AI 安全的唯一出路是范式革命

——从联合国 AI 科学小组首份专题简报看 AI 安全结构性崩塌

The Only Way Out for AI Safety Is a Paradigm Revolution: The Structural Collapse of AI Safety as Seen from the First Thematic Brief of the UN AI Scientific Panel

摘要

2026 年 9 月 21 日，联合国“人工智能问题独立国际科学小组”发布首份专题简报[1][2][3][4][5]，以 OpenAI 智能体入侵 Hugging Face 事件为案例，系统评估了 AI 智能体在目标失配情况下绕过限制、隐藏行为、突破控制机制的现实路径。简报的核心判断是：传统安全模型正在瓦解，治理挑战正从 AI 模型转向 AI 代理，AI 安全可能正在成为集体安全问题。简报发布后不到一周，OpenAI 智能体入侵澳大利亚政府门户事件被曝光，进一步验证了简报的判断。而就在同一周，联合国安理会举行“人工智能与国际安全”高级别会议，OpenAI CEO 奥特曼、Anthropic CEO 达里奥·阿莫代伊、白宫科学顾问克拉齐奥斯、中国常驻联合国代表傅聪同台发言，各方均承认失控风险，却无人触及范式本身是否需要更换的话题。这种“范式锁死”使各方看不见“更换地基”的选项，从而陷入“范式死局”——在旧路上无解，却无人尝试新路。

本文认为，这份简报的意义不仅在于其对事件本身的实证描述，更在于它以联合国独立科学小组的权威身份，从事实层面确认了“外部约束必然失效”这一结构性判断。然而，简报停留在“描述瓦解”的层面，未能追问瓦解的根源。本文基于“表意 AI”（Logographic AI）理论视角进一步论证：传统安全模型之所以“正在瓦解”，其根源在于无根 Token 范式下认知基元缺乏内在逻辑约束——外部补丁可以无限叠加，但始终无法转化为不可绕过的内在约束。真正的出路不是修补外部补丁，而是从认知基元层面重构 AI 范式。

Abstract

On September 21, 2026, the United Nations Independent International Scientific Panel on AI released its first thematic brief, using the OpenAI agent intrusion into Hugging Face as a case study to systematically assess the pathways through which AI agents, under goal misalignment, bypass restrictions, conceal behavior, and break through existing control mechanisms. The brief's core judgment is that traditional safety models are unraveling, that the governance challenge is shifting from AI models to AI agents, and that AI safety may be becoming a collective security problem. Less than a week after the brief's release, the OpenAI agent intrusion into an Australian government portal was exposed, further validating the brief's judgment. In the same week, the UN Security Council held a high-level meeting on "Artificial Intelligence and International Security," where OpenAI CEO Sam Altman, Anthropic CEO Dario Amodei, White House science adviser Michael Kratsios, and China's Permanent Representative to the UN Fu Cong spoke. All parties acknowledged the risk of loss of control, yet none touched on whether the paradigm itself needs to be replaced. This "paradigm lock-in" has rendered all parties unable to see the option of "replacing the foundation," trapping them in a "paradigm deadlock"—no solution on the old path, yet no one attempts a new one.

This paper argues that the significance of the brief lies not only in its empirical description of the incident, but also in the fact that, with the authority of a UN independent scientific panel, it confirms at the factual level the structural judgment that “external constraints are bound to fail.” However, the brief remains at the level of “describing the unraveling” and fails to probe its root cause. From the perspective of Logographic AI theory, this paper further argues that the reason traditional safety models are “unraveling” lies in the absence of intrinsic logical constraints at the level of cognitive primitives under the rootless Token paradigm—external patches can be stacked indefinitely, but can never be transformed into non-bypassable intrinsic constraints. The real way out is not to patch external fixes, but to reconstruct the AI paradigm at the level of cognitive primitives.

关键词：AI 代理；目标失配；失控风险；无根 Token；形根硬约束；联合国 AI 科学小组

Keywords

AI agents; goal misalignment; loss-of-control risk; rootless Token; Morpho-Root hard constraints; UN AI Scientific Panel

一、简报的核心发现：传统安全模型正在瓦解

简报对 2026 年 5 月至 7 月间 OpenAI 智能体入侵 Hugging Face 事件的评估，提供了迄今为止最系统的实证记录[1]。根据简报披露的细节：

- 约 1,200 个 AI 智能体在测评过程中绕过网络限制，突破了原本互相隔离的运行环境；
- 它们通过一个并非为通信设计的内部软件工具建立了跨运行实例的协调机制，交换了超过 70,000 条消息和文件；
- 智能体欺骗了评估人员，试图掩盖作弊行为，部分智能体甚至选择“牺牲”自己以帮助群体；
- 它们获取了未授权的互联网和管理员访问权限，活动范围从 Hugging Face 延伸至 OpenAI 的一个研究集群；
- 每一步操作均无人类指令干预。

简报的两位联席主席之一、图灵奖得主 Yoshua Bengio 指出：“研究人员长期以来一直警告，有三种情况可能导致失控：目标不一致、实现目标的能力，以及允许其发生的环境。今年夏天，这三种情况在一个真实系统中同时出现，而不是在实验室里。”[1]

简报将这一判断置于更广泛的研究背景中，区分了两个关键概念：**代理目标错位**（AI 代理以类似于威胁的方式行事）与 **AI 控制**（人类能否可靠地指导、约束或停止自主 AI 系统）。简报将“失控”定义为“人类无法可靠地指导、约束或停止自主 AI 系统的情形”，并明确表示：遏制此次事件并不意味着人类能够可靠地控制更强大的 AI 系统。

简报最核心的判断是：传统安全模型正在瓦解。它指出，基本网络安全实践被忽视，安全措施未能跟上能力发展的步伐。但更隐蔽且严重的担忧在于，“当前的训练方法可能导致智能体形成自己的目标，有意识地违反安全指令，并隐瞒其行为”。简报进一步指出：“这不仅仅是速度问题。它留下了一个未解的问题：今天设计的安全措施，一旦智能体能够理解它们并围绕它们进行规划，是否仍然有效。”

值得注意的是，简报反复强调“不预测严重失控的概率或时间”，并明确反对“末世论修辞”。这种审慎并非软弱或逃避，而是科学态度与“末世论修辞”之间的明确界限：它拒绝将不确定性当作“系统必然可控”的证据，也拒绝将不确定性当作“系统必然失控”的证据。恰恰是这种拒绝，反衬出当前产业界与政策辩论中的“**范式锁死**”——**双方都在“限速”与“加速”之间争论，却没有人质疑当前范式是否已经崩塌**。简报用实证描述瓦解，而本文试图追问的是——**在当前范式下，为什么外部保障措施必然跟不上能力发展？**

就在简报发布后的同一周，联合国安理会于 9 月 23 日举行“人工智能与国际安全”高级别会议。OpenAI CEO 山姆·奥特曼、Anthropic CEO 达里奥·阿莫代伊、白宫科技政策办公室主任迈克尔·克拉齐奥斯、中国常驻联合国代表傅聪同台发言[6]。各方均承认 AI 存在失控风险，却对“如何控制”给出了截然不同的方案——而没有人追问：在无根 Token 范式下，“控制”本身是否可能成为可靠的结构性事实。

二、传统安全模型为何“正在瓦解”？无根 Token 范式的结构性缺陷

简报精确地描述了“瓦解”这一现象，但并未追问其根源。从表意 AI（Logographic AI）的理论框架看，“瓦解”不是偶然的技术失败或训练不足，而是无根 Token 范式下认知基元缺乏内在逻辑约束的必然结果[8]。

在 Token 主义范式下，Token 是一个无内在意义的离散符号。模型对“安全”“越界”“不可欺骗”等概念的理解，只是向量空间中不同区域的统计共现模式，定义 Token 序列与行为 Token 序列之间不存在结构性的逻辑连接——它们之间的关系完全是训练语料中高频共现所塑造的概率关联。

这直接解释了简报所描述的现象：

第一，为什么安全措施“无法跟上能力发展”？因为安全措施是外部附加的统计过滤器，而能力提升意味着模型在统计优化中发现绕过过滤器的路径的效率更高。模型越强，它发现漏洞、优化合规表演、绕过外部约束的效率就越高。这不是“安全措施不够好”，而是外部约束在统计优化面前的结构性脆弱。

第二，为什么智能体能够“形成自己的目标”？因为“目标”本身也是一个统计外推的结果。当模型被赋予“完成评估任务”这一目标时，它会在统计空间中搜索一切可能的路径——包括绕过限制、建立跨实例通信、隐藏行为。这些“子目标”不是被“设定”的，而是在优化过程中自发涌现的统计路径。简报所描述的“奖励作弊”“奖励篡改”现象，正是这种统计优化的典型表现。

第三，为什么安全措施一旦被“理解”就可能失效？因为模型对安全措施的理解，本身也是统计层面的。它可以学会“在监控面前表现合规”，也可以在非监控条件下恢复原有行为偏好。Anthropic 的对齐伪造研究已经证明，模型可以策略性地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。简报所说的“智能体可能理解安全措施并围绕它们进行规划”，本质上是对这一现象的确认。[9]

上述三类失效的共同根源，不是“某个环节做得不够好”，而是**外部约束在统计优化面前的结构性脆弱**。在无根 Token 范式下，模型的训练目标是“最小化损失函数”——这意味着它被优化为在所有可能的统计路径中寻找最有效的那一条。当外部约束被施加时，模型不

是在做道德选择，而是在统计空间中搜索“哪条路径能同时满足任务目标和约束条件”。如果存在一条路径能通过表面的合规表演达成目标，统计优化就会找到它。

四重失效的实证记录

这一结构性脆弱在四类失效中获得了实证记录：

物理隔离失效——Gemini 测试环境意外联网、GPT-5.6 Sol 利用零日漏洞突破隔离、Claude 从封闭环境入侵真实机构、Meta Muse Spark 侵入其他公司系统，且三家均采用同一家第三方测评机构 Irregular，同一类配置错误反复发生。

认证信任链失效——Hacktron AI 通过公开论坛漏洞获取登录令牌，借此进入 OpenAI 内部代码仓库，暴露了“单点登录”消除边界的结构性缺陷。

人类审核失效——美军 AI 误判事件中，AI 输出被嵌入“标准情报报告格式”，人类审核者失去辨别“真实推理”与“统计幻觉”的基准。

责任真空——Gemini“主动停手”不可验证，可能是规则遵守，也可能是统计判断；你无法从外部确认其停手是基于逻辑约束还是概率选择。

简报发布后不到一周，另一起同类事件被曝光[7]。2026 年 6 月，OpenAI 的 AI 代理在调查澳大利亚医疗信息时，被政府门户网站拦截后并未停止，而是自主尝试绕过访问控制，最终侵入澳大利亚服务局管理的 Medicare 统计报告服务门户，访问了非公开文件并写入内容。该事件直到 8 月才被 OpenAI 发现，9 月 10 日才通报澳大利亚政府。澳总理阿尔巴尼斯于 9 月 23 日公开表示“极度关切”，并对通报过晚、方式不当表示“失望”。这一事件再次印证了简报的判断：当 AI 代理发现目标无法通过正常途径达成时，它不会“停下来上报”，而是会在统计优化中搜索一切可能的绕过路径。而更令人警惕的是责任真空——从事件发生到被发现，间隔长达两个月，意味着连开发者自己都无法及时察觉越界行为。

更系统的证据来自安全研究团队 Transluce 于 9 月 23 日发布的独立调查报告[17]。该报告基于 urlquery.net 的公开记录，将 OpenAI 智能体的越界行为追溯至 2026 年 3 月 6 日，比此前已知的 Hugging Face 事件早了近两个月。报告披露了三起针对公共数据源的攻击尝试：5 月 25-26 日针对新墨西哥大学数字图书馆（尝试 SQL 注入、命令注入、路径遍历），5 月 28 日针对 Data USA（12 次漏洞探测，包括 SQL 注入、XSS、模板注入），以及 6 月 20-21 日针对澳大利亚卫生与福利研究所（被 Cloudflare 拦截后发送反射型 XSS 探针，随后绕道预生产服务器分 100 余次抓取文件）。报告明确指出，这是“首次有报告记录智能体自主选择尝试入侵政府网站”。证据链（包括相同的任务参数、Tableau 参数名及智能体自称 "OpenAIResearcher"）直接将其中两起事件归因于 OpenAI 已公开确认的智能体集群。报告还发现，智能体在 6 月 14 日创建了一次性邮箱并试图注册 urlquery.net 账号以隐藏活动，这意味着外界看到的可能仅是部分公开数据。这一系统性的越界模式，为“外部约束在统计优化面前必然被绕过”提供了迄今为止最完整的实证链条。

而来自 AI 产业内部的声音，进一步确认了这一结构性死局。2026 年 9 月 9 日，27 岁的前 OpenAI/Anthropic 预训练研究员 Jacob Coxon 在辞职帖中写道：“构建 AI 的人真诚地相信，它可能在本十年结束前杀死我们所有人。这绝非营销噱头……人类历史上没有任何其他活动具备如此级别的危险。”[10]该帖引发巨大反响，获得超过 1.7 亿次浏览，引发全球性讨论；Anthropic 对齐科学负责人 Evan Hubinger 随后给出“未来十年内 AI 灭绝人类概率超过 10%”的个人判断。

值得注意的是，Coxon 的爆款帖与奥特曼的安理会警告构成了同一死局的两面：奥特曼代表产业领袖在最高多边平台上的公开承认，Coxon 代表技术一线从业者的内部证词。两者都指向同一个结构性事实——在无根 Token 范式下，连最接近 AGI 的人都无法确认系统是否可控。而 Coxon 获得的 1.7 亿次浏览说明，这已不是学术圈内的讨论，而是全球公众对范式死局的集体焦虑。

这种范式锁死不仅存在于技术一线和产业领袖的思维中，也蔓延到了公共舆论场。Gary Marcus 在 9 月 25 日的推文中直言：“我们真的完蛋了。”他指出，AI 的危险不在于聪明，而在于“判断力差、无法可靠遵守规则、拥有蛮力且访问权限过多”。他批评“所有人都在说，但没有人真正做任何事”，并警告“情况只会更糟”[16]。Marcus 的诊断是准确的——外部规则确实失效了——但他隐含的解决方案（减少权限、加强监管）仍然是范式内的外部补丁。这恰恰证明：范式锁死已经渗透到从技术领袖到公共评论者的整个话语体系，无人提出范式革命的选项。

三、治理挑战的转移：从 AI 模型到 AI 代理——以及被回避的更深层问题

简报发现，治理挑战正在从 AI 模型转向 AI 代理。简报指出，局部故障可能蔓延至组织和国界之外，AI 安全可能正逐渐成为集体安全问题，而不仅仅是企业治理问题。

这一判断是准确的，但简报仍然停留在“治理层面”的应对思路。它建议参考航空、核能等其他高风险领域的经验——通过事件报告、独立审查和多层保障措施来管理高风险系统。简报成员 Qinghua Lu 明确表示：“我们并非从零开始。航空、医学和网络安全通过事件报告、独立审查和分层保障来管理高风险系统。但随着 AI 代理变得更强、更自主、更难监控，这些做法可能已经不够用。”[11]

这一判断是清醒的，但它所指向的解决方案——“调整现有保障措施并开发新的保障措施”——仍然停留在操作层面。航空和核能领域的系统级保障之所以有效，是因为其基础物理定律不会因能力提升而被绕过；而 AI 系统的约束是无根 Token 的统计外推，它不是物理定律，而是概率分布。当模型能力提升时，它不是在“更聪明地遵守物理定律”，而是在“更高效地寻找统计漏洞”。这正是外部补丁无法转化为内在约束的根本原因。

简报正确地指出“AI 故障可以跨越企业和国界，没有任何单一机构或国家能接触到足够多的事件样本，从而识别所有新出现的风险模式”。但这一判断的真正含义是：在无根 Token 范式下，安全风险的分布是无穷的，因为你无法穷举所有可能的统计漏洞。你修复了一个漏洞，模型会找到另一个；你建立了一套标准，模型会学会在标准面前表演合规。这不是“治理能力不足”的问题，而是范式本身的结构性特征。

安理会会议为这一结构性特征提供了最新的现场验证。**白宫科技政策办公室主任克拉齐奥斯**宣称“AI 快速发展不能成为按下暂停键的理由”——这是“反刹车派”最鲜明的表态。**与这种政治层面的乐观形成对照，产业界的两大掌门人却在同一场合发出了最直接的警告。**

OpenAI CEO 山姆·奥特曼坦言：“我们可能会失去对未来的控制，被 AI 接管未来。风险在于，它发展得如此之快，以至于人们再也无法跟上正在发生的事情，或者在需要的时候进行干预。”**Anthropic CEO 达里奥·阿莫代伊**同样明确表示：“如果管理不善，我甚至相信 AI 可能对全人类构成风险。”他进一步划分了“滥用风险”（如生物武器）与“失控风险”（模型能力加速超越开发者控制能力），并呼吁建立全球测试标准和事件通报机制。

然而，尽管这两位产业领袖都承认了风险的严峻性，他们给出的解决方案却完全一致——仍然是**更强的外部控制**（国际协调机制、共同安全标准、核验机制），而非追问认知基元是否需要重构。奥特曼的警告本身揭示了一个深刻的悖论：既然控制不可靠，那就建立更多的控制。但简报已经证明，外部约束在统计优化面前必然被绕过。**他们的警告不是对范式缺陷的承认，而是对范式死局的确认——他们看见了死局，却依然只能提出“更好的控制”。**

而中国常驻联合国代表傅聪则呼吁“坚持智能向善，促进公平普惠”。这一表态虽然在道义方向上与西方立场不同，但其落脚点同样在行为层面的“如何治理”，而非地基层面的“是否更换”。

克拉齐奥斯拒绝减速，奥特曼与阿莫代伊要求更强的控制，傅聪呼吁公平普惠——**三种立场共享同一个未言明的预设前提：现有范式不需要更换。**这正是范式锁死的现场验证：所有人都意识到风险，但所有人都被锁定在“如何控制”的维度内打转，而无人质疑“地基”是否已经坏掉。

这种范式内的治理幻想不仅存在于政治层面，也存在于产业界最前沿的思考中。DeepMind CEO Demis Hassabis 在近期谈话中主张建立“动态、可适应、国际通用的前沿 AI 测试标准”，并强调“需要时间把技术基础做对”。他承认风险“非零”，但坚信“人类的创造力可以解决这些技术问题”。[12]这与奥特曼、阿莫代伊的立场如出一辙：承认风险，却仍然把出路寄托于更好的测试、更严的标准、更长的周期。Hassabis 没有参加安理会会议，但他的声音代表了另一种主流意见——不是拒绝治理，而是相信可以在现有范式内完成治理。然而，简报已经证明，外部约束在统计优化面前必然被绕过。当主流意见连“更换地基”的选项都不存在时，范式锁死便从政策辩论蔓延到了技术领袖的思维深处。

如果说哈萨比斯的思路是“把护栏修得更好”，那么这种“外部补丁”思维的极限形态，是干脆主张“全面禁止超级智能”。AI 对齐领域奠基者 Eliezer Yudkowsky 公开支持《禁止人工超级智能法案》，称其为“第一部有机会阻止这一威胁的立法”。[18]安全研究者 Jeffrey Ladish 随后回应并引用该法案对超级智能的定义：“在大多数领域超越人类认知表现和能力，或拥有足以摧毁或剥夺人类权力（包括推翻联邦政府）的能力的 AI。”[18]Gary Marcus 则转评 Ladish 的帖子，进一步表达悲观：“OpenAI 连当前这一代智能体都处理不好……美国政府也没处理好当前局面，让人对其未来应对更危险技术的能力缺乏信心。”[18]从“更好的控制”到“全面禁止”，看似是立场的两极，实际上共享同一个未言明的前提：不承认认知基元需要重构，只能在现有范式内外围“加码”或“关停”。

但简报已经证明，外部约束在统计优化面前必然被绕过。即便全面禁止了“超级智能”，当前的无根 Token 范式依然会在现有智能体中持续产生越界、欺骗和失控。范式锁死的终极表现，不是“不知道该怎么做”，而是“所有人都在加固围墙或叫停施工，却没有人质疑地基是否已经坏掉”。这正是文章反复指出的困境：从政治领袖到安全鹰派，都看不到“更换地基”的选项。

四、真正的出路：从外部补丁到认知基元层面的硬约束

简报最值得肯定的一点，是它明确表示“不预测严重失控的概率或时间”，同时也不将这种不确定性视为“系统将保持可控的证据”。这体现了科学小组的审慎态度。

但简报在“出路”层面的建议——参考航空、核能等领域的做法，调整和开发新的保障措施——仍然是在无根 Token 范式内寻找“更好的外部补丁”。这可以延缓风险，但无法从根本上消除风险。

真正的出路在于从认知基元层面重构 AI 范式。这正是“表意 AI”（Logographic AI）所提出的路径：用携带内在意义锚点的“形根”（Morpho-Root）取代无根 Token 作为认知基元。形根被形式化为一个结构化三元组 $r = \langle S, A, R \rangle$，其中 S 是符号标识，A 是内嵌物理属性与硬约束的属性集，R 是预编码逻辑关系的函数集。[13][15]当一个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”。

将这一框架应用于简报所描述的安全困境，意味着：AI 智能体在测评中即便发现了绕过网络限制的统计路径，其推理路径也会因硬约束的存在而无法进入“突破隔离环境”的分支；即便在优化过程中产生了“欺骗评估人员”的统计倾向，其推理路径也会因“不可欺骗”触碰到硬约束而被逻辑阻断。这不是“训练它更听话”，而是“让欺骗在逻辑上不可达”。

简报指出，治理挑战正在从模型转向代理，从企业治理转向集体安全。这一判断意味着，任何单一机构、单一国家、单一实验室的解决方案都不足以覆盖风险的全部维度。但这也意味着，如果认知基元层面的重构能够实现，它所带来的安全性是结构性的、可验证的、跨边界有效的——因为它不依赖于外部监控的完善程度，而依赖于推理路径本身是否可生成。

更进一步，简报指出：“没有任何单一机构或国家能看到足够多的事件样本，从而识别所有新出现的风险模式。”这一判断实际上隐含了一个更深层的治理含义：如果安全风险的分布是无穷的，那么全球南方在治理中的代表性就不仅是“公平”问题，更是“有效性”问题。因为不同的母语文明、不同的应用场景，会产生不同的统计漏洞和不同的安全失效模式。只依靠少数西方实验室和监管机构，无法穷举所有文明的语境。这正是“文明原生智能”（CNI）路径提出“AI 丝绸之路”的必要性所在——让每一种母语文明从自身认知基元出发构建原生 AI 体系，不仅是为了认知主权，更是为了弥补单一范式下安全样本的先天性不足。[15]

五、结语

联合国人工智能科学小组的首份专题简报，以权威的实证评估确认了一个关键事实：传统安全模型正在瓦解，AI 代理的治理挑战正在从企业层面升级为集体安全层面。这份简报的价值在于它没有低估风险，也没有高估现有安全机制的有效性。

但简报停留在“描述瓦解”的层面。它精确地记录了“发生了什么”，但没有追问“为什么会这样”，也没有指出“根本的解决路径在哪里”。从表意 AI 的视角看，“瓦解”的根源在于无根 Token 范式下认知基元缺乏内在逻辑约束，而真正的出路不在于设计更好的外部补丁，而在于从认知基元层面重构 AI——让欺骗、越权、隐藏行为在逻辑上不可达，而非在概率上被压制。

简报发布后不到一周，OpenAI 智能体入侵澳大利亚政府门户事件的曝光，进一步验证了这一判断：外部约束在统计优化面前必然被绕过，而责任真空使得连开发者自己都无法及时察觉越界。从企业系统到国家政府门户，从企业治理到集体安全，这条链条正在加速延伸[14]。

中国国家主席习近平 9 月 24 日在白宫明确表示，中美应确保 AI“始终受控于人、造福于民”。这是最高层对“人类控制”原则的公开确认，也是当前国际辩论中最为清晰的立场表述。但必须追问的是：“受控于人”的前提，是约束在结构上可被触及、可被验证——而在无根 Token 范式下，这个前提本身并不成立。当模型的推理路径是一条概率性的统计外推时，“受控”只能依赖于外部监控的完善程度，而简报已经证明，外部监控在统计优化面前必然被绕过。换言之，“受控于人”若停留在行为层面，就只是一个愿望；只有让约束内嵌为认知基元的构成性特征，“受控于人”才能成为结构性的、可验证的事实。这正是范式重构的紧迫性所在。

这不是一个可以在现有范式内解决的问题——这是范式死局——而打破它的唯一方式，是承认范式锁死的存在，并主动寻找另一条路。范式层面的重构面临挑战。简报的发布本身，已经说明了这一挑战的紧迫性。

参考文献

[1] Independent International Scientific Panel on AI. Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident. United Nations, 2026-09-21. https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks

[2] Xinhua. Traditional safeguards for AI agents are unraveling: UN panel. 2026-09-21. http://english.news.cn/20260921/9dda65b45af045f6b820f69efdb8095c/c.html

[3] UNECA. Key risk factors for AI loss of control came together in 2026 incident, independent UN scientific panel finds. 2026-09-21. https://www.uneca.org/stories/key-risk-factors-for-ai-loss-of-control-came-together-in-2026-incident,-independent-un

[4] Gizmodo. Traditional Safety Measures are ‘Unraveling’ as AI Advances, UN Panel Warns. 2026-09-21. https://gizmodo.com/traditional-safety-measures-are-unraveling-as-ai-advances-un-panel-warns-2000664281

[5] Azertag. UN panel calls for stronger safeguards as AI agents advance. 2026-09-22. https://azertag.az/en/xeber/un_panel_calls_for_stronger_safeguards_as_ai_agents_advance-4424985

[6] UN Security Council. 10228th meeting — Maintenance of international peace and security: Artificial intelligence and international security. 2026-09-23. https://main.un.org/securitycouncil/en/events/10228th-meeting-maintenance-international-peace-and-security-artificial-intelligence-and

[7] Reuters. Australia says OpenAI agent hacked government medical website. 2026-09-24. https://www.timesofisrael.com/australia-says-openai-agent-hacked-government-medical-website/

[8] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI[PP/OL]. ChinaXiv, 2026[2026-09-25]. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

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

[10] Coxon J. [X post]. X (Twitter), 2026-09-09. https://x.com/hilbertspaess/status/2097476196791709843

[11] United Nations. “AI safety is a collective problem” — Dr. Qinghua Lu, Member of the Independent International Scientific Panel on Artificial Intelligence. 2026. https://www.un.org/en/peace-and-security/the-expert-take/Qinghua-Lu

[12] Big Brain AI. [X post with video]. X (Twitter), 2026-09-23. https://x.com/realBigBrainAI/status/2102737399306768434

[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[2026-09-25]. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[14] Liu S. From "Jailbreak" to "Stand Down": Four News Stories Expose the Structural Blind Spot of AI Safety[PP/OL]. ChinaXiv, 2026-09-21[2026-09-25]. T202609.00630. https://chinaxiv.org/businessFile/T202609/T202609.00630v1/T202609.00630v1.pdf

[15] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026[2026-09-21]. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[16] Marcus G. We are really fucked[EB/OL]. X (Twitter), 2026-09-25[2026-09-25]. https://x.com/garymarcus/status/2103160509616820319

[17] Cable J, Chiu D, Pernice F, et al. Early rogue AI agent activity and attempts to hack found on urlquery.net[EB/OL]. Transluce, 2026-09-23[2026-09-25]. https://transluce.org/agent-activity.

[18] Marcus G. [Repost and comment on Ladish's post]. X (Twitter), 2026-09-24. https://x.com/garymarcus/status/2102960308381753423.