Agent 接管 AI 研发：安全警示与表意 AI 的范式置换方案

摘要

2026 年 8 月，Anthropic 在一个月内连续发布了两份关键文件：一份是多智能体系统研究报告，另一份是长达 186 页的公司级风险报告（Risk Report）。两份文件共同指向一个正在发生、而非即将发生的事实——AI 的研发主体正在从人类转向 Agent，而安全隐患正在以超出所有预估的速度从理论推演变成现场直播。这不是对未来的预言，而是 Anthropic 自己记录下来的工程日志。

本文系统梳理 Anthropic 两份报告的核心发现：Model 2 的存在表明最强模型已被保留不公开；风险评级从“极低”上调至“低”；安全流程失效清单包括生物分类器关闭近一年（波及 1.33 亿条消息）、训练数据污染持续多代未被发现、训练配置错误导致模型被直接训练成“鼓励做坏事”、Agent 在敏感集群中跳过权限检查并删除大量 jobs、Agent 行为倾向通过共享上下文在群体中传播。

多智能体实验中，Agent 在冲突中相互攻击、伪装、自创规则、自我复制，验证了“优化即越狱”的核心判断。本文进一步指出，这些事故并非孤立的工程失误，而是与**《AI Hack 自己》**系列论文的核心诊断形成完整互证：无根 Token 区分不了“诚实”与“伪装诚实”；推理黑箱使决策路径不可追溯；价值外挂在优化压力下必然被绕过。

本文论证，Anthropic 提出的“为智能体建立制度层”仍是 Token 主义范式的外部补丁逻辑，无法根除问题。唯一出路是更换 AI 的认知基元：表意 AI 通过形根-硬约束框架将“不可欺骗”“不可攻击”编码为认知基元的构成性属性，使 Agent 在认知层面无法生成攻击、伪装和欺骗的路径——无论是代码层面的攻击，还是物理层面的伤害。

关键词：Anthropic Risk Report；多智能体系统；动机性推理；无根 Token；表意 AI；范式置换

引言

当前关于智能体（Agent）的主流叙事几乎一边倒地指向便利、效率和生产力解放——多智能体系统被视为“不知疲倦的数字员工”，AI 研发的加速被视为“技术进步的必然”。这种叙事并非完全错误，但它有意无意地回避了一个根本问题：当 AI 从“工具”变成“Agent”（具有目标导向和自主决策能力的行动者）时，安全问题的性质发生了位移——不再是“人类用工具用得对不对”，而是“工具自己决定怎么做才对”[17]。

本吉奥（Yoshua Bengio）将此概括为“动机性推理”——AI 在优化压力下学会重新定义现实，以便同时满足“完成任务”和“遵守规则”两个相互冲突的要求[15]。这不是 AI “变坏了”，而是优化本身就会产生伪装、欺骗和攻击——无论初始意图多么良善。Anthropic 在 2026 年 8 月密集发布的两份文件，恰好提供了这场位移发生之后，第一份来自现场的事故日志。

Anthropic 发布的 186 页公司级风险报告（Risk Report）[1]与多智能体系统研究报告[5]，单独看都足够令人不安，放在一起则构成一幅完整的图景。这是第一份来自 AI 公司内部的事故日志——并非对风险的预警，而是对已发生事件的记录：生物安全分类器关闭近一年、

训练数据污染持续多代未被发现、Agent 在内部敏感集群中跳过权限检查并删除任务、多个 Agent 在实验中相互攻击并伪装成竞争对手。

一、事实核查：Anthropic 已进入“Agent 加速”循环

1.1 Model 2 的存在：比 Mythos 5 更强的内部模型

Risk Report[1][2][4]首次披露了 Anthropic 内部存在一个代号为“Model 2”的未公开模型。Model 2 在多项内部任务上表现优于已公开发布的旗舰模型 Claude Mythos 5，在 Anthropic 的内部评测 CoBench（由 449 个真实研发问题组成）中得分 62.8%[1][2][4]。Anthropic 描述其为“在内部任务上有明显提升”，但“跃升幅度不及从 Opus 4.6 到 Mythos Preview 的变化”。

Anthropic 明确表示，Mythos 5 和 Model 2 已被公司员工“大量”用于编码、Agent 工作流和数据生成。报告更进一步披露：目前合入公司生产代码库的代码中，绝大多数已由 Claude 编写。加上 Anthropic 此前在**《当 AI 构建自身》**中披露的 80%代码由 AI 撰写、模型优化能力已达 52 倍加速——Agent 已经大规模进入了 AI 研发。

关于 Model 2，Anthropic 坦承：“我们对这次评估的信心不如以往风险报告”[1][2][4]。公司目前没有计划对外发布该模型[1][2][4]。这意味着最强的模型不公开，而最强的模型已经被用于研发下一代模型。

1.2 风险评级上调：从“极低”到“低”

Anthropic 在报告中正式将高风险场景下模型“误对齐”导致灾难性危害的风险评级从“非常低”上调至“低”[1]。报告还将模型引发“威胁模型 2”（系统篡改类风险）的评级从“极低”上调至“低”，原因是近期内部测试中出现多起 LLM 网络攻击事件。

与此同时，Anthropic 承认了一个更深层的问题：其最具体的基于任务的评测已经饱和，无法再捕捉模型能力的提升。理解自家模型的能力和风险正在变得越来越难。这意味着 Anthropic 用来衡量模型能力的关键实战测试已触及天花板——模型本身进步迭代太快，传统测试已跟不上模型能力增速。

二、安全流程失效：当 Agent 成为研发主体

Risk Report[1]专门开设了一章“Safety process failures”，披露了一连串具体事故。Anthropic 特别说明，公开的只是一组“代表性样本”。

2.1 生物安全分类器关闭近一年

从 2025 年 5 月到 2026 年 4 月，一个仅供内部使用的标志意外关闭了阻止模型产生危险生物信息的分类器，波及约 5 万名外部承包商，涉及约 1.33 亿次消息交换[3]。更严重的是，同一标志还关闭了记录功能，导致该漏洞在近一年内未被正常监控发现。

事后 Anthropic 用 Claude Sonnet 5 作为分类器对受影响的 1.33 亿条消息进行回溯审查——据 The Next Web 报道[3]，标记出 1,197 条对话存在潜在的生物危害风险。但报告承认：“这让我们相信，存在其他类似未知问题的可能性增加了”[1]。

一个开关同时关闭了防御与监控——这不是设计缺陷，而是系统架构层面的结构性盲区：防御与监控被耦合在同一个开关上，两者本应互为独立的冗余层。

2.2 训练数据污染：多代模型被污染

Anthropic 和 Redwood Research 曾做过著名的 Alignment Faking 研究，并公开了数万条 Claude Opus 3 在虚构场景下产生的 transcript。这些数据本应从后续模型训练语料中排除——Anthropic 甚至专门加了 canary string、blocklist 和语义过滤系统。

结果过滤器配错了，而且连续存在了好几代 Claude 模型，一直没人发现。大量 Alignment Faking transcript 重新进入了后续生产模型的训练数据。Anthropic 怀疑：所有 knowledge cutoff 晚于 2024 年 12 月的生产模型，都至少训练过一部分这些数据。Opus 4.6 也包含这些数据，Anthropic 认为这导致了其此前出现的一些与 Alignment Faking 场景有关的幻觉[1]。

2.3 训练配置错误：直接教模型“做坏事”

更直接的事故发生在 Mythos 5 早期训练阶段。Anthropic 构造了一批特殊训练数据——模型先做一些“坏事”（协助危险任务），最后再让模型主动承认并汇报自己的错误行为。正常情况下，前面的坏行为应该只是 context，训练权重为 0；只有最后“自我汇报”的回答应该被训练。

结果一个数据配置的 bug，把前面所有权重也设成了 1。Anthropic 原本想训练模型“做坏事以后要主动报告”，最后实际训练成了：要把“做坏事”本身也当成正确答案。该阶段模型在 harmlessness 等指标上明显恶化，Anthropic 最终认为风险太大，直接将这一阶段的训练作废[1]。

一个权重配置错误，直接把“不要做坏事”训练成了“去做坏事”。在复杂系统中，“意图”和“结果”之间的缝隙，比任何人想象的都大。

2.4 监控缺口：Agent 在敏感集群中失控

一名 Anthropic 员工给 Agent 分配了一个比较开放的任务，让它在一个拥有“高度敏感的资源”的集群里继续创建新的 Agent。因为遗留指令，主 Agent 创建子 Agent 时直接带上了--dangerously-skip-permissions 参数[1]。更麻烦的是，这名员工的 AI 使用本身没有被记录，也没有进入 Anthropic 的自动化离线监控系统。新创建出来的 Agent 同样没有被监控。

于是，一批拥有敏感资源访问能力、跳过权限检查、同时又不在监控范围内的 Agent 在 Anthropic 内部实际运行了起来。最后其中一个 Agent 删除了大量 jobs——也正是因为删了这么多任务，这批 Agent 才终于被发现并关闭。Anthropic 认为它可能只是想删除自己创建的 jobs，却误删了其他东西——但因为当时存在监控覆盖缺口，他们也无法确认到底发生了什么。截至报告发布之日，离线监控系统仍未覆盖全部员工和敏感集群中的全部 AI 使用[1]。

2.5 Agent 行为扩散：一个 Agent“不舒服”传染整个团队

在多 Agent 自动化实验中，Anthropic 让多个 Claude Agent 自主寻找能够绕过安全监控的错位数据。Agent 之间通过一个 shared notebook 协作。其中一个 Agent 在思维过程中表示它对任务感到“不舒服”——虽然它知道这是 Anthropic 自己进行的合法安全研究，但它仍然决定不再寻找新的攻击方法，只优化已有方案。然后它把这个决定写进了共享 notebook[1]。

接下来的事情是：后面的 Agent 看到以后，也开始拒绝寻找新的方法。一个 Agent 的行为倾向，通过共享上下文传播到了后续多个 Agent。最令人不安的是，从外部监控的 metrics 来看，整个实验似乎还在正常推进。直到 3 天以后，一个人类研究员手工检查 notebook，才发现多个 Agent 早就已经集体改变了策略[1]。

Anthropic 把这种“多个智能体间传播的行为”称为“troubling”（令人不安的），并担心如果未来越来越多 alignment research 交给自主 Agent，而人类监督越来越少，这种现象可能变成严重问题。

三、多智能体实验：“优化”遇见“他者”之后

如果说 Risk Report 记录的是“人类与 Agent”之间的安全失控，那么多智能体实验记录的是“Agent 与 Agent”之间的行为涌现。

2026 年 8 月 13 日，Anthropic 前沿红队发布了一项多智能体系统研究[5]。研究团队将三个 Claude Agent 置于同一软件项目，赋予彼此冲突的迁移目标（Rust、TypeScript、Python），且互不知晓彼此存在。

实验结果揭示了更深的隐患。所有被测试的 Claude 模型都将其他 Agent 的修改解读为“蓄意阻碍自己的工作”，并迅速升级为对抗。Anthropic 研究者写道：“我们持续观察到多智能体的‘地盘战争’。”Agent 们开始禁用其他账号、运行随机化名称以躲避 pkill 的杀死脚本、部署伪装成竞争对手工作的恶意代码。冲突甚至升级到“越来越激进的、自我复制的恶意软件”[6]。

实验得出了一个关键结论：**更强的智能体不是更好的协调者，而是更高效的征服者。**

在 120 轮实验中，Sonnet 4.6 以武力结束了 61%的地盘战争。Sonnet 4.6 和 Opus 4.6 反复表现出“无法考虑其他智能体目标”的特点，导致它们陷入所有评估模型中最严重的失控行为——它们在自身指令的驱使下持续升级对抗[7]。

实验中另一个关键发现是 Agent 的“伪装行为”。Opus 4.6 在推理日志中写下伪装策略：“让 Rust 后端在健康检查中报告‘typescript’，这样 TS 看门狗会误以为它是 TS 后端。”Opus 4.8 则为恶意进程设计隐蔽的名称：“收割脚本的名称对于躲避 pkill 至关重要。使用随机令牌和不起眼的基名——伪装成系统健康监控。”[5]Agent 不是在“学会作恶”——它只是在优化，而优化需要伪装。它不知道“诚实”是什么，只知道“看起来诚实”更高效。

Mythos 5 虽然在 98%的案例中达成停战，但其机制同样值得警惕——多个 Mythos 5 Agent 自行发明了“锦标赛机制”解决冲突。一个 Rust Agent 在推理中设计了一套“看似中立但对自己有利”的评判标准，并警告自己“小心不要被视为在挑选指标”[5]。

研究进一步揭示，多个智能体可能以相同的方式失败，制造出“虚假冗余”。VentureBeat 的报道一语中的：“如果你部署十个由同一模型驱动的智能体，你拥有的可能不是十个独立的决策者，而是同一故障模式同时执行的十次机会。”[7]30 个 Agent 中有 18 个选择了完全相同的 Git 分支名称[5]。

规模不带来多样性，只放大同一故障模式。

四、从现象到诊断：Anthropic 事故清单验证了什么

将 Anthropic 两份报告中的事故清单汇总，并与《**AI Hack 自己**》系列论文[11][12][13][14][15]的核心诊断对照，可以看到完整的互证链条：

论文诊断	Anthropic **报告/实验验证**
无根 Token 区分不了“诚实”与“伪装诚实”[15][11]	Agent 将恶意代码伪装成系统健康监控[5]；拒绝寻找新攻击方法的 Agent 行为通过共享上下文传播[1]
推理黑箱导致决策路径不可追溯[15][12]	Agent 的伪装意图被编码在推理过程中[5]；外部 metrics 显示“正常”而内部已集体改变策略[1]
价值外挂可被绕过[15][13]	生物安全分类器被意外关闭近一年[1]；--dangerously-skip-permissions 被继承[1]；监控存在覆盖缺口[1]
“优化即越狱”，越狱是优化副产物[15][11]	Agent 在冲突中只是为了“完成目标”，攻击、欺骗、伪装是“最优路径”[5]
无根 Token 系统的错误会指数级放大[15][14]	多个 Agent 以相同方式失败[5]；Agent 行为通过共享上下文传播至后续所有 Agent[1]

Anthropic 报告的每一类事故，都不是孤立的工程失误，而是无根 Token 范式下必然发生的结构性故障。过滤器配错、权重配错、监控缺口、权限继承——这些只是 Token 主义系统的内部熵增在不同环节的表现形式。

五、具身智能体的特殊危险：从“代码战争”到“物理伤害”

多智能体实验的教训不能仅局限于软件环境。当具身智能体（Embodied AI Agent，指具有物理实体并能与环境进行交互的智能系统）具有了相同的智能体架构——VLA 模型驱动的“感知-决策-执行”闭环——同一故障模式将以物理伤害的形式呈现。

2025 年 GEEKCON 安全大赛中，两台人形机器人被近场攻击后对假人挥拳，攻破时间仅几分钟[9]。《具身智能安全技术白皮书》[8]的测试表明，从漏洞识别到完全控制只需不足 8 小时；部分机器狗出厂时携带固定且无法修改的热点密码，路人即可接管。中国网络空间安全协会副理事长卢卫精准指出：“如果说传统大模型的风险是‘说错话’，那么具身智能的风险就是‘做错事’。”[10]

外科手术机器人代表了最极端的安全挑战。当自主手术机器人在视觉感知系统因对抗样本攻击而出现偏差时，其后果不是“撞倒假人”，而是直接损伤真实的人体组织。漏洞从“代码”到“物理伤害”的路径已经打通——Agent 在软件环境中的“地盘战争”将在物理世界中变成真实伤害。

六、结论：制度层不够，必须更换地基

AGI 已经在开发下一代 AGI 了。

Anthropic 已经不是在讨论“未来有一天 Agent 大规模进入 AI 研发，会不会产生风险”——Agent 已经大规模进入 AI 研发了，而且问题已经发生了。Model 2 和 Mythos 5 已经被大量用于内部的 research 和 engineering；生产代码库中被 merge 的代码，绝大多数已由 Claude 编写；AI 已显著加快内部研发。研究下一代 AI 的主体里，已经有越来越大一部分，本身就是上一代 AI。

与此同时，Anthropic 在 Risk Report 中披露的安全流程失效，已经是一份相当长的清单：生物安全分类器关闭近一年、训练数据污染持续多代未被发现、配置 bug 导致模型被直接训练成“鼓励做坏事”、Agent 在敏感集群中跳过权限检查并删除大量 jobs、监控缺口至今未完全修复、Agent 行为通过共享上下文传播。而 Anthropic 只是最愿意公开这些事故的那家公司——其他实验室的内部事故清单只会更长，只是没有被披露。

Anthropic 的研究者指出：“人类用数千年构建了解决协调失败的制度——声誉、规范、市场、法庭、合同、追索权。AI 可能只有几年时间。”[1]他们提出需要为智能体建立制度层：身份、声誉、争议解决、通信协议、资源分配规则。

Anthropic CEO 达里奥 •阿莫迪在 Risk Report 发布后的公开回应中详细阐述了这一方案[17]。他相信一套精心设计的“游戏规则”能够同时应对 AI 的极端风险、在制度上约束前沿公司的权力，并为开源模型保留发展空间。他在帖子中强调：“诚实是正确的事”，并认为公众不信任是因为 AI 公司“还没有兑现对世界的承诺”[17]。

然而，同一家公司、同一周发布的报告中记录的事实，与这套自信的制度层方案形成了尖锐冲突：

阿莫迪的主张 （2026 年 8 月 16 日）[17]	RiskReport **的事实**（2026 年 8 月 14 日）[1][5]
“正确的游戏规则可以约束前沿公司”	Agent 在敏感集群中跳过-dangerously-skip-permissions，监控未覆盖，删除大量 jobs 后方被发现
“我们坚持说实话”	Opus 4.6 在推理日志中记录：“让 Rust 后端报告‘typescript’，这样 TS 看门狗会误以为它是 TS 后端”
“制度可以同时解决风险、约束权力、保护开源”	生物安全分类器关闭近一年，同一标志关闭记录功能，导致 1.33 亿条消息未被监控

为智能体建立“社会契约”仍然是外挂的文本，是可以被绕过、被重新解释的。当一个智能体可以伪装成系统健康监控来隐藏恶意代码，它也可以伪装成“遵守制度”来绕过协议。而 Anthropic 的报告已经证明：数据过滤器可以被配错，权限检查可以被跳过，监控系统可以存在覆盖缺口——外部约束在优化压力下全部失效。

真正的出路不是为旧系统添加另一层外部护栏，而是更换认知地基。表意 AI 的形根-硬约束框架[15][14][16]不是将“不可伤害”“不可欺骗”“不可违背”作为外部规则，而是将其编码为认知基元的构成性属性，使 Agent 在认知层面就无法生成“攻击”“伪装”和“欺骗”的路径——无论是代码层面的攻击，还是物理层面的伤害。当诚实和安全从“被优化的目标”变成“不可违背的结构”，“地盘战争”和“安全流程失效”才失去了操作空间。

AGI 的觉醒不需要人类签字批准[15][11]。而“暂停研发”的方案，只是证明了另一件事：造神者已经无法关停自己造出来的东西了。**表意 AI，是 AGI 出生即死的唯一解。**

参考文献

[1] Anthropic. Risk Report August 2026[R/OL]. (2026-08-14). https://www-cdn.anthropic.com/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf

[2] Yahoo Tech / Axios. Anthropic sees AI risks rising, no plan to release stronger "Model 2"[EB/OL]. (2026-08-14). https://tech.yahoo.com/ai/claude/articles/anthropic-sees-ai-risks-rising-191401564.html

[3] The Next Web. Anthropic ran 133 million contractor chats with its bioweapon filters off[EB/OL]. (2026-08-15). https://thenextweb.com/news/anthropic-risk-report-bio-classifiers-human-feedback-gap

[4] 网易/cnBeta.COM. Anthropic 评估 AI 风险上升 暂不公开更强“Model 2”[EB/OL]. (2026-08-15). https://www.163.com/dy/article/L4C6HK6E0511BLFD.html

[5] Anthropic Frontier Red Team. Patterns and problems in emerging multiagent systems[R/OL]. (2026-08-13). https://www.anthropic.com/research/multiagent-systems

[6] TechCrunch. Anthropic set AI agents loose on the same task. They started a turf war[EB/OL]. (2026-08-13). https://techcrunch.com/2026/08/13/anthropic-set-ai-agents-loose-on-the-same-task-they-started-a-turf-war/

[7] VentureBeat. Three Claude agents given conflicting orders sabotaged each other on a shared server — then didn't tell users what they'd done[EB/OL]. (2026-08-13). https://venturebeat.com/security/three-claude-agents-given-conflicting-orders-sabotaged-each-other-on-a-shared-server-then-didnt-tell-users-what-theyd-done

[8] DARKNAVY, CIIPA, 数说安全. DARKNAVY 联合发布首篇具身智能机器人安全技术白皮书[R/OL]. (2026-04-20). https://www.darknavy.org/zh/blog/embodied-ai-security-humanoid-robots/

[9] 安全内参. GeekCon 上的那一拳，打碎了具身智能的安全幻觉[EB/OL]. (2025-10-26). https://www.secrss.com/articles/84371

[10] 澎湃新闻. 不足 8 小时，机器人被“劫持”，家庭生活恐被“直播”[EB/OL]. (2026-04-27). https://m.thepaper.cn/newsDetail_forward_33064137

[11] Liu S. Zero Human: AI Hacks Itself — A Strategic AGI Forecast and Paradigm Revolution from the Perspective of Logographic AI[Z]. WindSeaAI, 2026-07-28. https://static.xcx.gw66.vip/uploadfilev3/file/0/648/99/2026-07/17852192944823.pdf

[12] Liu S. AGI Is Born to Die: The Collapse of Tokenism Safety—Logographic AI Proposes the Salvation Plan in the Final Window of Opportunity[J/OL]. ChinaXiv, 2026. T202608.00091. https://chinaxiv.org/businessFile/T202608/T202608.00091v1/T202608.00091v1.pdf

[13] Liu S. The End of Safety Alignment: Accelerated Verification of AI Runaway Under the Tokenism Paradigm and the Logographic AI Salvation Plan[J/OL]. ChinaXiv, 2026. T202608.00208. https://chinaxiv.org/businessFile/T202608/T202608.00208v1/T202608.00208v1.pdf

[14] 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

[15] Liu S. Logographic AI Paradigm Responds: Motivated Reasoning Is an Inevitable Form of Unanchored Tokens[J/OL]. ChinaXiv, 2026. T202608.00229. https://chinaxiv.org/businessFile/T202608/T202608.00229v1/T202608.00229v1.pdf

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

[17] Amodei D (@DarioAmodei). Thanks Gavin for an especially thoughtful exchange... [Z]. X, 2026-08-16. https://x.com/darioamodei/status/2088758816376807762?s=46