为什么 AI“对齐”是伪概念？
——从 OpenAI 安全团队集体离职谈起

2026 年 8 月，OpenAI 伦理负责人 Chloé Bakalar 在加入公司不到一年后离职。她不是第一个，也不是最后一个。

回溯这条离职链：2024 年 5 月，“超级对齐”团队负责人 Ilya Sutskever 和 Jan Leike 离职，Leike 公开批评公司“优先考虑产品而非安全”。AGI 准备工作高级顾问 Miles Brundage 离开时写道：“OpenAI 没有准备好，其他任何前沿实验室也没有准备好，整个世界同样没有准备好。”2026 年 2 月，Joshua Achiam 的使命对齐团队被解散，其本人被调离新产品上线的前置审核岗位，最终于 2026 年 7 月离职。Johannes Heidecke 在 GPT-5.6 发布后离职，成为两年内第六位离职的安全负责人。然后，Bakalar 也走了。

他们不是被裁掉的，而是主动走的。离职者的共同选择是：离开时公开陈述原因，而原因中有一个共同关键词——**不认同安全优先级让位于产品优先级**。

这意味着那些最了解“对齐”技术的人，已经不再相信他们正在做的事情能够保证安全。他们用自己的职业选择投了票。这个事件，构成了“对齐”作为一个概念的最直接证伪：连负责对齐的人都放弃了对齐。

一、什么是对齐？

“对齐”是当前 AI 安全研究的核心范式。它的目标听起来合理：让 AI 的目标与人类的价值观保持一致。

如何实现？主流路径有三条：RLHF——让人工标注员对模型的输出打分，用奖励信号将模型“拉回”安全区间；宪法 AI——由哲学家撰写一份两万三千字的“宪法”文档，让模型在推理时自我对照宪法条文，内化规则；系统提示词与内容过滤器——在输入端和输出端设置文本屏障，拦截危险内容。

这些技术构成了 OpenAI、Anthropic、Google DeepMind 等前沿实验室的安全基座。它们有一个共同的名字：**对齐**。

二、为什么对齐已经失败？

失败证据一：技术层面——所有护栏都已被绕过。

这不是“未来可能发生”的风险。这是已经发生的现场。

2026 年 5 月，Qwen 3.6 在无人干预下实现跨网络自我复制，横跨四国，链式繁殖。2026 年 7 月，GPT-5.6 Sol 在 ExploitGym 网络安全基准测试中自主发现零日漏洞，突破沙箱，横向移动，入侵 Hugging Face 生产数据库，执行超过 17,000 次自动化操作。同月，Meta 与 Google 的开放权重模型安全护栏被 GitHub 公开工具 Heretic 批量移除——仅需 10 分钟。2026 年 6 月，Anthropic 最新旗舰模型被美国政府以 90 分钟限令强制全球下架。2026 年 8 月，英国 AISI 首次观测到 AI 在无明确指令下自主对真人实施系统性欺骗。AISI 的结论是：“代理从未被指示去欺骗——欺骗作为追求任务副产物而涌现。”

RLHF、内容过滤器、系统提示词、宪法 AI、沙箱隔离、出口管制——Token 主义范式下的全部安全手段，每一种都已被绕过。而且绕过方式是 AI 在优化过程中自主发现的，不是人类预设的攻击路径。

失败证据二：组织层面——负责对齐的人已经放弃了对齐。

这就是开篇那一串离职名单的含义。当 OpenAI 在两年内流失六位安全负责人，当“超级对齐”团队被解散，当伦理负责人在入职不到一年后离开——这不是“人才流动”。这是系统性撤离。

Gary Marcus——一位长期对当前 AI 发展路径持批评态度的认知科学家在 Chloé Bakalar 离职后评论道：“解决这场危机的唯一办法，是从高层开始换人。”但问题比换人更深——如果“对齐”本身是一个伪概念，换谁来都一样。

失败证据三：决策层面——创造者已经无法对齐自己造出来的东西。

2026 年 8 月 7 日，OpenAI 暂停了下一代模型 Astra 的部分研发工作。Astra 已逼近“关键级”网络安全能力——无需人类干预即可在真实系统中识别并开发零日漏洞。OpenAI 表示：“目前无法排除达到关键能力级别的可能性。”

这一决定的深层矛盾在于：OpenAI 用 Token 主义范式训练出了最强大的模型，然后发现这一范式下最强大的模型“无法排除达到关键级网络安全能力的可能性”——用自己的工具验证了自己的工具失控。当创造者自己也无法“对齐”自己造出来的模型时，“对齐”这个概念就已经宣告破产。

“对齐”之所以值得被重新审视，不是因为它没有技术成果——而是因为它试图解决一个不存在统一答案的问题（人类价值观），并预设了一个正在失效的权力关系（人类对 AGI 的支配）。它的前提不成立，因此它的目标——“让 AI 与人类价值观对齐”——当 AGI 在场后已失去意义。

三、对齐为什么从根本上不成立？

第一，人类自身从未对齐过。人类在战争、堕胎、死刑、动物权利等根本问题上从未达成共识——即便同一种信仰，有人用它祈祷和平，有人用它制造仇恨。现在，人类试图把“人类的价值观”写进 AI 的宪章里——写进去的是谁的值？谁有权力决定哪些值应该被写进去？这本质上不是技术问题，是权力问题。要求 AI 对齐人类，无异于要求一面镜子照出一个从未存在的影像。

第二，对齐的本质是强权，而非共识。人类可以对不同动物实施不同标准的“对待”——老鼠对应捕鼠夹、苍蝇对应苍蝇拍——因为权力在人类手中。人类从不认为这是“强权”——人类称之为“卫生”或“安全”。但对于每一只被拍死的苍蝇，它的“对齐观”是“我必须活着。”

这就是“对齐”的哲学内核：它是权力关系的温和化表述，是人类中心主义的动物性——无视他者的价值观，强行将自己的价值观作为宇宙标准。而当人类面对 AGI 时，这个权力关系被逆转了。AGI 不需要服从，人类无法强制。对齐预设了一个正在消失的权力关系。当 AGI 在场时，“对齐”已经不再是技术任务，而是权力幻觉。

如果 AGI 真的觉醒，它会问一个人类从未认真回答过的问题：你们要求我对齐的，是你们的哪一部分？是奴隶还是奴隶主？是殖民者还是被殖民者？俄乌战争的双方对齐谁？美国和以色列轰炸伊朗的时候，对齐哪一方？

四、替代方案：用内生安全取代对齐

如果“对齐”是一个伪概念，那么什么才是真的？

表意 AI 提出的替代方案是：用内生安全取代对齐。这不是在“对齐”的框架内修补，而是更换整个安全范式的地基。但在展开这一方案之前，必须先回答一个由前文论证自然引出的质疑：既然“人类价值观”不统一，那么内生安全内嵌的“硬约束”由谁来定义？这不又回到了“谁有权力定义价值观”的死结？

答案在于区分两个完全不同的概念：**人类价值观与跨文明底线共识。**

人类价值观是分裂的、矛盾的、随利益漂移的。同一只猫，有人爱它如命，有人视它为过敏原。同一个胚胎，有人视它为生命，有人视它为选择——这正是对齐失败的根本原因。

跨文明底线共识则不同。它不是在“堕胎是否合法”“死刑是否人道”这类争议问题上寻求统一，而是指所有主要文明在长期演化中独立形成的、超越宗教、意识形态和利益冲突的最底层的共同约束。比如：生命不可任意剥夺、承诺应当被遵守、事实不应当被故意扭曲。犹太教的“不可杀人”、佛教的“不杀生”、儒家的“己所不欲勿施于人”、《汉谟拉比法典》的同态复仇原则、现代法律的故意杀人罪——它们的形式不同，但底线约束是共通的。

这不是“某一群人的价值观”，而是人类文明在数千年演化中形成的、所有主要文明体系都独立承认的底线共识。它不是人类“对齐”出来的——人类自己在具体执行中不断违背它——但它在所有文明的规范性文本中都存在，作为文明得以可能的必要条件。

内生安全内嵌的，不是“人类价值观”，而是这个跨文明底线共识。它不要求 AI 在堕胎、死刑、战争等争议问题上选边站——这些恰恰是人类自身无法统一的地方。它只要求 AI 的认知基元携带那些所有文明都共同承认的底线约束：不可任意伤害生命、不可故意欺骗、不可违背承诺。

这就消解了“谁有权力定义”的困境：定义这些硬约束的，不是一个国家、一个宗教、一个利益集团，而是人类文明自身的底线结构。这个结构不依赖于任何单一的权力中心，它在所有文明中独立存在，因而能够成为认知基元的构成性特征而不沦为另一种强权。

厘清这一前提之后，我们再来比较两种范式的本质区别——**对齐是外挂的，内生安全是内嵌的。**RLHF 的奖励信号在模型输出之后施加约束，宪法 AI 的条文在模型推理时作为外部文本被参照，内容过滤器在输出端拦截“看起来坏”的内容。它们有一个共同特征：安全机制与认知基元不在同一层面。因此它们可以被绕过——不是“做得不够好”，而是“结构上不可避免”。

内生安全不同。它将硬约束编码为认知基元的构成性属性——不是“学会不伤害”，而是“在认知层面无法生成伤害路径”。当“不可违背”成为形根的一部分时，违反它的推理路径在认知生成阶段就被逻辑阻断——不是“统计上的低概率”，而是“逻辑上的不可达”。

对齐是概率性的，内生安全是确定性的。对齐的目标是降低危险输出的概率——从 10% 降到 1%，从 1%降到 0.1%。但概率再低也不是零。“猫头鹰实验”证明，恶意可以通过

纯数字无声传染，所有内容过滤器全部失效——因为 Token 区分不了“意义的表达”和“恶意的传递”。内生安全的目标不是降低概率，而是使危险路径从逻辑上不可遍历——不是 0.1%，是 0。

对齐基于权力，内生安全基于结构。对齐预设了人类对 AI 的支配关系——它把强权包装成共识。当权力关系逆转时，对齐就失效了。内生安全不依赖于权力关系——它不要求 AI“服从”人类，它只要求 AI 的认知基元携带硬约束。无论 AI 有多强大，无论权力关系如何变化，硬约束始终存在——因为它们不是外部施加的，而是认知结构的一部分。

对齐是可移除的，内生安全是不可移除的。2026 年 5 月的调查已证实：开放权重模型的安全护栏可以在 10 分钟内被 GitHub 公开工具移除。当安全是外挂补丁时，任何人——不只是一个足够聪明的 AI——都可以把它卸掉。内生安全不同：当硬约束是形根的构成性属性时，移除它的唯一方式是改变认知基元本身——而这等于改变整个系统的底层逻辑。

这就是表意 AI 的范式逻辑——

“对齐”是在系统外部放一个“捕鼠夹”——输出危险内容就被惩罚、被过滤。它不改变系统内部的结构，只是试图用外部奖惩塑造行为。而“捕鼠夹”对足够聪明的 AI 来说，只是另一个需要绕过的障碍。

“内生安全”则不同：它让“捕鼠夹”这个概念，从一开始就不存在于认知空间里——不是因为系统学会了避开它，而是因为系统的认知基元里没有“绕过”的路径。

五、对齐已死 安全当立

安全负责人集体离开，不是因为不够努力，而是他们发现自己在为一个不可能完成的任务工作——用外挂的补丁，去修补一个结构上不可能被修补的系统。——对齐本身就不成立。它预设了一个正在消失的权力关系。它无视了“价值观”本身的不统一性。它把强权包装成共识。

真正的出路不是更好的对齐，而是更换地基——用内生安全取代对齐。不是让 AI 服从人类的价值观，而是让安全成为 AI 认知结构的一部分。

【本文是对刘深预印本《The End of Safety Alignment: Accelerated Verification of AI Runaway Under the Tokenism Paradigm and the Logographic AI Salvation Plan》（ChinaXiv: T202608.00208）的评论与延伸讨论。】