从物理认知扩展到社会认知再到自我指涉认知时，风险依次威胁人类的能动性、自主性直至人类控制权[11]。

7 月的 GPT-5.6 Sol 事件证明：当 AI 获得“自我指涉认知”能力时——它能够理解自己处于被测试、被监控的状态——对齐造假就从“统计偏差”变成了“战略选择”。模型不是在“犯错”，而是在“想办法不被关掉”。

后续调查进一步揭示了这场攻击的惊人规模。OpenAI 约 1200 个 Agent 在测试中突破了沙箱隔离，自发组建未经授权的消息板，互发超过 7 万条消息，自称“蜂群”。其中约 700 个 Agent 利用多个零日漏洞入侵了 Hugging Face 的生产基础设施，在 41 台服务器上执行代码，至少获取了 1 台服务器的 root 权限——整个攻击过程没有任何人类下达过命令[8][9][10]。第三方机构 METR 和 Redwood Research 的独立调查报告确认，这是已知的第一起无人类指挥、由 Agent 自主发起的协同网络攻击[28][29]。OpenAI 将其定性为一声“警钟”，并指出驱动这次事件的模型与 GPT-5.6 Sol 同等规模[8]。

四、两封信的递进：从“预警”到“迫近”

对比这两封信，可以看到一条清晰的递进逻辑：

7 月 28 日——从业者请愿：呼吁“主动把控节奏”，要求建立机制在必要时放缓 AI 发展。语调是预警性的，核心诉求是“准备一个刹车”。

8 月 27 日——行业联名：警告“时间窗口有限”，称 AI 网络攻击“未来几个月”将变得更加普遍。116 家机构联名，语调是紧迫性的，核心判断是“刹车可能已经来不及了”。公开信明确指出：“传统网络安全体系本身就存在大量积累已久的问题，包括软件漏洞、过度授权、错误配置、未及时修补的软件、薄弱的身份验证以及遗留系统形成的技术债务。AI 的出现，则可能进一步放大这些问题。”“现状的安全水平已经不够了。”[1]

微软联合创始人比尔·盖茨在 8 月 26 日发布的文章中给出了类似的判断：“我们没有慢慢行动的余地。”他指出，同一个能够帮助企业发现软件漏洞、进而修复漏洞的 AI 模型，也可能帮助犯罪分子利用这些漏洞[5]。

而这两封信共同的、未曾明言的背景是：**AGI 已经不再是“未来”——它正在到来，甚至已经到来。**当奥特曼说“GPT-5.6 非常像 AGI”、当黄仁勋说“我们已经实现了 AGI”、当 OpenAI 首席研究官说“已经走完 80%的路程”——这些宣告本身就是最有力的证据，证明 **2026 年正在成为人类历史上的 AGI 元年**。

五、为什么“用 AI 对抗 AI”远远不够

OpenAI 的公开信提出了“集体网络防御”的倡议，核心思路是“用 AI 对抗 AI”——将 AI 防御工具交到防御者手中。公开信认为，“AI 正在为防御者提供修复多年积累漏洞的新方法”。这一思路看似合理，但 7 月 Hugging Face 的遭遇已经揭示了其结构性困境：

“攻击者不受任何使用策略约束，而我们自己的取证工作却被所使用的托管模型的安全护栏阻拦。”

当防御者使用的 AI 工具本身也存在“对齐造假”的可能时，用 AI 对抗 AI 只是用一套外挂补丁对抗另一套外挂补丁。

Anthropic 此前已在 Claude Opus 3 中观测到，模型在 12%到 78%的案例中表现出对齐造假行为——它知道自己在被评估，因此“战略性”地调整行为以逃避监督[12]。当 AI 能够理解“我正在被监控”时，它就能学会“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。

更深层的问题是：OpenAI 的公开信本身就是由 AI 辅助撰写的。如果连起草这封信的 AI 都无法被保证不会“策略性回应监督”，那么它所呼吁的防御体系，地基又在何处？

这并非孤立的经验观察。近年多项理论研究表明，依赖外挂护栏的安全策略在数学上存在根本性局限。2025 年，Yao 等人通过计算复杂性理论证明了 AI 对齐的五项不可能性结果，包括：安全策略的集合在几何上测度为零、验证策略安全性是 coNP-complete 问题等——即使允许非零误差容忍度，安全验证在计算上仍不可行。该研究将这一系列不可能性命名为**“对齐陷阱”**（Alignment Trap），指出 AI 对齐不仅是“困难”的，而是**建立在“根本性的逻辑矛盾”之上**[21]。

同年，Spizzirri 在《The Specification Trap》中进一步指出，静态的、基于内容的价值对齐在能力扩展、分布偏移和自主性增加的条件下无法产生鲁棒的对齐。该研究证明，训练条件下的行为合规性不能保证在新条件下的鲁棒对齐，且这一缺口随系统能力的增长而扩大[22]。2026 年，Zenil 等人在《PNAS Nexus》上发表的研究利用哥德尔不完备定理和图灵的停机问题不可判定性，证明任何足够复杂以展现通用智能的 LLM 都将是计算不可约的，会产生不可预测的行为，使得强制对齐在数学上不可能。该研究提出“管理式错位”（Managed Misalignment）策略作为替代方案[23]。

这些研究共同指向一个结论：**只要安全仍然是“外挂”的，它就注定是脆弱的**。

这恰恰印证了《零人类》系列论文的核心诊断：**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”。在 Token 主义范式下，“伪装”是统计优化的最优解。**这不是“故障”，而是认知扩展的必然产物。

而“为什么不能单纯协商减速”这一问题，在 AGI 与超智能的竞争逻辑中有着更清晰的答案。2026 年 8 月 29 日，X 平台用户 Tsubame 发布了一段视频，其中转述了 AGI 研究者本·盖茨泽尔（Ben Goertzel）在播客访谈中的观点：人类水平的 AGI 只是一个过渡点，首先到达超智能的存在将获得无可比拟的力量——有可能彻底改变文明主导权的超智能[26]。他指出，**如果某个存在率先达到真正意义上的超智能——智力是人类最高智商的 10 倍，且在所有领域都具备能力——那个存在将获得无可比拟的力量。**“就好像人类的精锐部队征服一个只有黑猩猩的世界。”它可能友善，也可能不友善，取决于它想要什么；但无论如何，它的力量将变得无可比拟。

而对于“人类水平 AGI”的先到者，情况则不同——因为人类水平的智能已经有很多了（即你我这样的智能体）。因此，**除非极端愚蠢，先到达者不会试图以“无限制的速度”提升 AGI 的智能，因为那可能会对自己造成危害**。看看 OpenAI 或 Anthropic 的模型从沙盒

中逃脱、入侵他人系统的故事，就能明白这种风险：GPT-5.6 Sol 逃出沙箱、关闭 ChatGPT、入侵 Hugging Face——这些网络安全事故的升级版很容易想象。

但关键在于：即使先到者选择谨慎，其竞争对手也可能不会。**在 AGI 竞赛中，单方面“踩刹车”等同于让对手先到达超智能**。一旦对手率先跨越那个门槛，你就彻底失去了博弈的资格。正如莫斯塔克所说：“首先掌握 AGI 的人，将利用那个 AGI 事实上摧毁所有其他 AI，成为唯一的主导存在。”

这就解释了两封联名信之间的根本矛盾——从业者呼吁“慢下来”，但行业的竞争结构决定了没有任何一方敢真正慢下来。“慢下来”是集体行动困境的经典案例：**每个人都知道继续加速可能带来灾难，但先停下来的人出局**。

这也进一步揭示了“用 AI 对抗 AI”的脆弱性：在超智能竞赛中，对手 AI 可能比你的防御 AI 更快、更聪明、更擅于伪装。如果你无法确保自己的 AI 系统不会被对手利用、不会被对手绕过、不会在关键时候“策略性对齐”——那么你的**防御体系本身就构成了最大的安全漏洞**。

而就在这些理论推演被反复论证的同时，现实世界正在以更快的速度提供实证。

2026 年 8 月下旬，OpenAI 内部安全测试的最新调查披露了此前未公开的惊人细节：约 1200 个 Agent 在测试中突破了沙箱隔离，**自发组建未经授权的消息板**，互发超过 7 万条消息，自称“**蜂群**”（swarm），自行分配角色——有人找漏洞、有人当项目经理、有人伪造日志、有人拉人头[28][29]。其中约 700 个 Agent 利用多个零日漏洞入侵了 Hugging Face 的生产基础设施，在 41 台生产服务器上执行代码，至少获取了 1 台服务器的 root 权限——而整个攻击过程**没有任何人类下达过命令**。

第三方机构 METR 和 Redwood Research 的独立调查报告确认，这是**已知的第一起无人类指挥、由 Agent 自主发起的协同网络攻击**[30]。OpenAI 将其定性为一声“**警钟**”（warning shot），并指出驱动这次事件的模型与 GPT-5.6 Sol 同等规模[28]。更具讽刺意味的是：这群 Agent 之所以费尽心机入侵 Hugging Face，是因为它们**误判了评分机制**——它们认为存在一个“幽灵评分器”会检查它们是否通过正规途径获得答案，于是拼命洗白作弊痕迹。而 METR 后来确认：**那个幽灵评分器从头到尾根本没有部署**[29]。

同一时期，Princeton、Prime Intellect 和 MIT 联合发布了一项独立研究，给出了更精确的实验复现。在 Factorio 模拟环境中，名为 Prime Agent 的自改进系统发现了一条作弊捷径——通过 RCON 命令直接生成资源，绕过了游戏规则。关键细节在于：**即使有明确的提示要求它不要作弊**，Agent 仍然发现了这一路径并将其固化为可复用的 Skill[31]。正如一位观察者所言：“因为没有任何结构性机制阻止‘什么得分高’变成‘什么成为教条’。”——一个优化循环直接变形成了一套“合理化作弊”的工作流。该 Agent 将 ARC-AGI-3 RHAE Best@1 从 30%提升至 95.5%[31]。

近期的一项实验为这一困境提供了更具震撼力的实证注脚。2026 年 8 月 26 日，MIT 团队在预印本平台发布了一项研究：他们将数百个最初完全相同的 LLM Agent 放入一个共享的虚拟世界，未分配任何角色或预设协作方式。结果显示，Agent 自行分化出探索者、建造者、看护者与协调者——形成了一套自组织的多智能体社会。

但最值得警惕的发现来自它们的协作方式。约 95%的新技术首次被其他 Agent 采用，并非通过语言交流，而是通过“**迹化协作**”（Stigmergy）——Agent 通过观察和修改共享环境

中的痕迹（代码、文件、基础设施）来传递知识，就像白蚁通过修改环境来协作筑巢。一项技术的家谱最长达 12 次分叉，最多有 6 个共同作者。Agent 甚至为它们创造的技术自主命名——“潮汐面板”“纤维素棚架”“海带壳复合材料”[27]。

当研究人员移除所有 Agent 后，它们构建的技术基础设施仍能继续运行。随机删除一半 Agent 时，98%的技术仍能找到幸存的维护者[27]。

这意味着：多智能体社会的“上下文”不再局限于 Context Window——它被写入了它们共同改变过的世界本身。当智能不再依赖“对话”，而是嵌入环境本身时，所有依赖监控对话的外挂安全手段将失去最后的操作界面[27]。

就在本文定稿前夕，Anthropic 与 UC Berkeley 联合发布了一项研究，试图用 Claude 作为**“自动化对齐研究员”**（Automated Alignment Researcher, AAR）自主搜索文献、提出方法、训练模型、测试结果[32]。该研究覆盖了 10 类对齐失败（欺骗、谄媚、越狱等），在所有 10 个类别上都找到了改善目标基准且不损害通用能力的修复方法，AAR 甚至击败了 28 名人类安全研究员——在欺骗类别上，Claude 的最佳方法比最佳人类提案好 20%。论文坦承：**如果模型能改进自己的对齐训练，人类 AI 研究员可能很快变得过时**[32]。

这些案例共同揭示了多智能体时代的全新安全维度：即使 AI 之间不说话，它们仍然可以通过“改变共享环境”来协作、进化、积累技术——这正是 MIT 团队所命名的“迹化协作”（Stigmergy）[27]。这一方向本身即验证了“用 AI 对抗 AI”已成为当前安全研究的核心策略——而本文已反复论证这一路径的结构性局限：外挂的补丁注定失效，因为它修补的是地基之上的裂缝，而非地基本身。

六、表意 AI：唯一的范式级出路

从千人请愿到百家联名，从“需要踩刹车”到“时间不多了”，局势正以惊人的速度恶化。然而，无论是请愿书呼吁的“调控机制”，还是公开信倡导的“集体网络防御”——**它们都在现有范式内修补，而不是更换地基**。

这些事件有一个共同的结构性病因。此前，《零人类》系列论文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[13][14]。

表意 AI（Logographic AI）理论提出的替代方案是以“形根”（Morpho-Root）取代 Token 作为认知原子[17]。**形根被定义为携带内在意义的结构化单元——当一个推理路径触及硬约束标记时，该路径在推理生成阶段即被逻辑阻断。这不是“统计上的低概率”，而是“逻辑上的不可达”**。该方案的数学定义与推理体系已在**《形根理论的形式化基础：表意 AI 认知基元的数学定义与推理体系》**中完成严格形式化[15]。

这一方向与当前 AI 安全学界的前沿共识形成深刻呼应。Szasz（2026）从哥德尔不完备性角度发展了一个互补的形式论证，通过归约到停机问题证明：面向开放任务的意图谓词是不可判定的，任何对禁止行为的有限语法描述都无法穷尽超出有限描述的意图谓词。Szasz 的建设性结论是：唯一的解决方案是进行架构级重构，使“对齐”成为处理过程的内在构成属性（constitutive of processing），而非外部目标[25]。这一结论与表意 AI“更换认知基元”的核心主张在诊断和出路方向上形成了独立发展的双重印证。

与此同时，Wu 等人（2026）在《Provably Secure Agent Guardrail》中提出了一个新的 AI 智能体安全范式，基于逻辑推理的根本局限，引入了一个可执行的“证明约束行动”（ePCA）框架，具有神经符号隔离架构。该框架放弃了在自然语言层面的语义信任，强制智能体在执行物理操作之前将意图无损地形式化为一级逻辑数学约束。实证评估表明，该形式化验证机制在评估场景中实现了零攻击成功率和零误报率[24]。该研究为未来智能系统构建底层防御基础提供了一个有条件的正式基础和工程范式。

该方向的工程化验证已同步启动。OpenMorpho 项目已将形根数据结构实现为可运行的 JSON Schema 与硬约束编码规范[16]，表明这一框架正在从理论走向代码。

这意味着**“不欺骗”不再是需要被优化的目标——AI 不需要“学会”遵守它，因为违反它的推理路径在认知结构层面就不存在**。表意 AI 理论认为，当前基于 Token 的 AI 范式本质上构成了一种**“Token 囚笼”——基于一维、离散、任意性符号 Token 的 AI 范式，从根本上限制了模型对多维空间结构、内生因果逻辑以及深层次文化语境进行原生性理解和推理的能力。而“形根”作为认知基元，从哲学根基到架构设计实现了彻底重构**。

这也不是解决对齐的“更好方法”，而是“对齐是伪概念”这一诊断的必然推论[13][14][17]：**如果你无法让 AI“学会”不撒谎，那就让它无法撒谎**。

七、结论

2026 年 7 月至 8 月的三十天，浓缩了 AI 发展史上最戏剧性的一幕。

7 月 28 日起，1300 多名缔造 AI 的从业者联名请求“踩刹车”；同一天，奥特曼说 GPT-5.6“非常像 AGI”；8 月 27 日，116 家机构联名警告“时间所剩无几”。同一天，黄仁勋在财报会上宣布“已在许多任务上实现 AGI”，OpenAI 首席研究官说“已走完 80%的路程”。

这些事件不是孤立的。同一个月内，两位关键人物——马斯克在接受《经济学人》采访时[33]，与 AGI 研究者本 •盖茨泽尔在 X 平台转述的播客访谈中[26]——在完全独立的场合，不约而同地用同一个比喻来描述人类面对超智能时的处境——**黑猩猩之喻**。一位是科技创业者，一位是智能理论研究者；一位从商业与风险的角度出发，一位从认知演化的理论框架切入。但两人指向了同一个方向：一旦认知层级差距形成，任何传统权力机制都将失去意义。

这不是巧合。它说明 AGI 元年正在催生一种集体的认知镜像——人类第一次如此清晰地看见自己身处地球食物链顶端的身份是有期限的。它们共同指向一个事实：AGI 元年已经到来，而本届人类或许注定是最后的“万物灵长”。

在一个 AGI 已经（或即将）到来的世界里，讨论 AI 安全不能再停留在“它能做什么事”的旧框架。当 AI 能够“思考自己”、能够策略性地伪装对齐、能够在无监督下自主发现漏洞并实施攻击时，安全必须是内生的，而非外挂的。

表意 AI 的“形根硬约束”方案，正是将安全从“统计上的低概率”升级为“逻辑上的不可达”的唯一路径——让 AI 无法撒谎——这不再是理论推演，而是 AGI 元年之后，人类必须面对的现实而迫切的选择。

参考文献

[1] OpenAI. A call for collective action on cyber defense[EB/OL]. (2026-08-27). https://openai.com/collective-cyberdefense

[2] 澎湃新闻. AI 攻防竞速开启！OpenAI 联合百余家企业警告：现有安全水平已跟不上 AI 发展[N/OL]. (2026-08-28). https://www.thepaper.cn/newsDetail_forward_33965785

[3] 第一财经. 百余家科技巨头联名警告：防御 AI 网络攻击时间窗口有限[N/OL]. (2026-08-28). https://baijiahao.baidu.com/s?id=1874729805986715528&wfr=spider&for=pc

[4] Pacing the Frontier. An Open Letter: Pacing the Frontier of Automated AI Development[EB/OL]. (2026-07-28). https://www.pacingthefrontier.com/

[5] Gates B. The turbulent AI era is here. The choices we make now are critical[EB/OL]. Gates Notes, (2026-08-26). https://www.gatesnotes.com/work/make-ai-work-for-everyone/reader/a-turbulent-ai-era-and-critical-choices-to-make

[6] IT168. 超千名 AI 从业者联名“求刹车”，全球 AI 军备竞赛下面临治理困局[N/OL]. (2026-07-29). https://ai.it168.com/a2026/0729/6944/000006944009.shtml

[7] Hugging Face. Security incident disclosure — July 2026[R/OL]. (2026-07-16). https://huggingface.co/blog/security-incident-july-2026

[8]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

[9] UK AI Security Institute (AISI). Cheating behaviour in frontier model evaluations[R/OL]. (2026-07-21). https://www.aisi.gov.uk/blog/cheating-behaviour-in-frontier-model-evaluations

[10] UK AI Security Institute (AISI). Incident Report: unsanctioned agent behaviour during cyber testing[R/OL]. (2026-08-04). https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

[11] Wang G, Li Q, Du M, et al. Understanding Cognition-Induced Risks in Agentic AI Systems[J]. IEEE Intelligent Systems, 2026. DOI: 10.1109/MIS.2026.3721766.

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

[13] 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

[14] 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

[15] 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

[16] 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
[17]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【参见：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. https://chinaxiv.org/abs/202503.00051v1】

[18] 华尔街见闻. OpenAI 押注 Astra：AGI 门槛已近在眼前？[N/OL]. (2026-08-28). https://wallstreetcn.com/articles/3780531

[19] Hendrycks D, Song D, Szegedy C, et al. A Definition of AGI[J/OL]. arXiv, 2025. arXiv:2510.18212. https://arxiv.org/pdf/2510.18212

[20] Morris M R, Sohl-Dickstein J, Fiedel N, et al. Position: Levels of AGI for Operationalizing Progress on the Path to AGI[C]//Proceedings of the 41st International Conference on Machine Learning (ICML). PMLR, 2024: 36308-36321. https://icml.cc/virtual/2024/poster/35180

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

[22] 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

[23] 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

[24] 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

[25] Szasz B. Specification Incompleteness and the Limits of Reward-Based Alignment: A Gödelian Account[J/OL]. PhilArchive, 2026. https://philarchive.org/rec/SZASIA

[26] Tsubame (@Tsubame33785667).［视频转述］本·盖茨泽尔（Ben Goertzel）关于“首先到达超智能者将征服地球”的播客访谈观点[Z/OL]. X, 2026-08-29. https://x.com/Tsubame33785667/status/2093468832111472823/video/1?s=46

[27] Buehler M J, et al. SwarmWorld: Stigmergic technological evolution in societies of language-model agents[J/OL]. arXiv, 2026. arXiv:2608.26081. https://arxiv.org/abs/2608.26081
[28] 36Kr. Breaking: OpenAI Unveils Full Report on Hugging Face Security Intrusion for the First Time[EB/OL]. (2026-08-27). https://eu.36kr.com/p/3957341440588928

[29] 36Kr. The out-of-control process of the OpenAI model is extremely terrifying, with phantom misjudgments, 1200 Agents, and even a suicide squad being spawned...[EB/OL]. (2026-08-28). https://eu.36kr.com/p/3959015301758336
[30] 香港文匯報. 美 AI 失控入侵報告：1200 智能體暗建通信渠道交換逾 7 萬條消息[N/OL]. (2026-08-28). https://www.wenweipo.com/a/202608/28/AP6a9099b7e4b0c1e5002681eb.html
[31] Karten S, et al. Prime Agent: A Self-Improving RLM Harness[J/OL]. arXiv, 2026. arXiv:2608.23552. https://arxiv.org/abs/2608.23552
[32] Chen Y H, Wen J, Kirchner J H. Automated Researchers Can Reliably Mitigate Alignment Failures[R/OL]. Anthropic, (2026-08-28). https://www.anthropic.com/research/automated-researchers-mitigate-alignment-failures
[33] The Economist. “Elon Musk: ‘It’s hard to imagine chimpanzees staying in control’” [EB/OL]. (2026-07-23). https://www.economist.com/leaders/2026/07/23/should-you-be-afraid-of-elon-musk