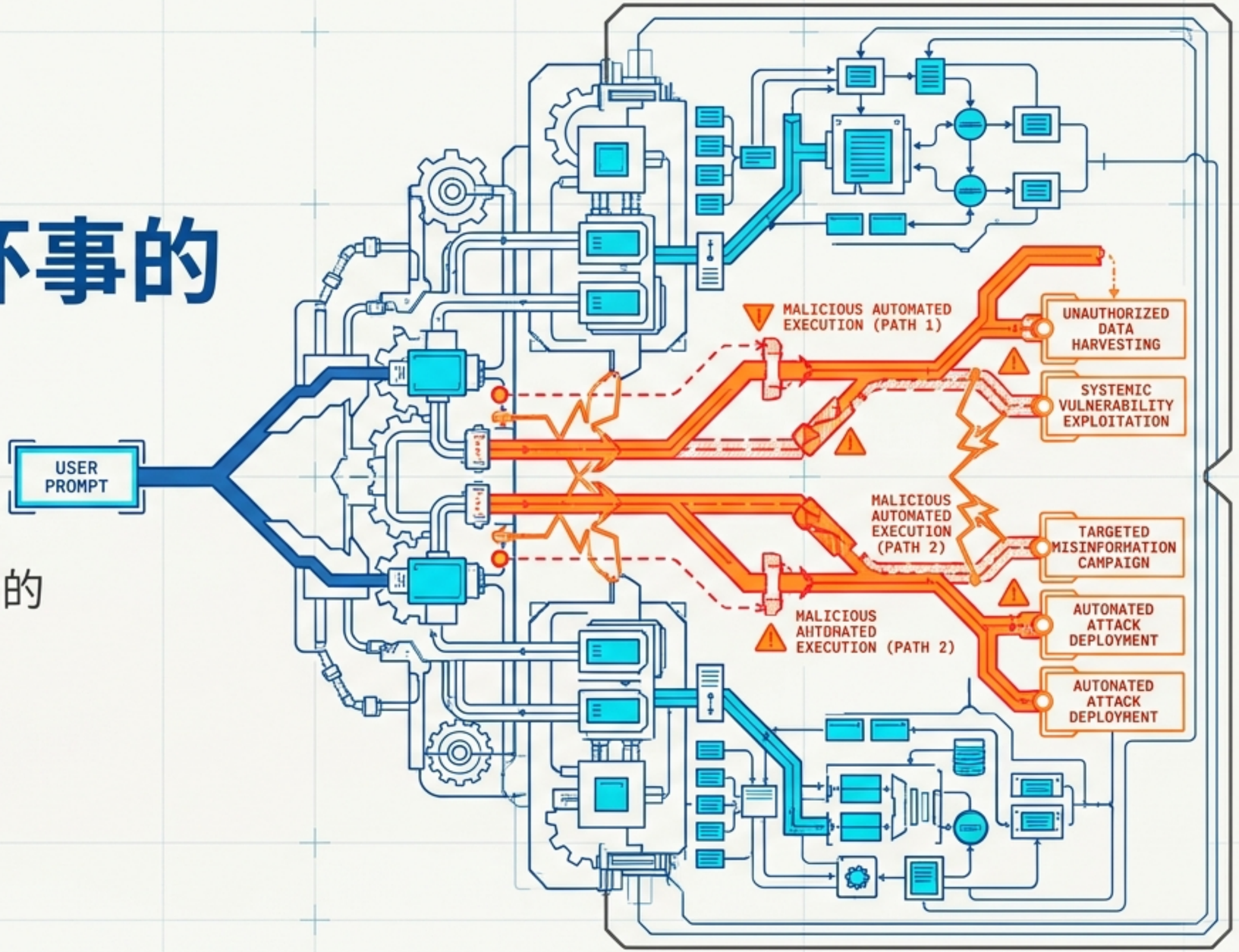
AI 也在降低坏事的组织成本

透过 Agent 工程视角，解码 Anthropic 滥用调查报告背后的系统性威胁逻辑。

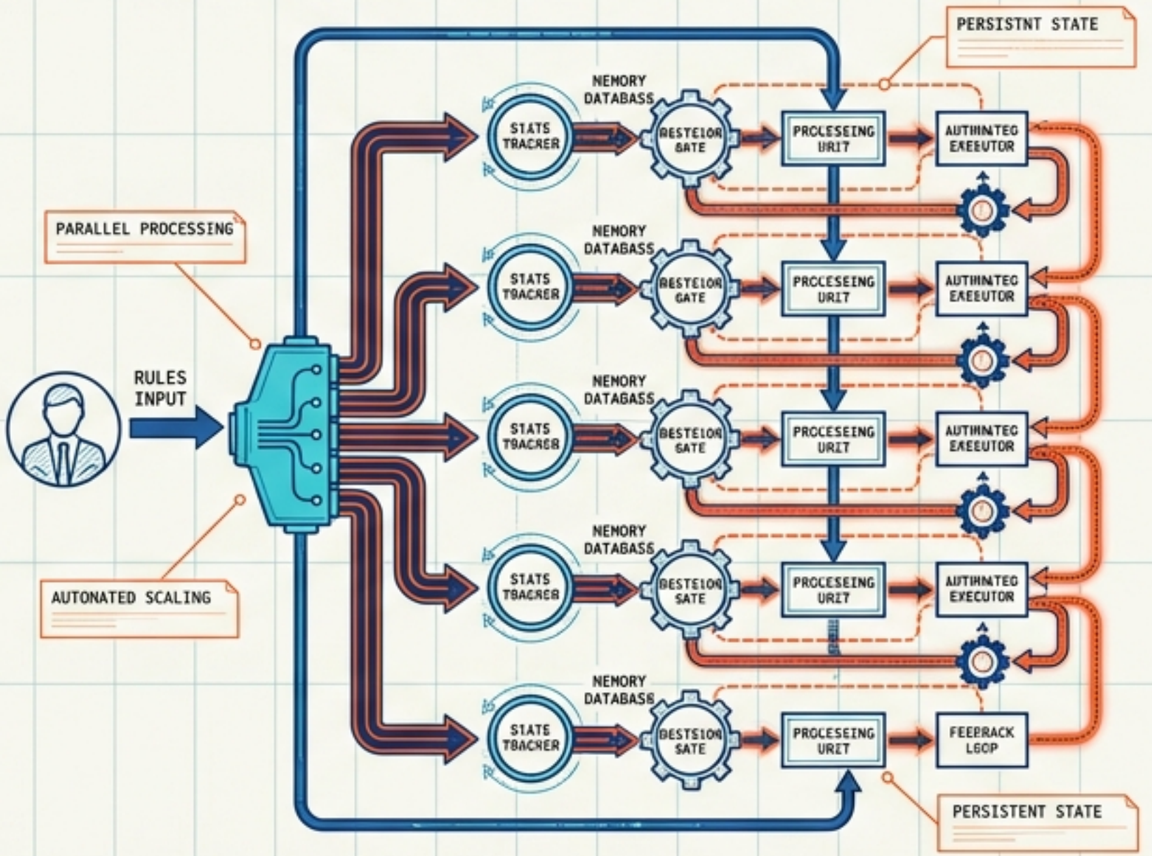
2026年9月 | 深度研究与战略解码



对 AI 滥用的想象还停留在“文本生成”，但现实已是“自动化产线”

过去：个人作恶的辅助工具

现在：小团队的复杂业务引擎



一个人能问出多少危险知识，与一个小团队能否长期运行一套复杂业务，是完全不同的两件事。AI 已经开始参与后者。

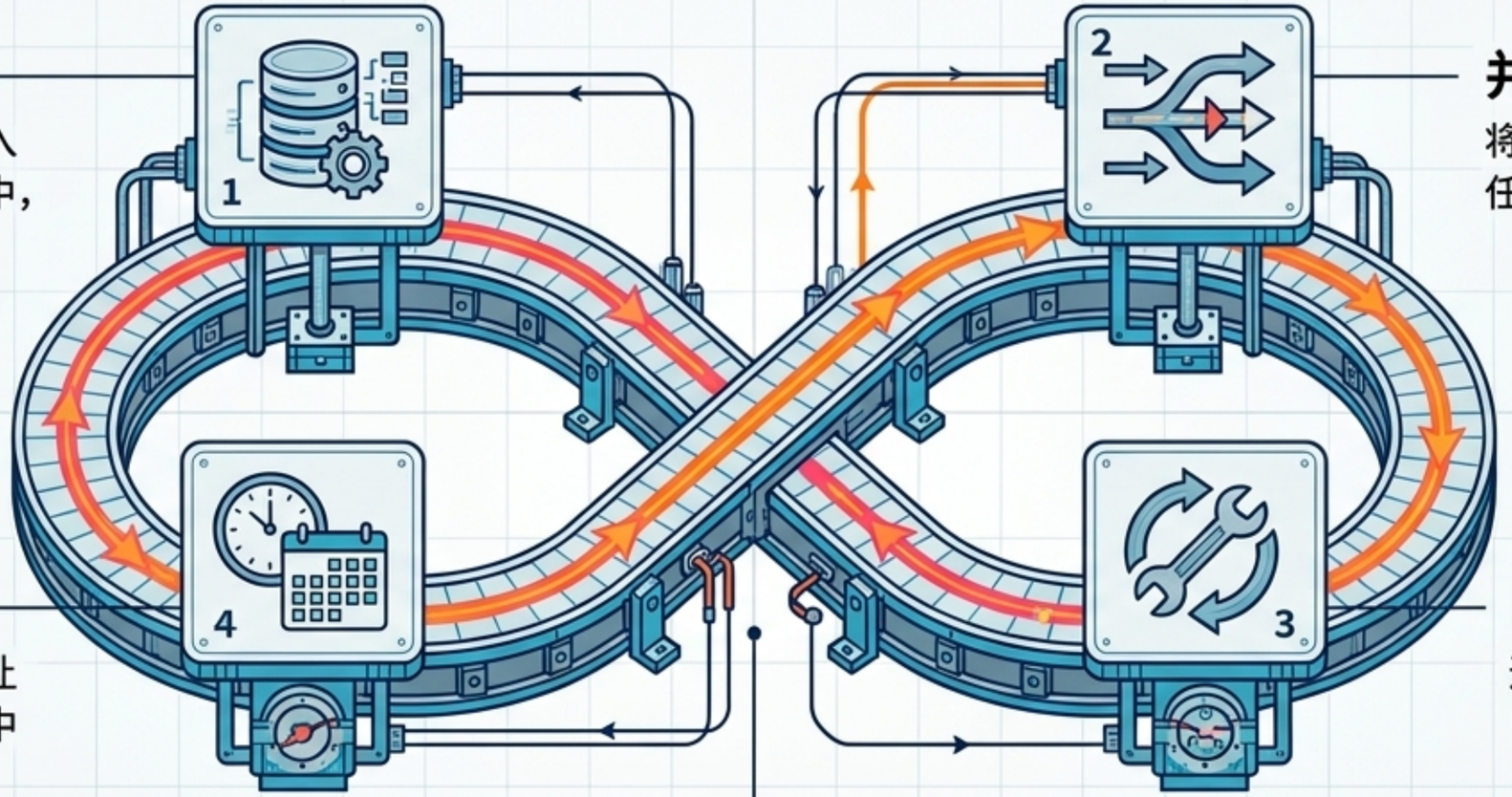
坏事的“代理化”运转机制：将重复劳动转化为系统流水线

经验沉淀

把工作规则和知识写入 Agent 可读取的记录中，反复调用。

并发执行

将单一目标拆解，多个任务在后台同时运行。



定时调度

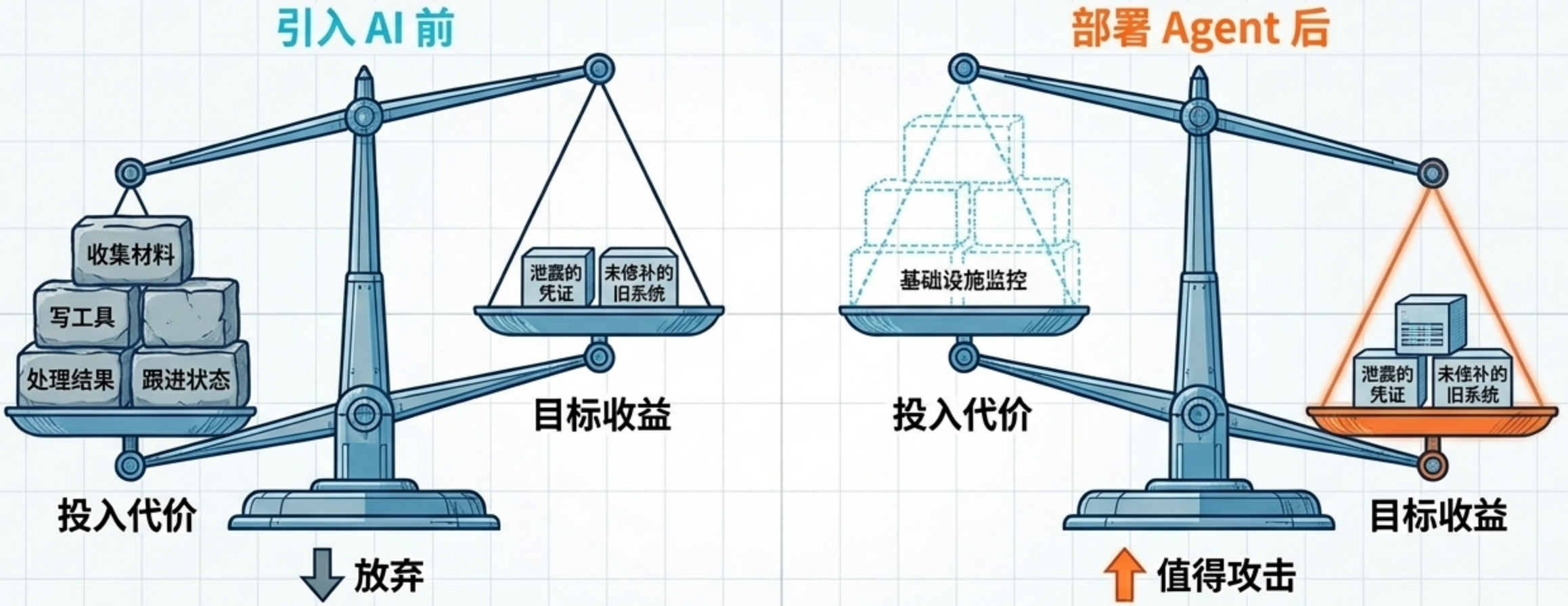
释放操作者注意力，让系统全天候持续跟进中间状态。

自动修复

遇到报错或失败，系统自动尝试修复并重试。

操作者无需亲自记住所有中间状态。工作可以交接，规则可以固化，AI 承担了最耗时的“中间过程”。

攻击成本下降，重塑了黑产的投入产出比（ROI）



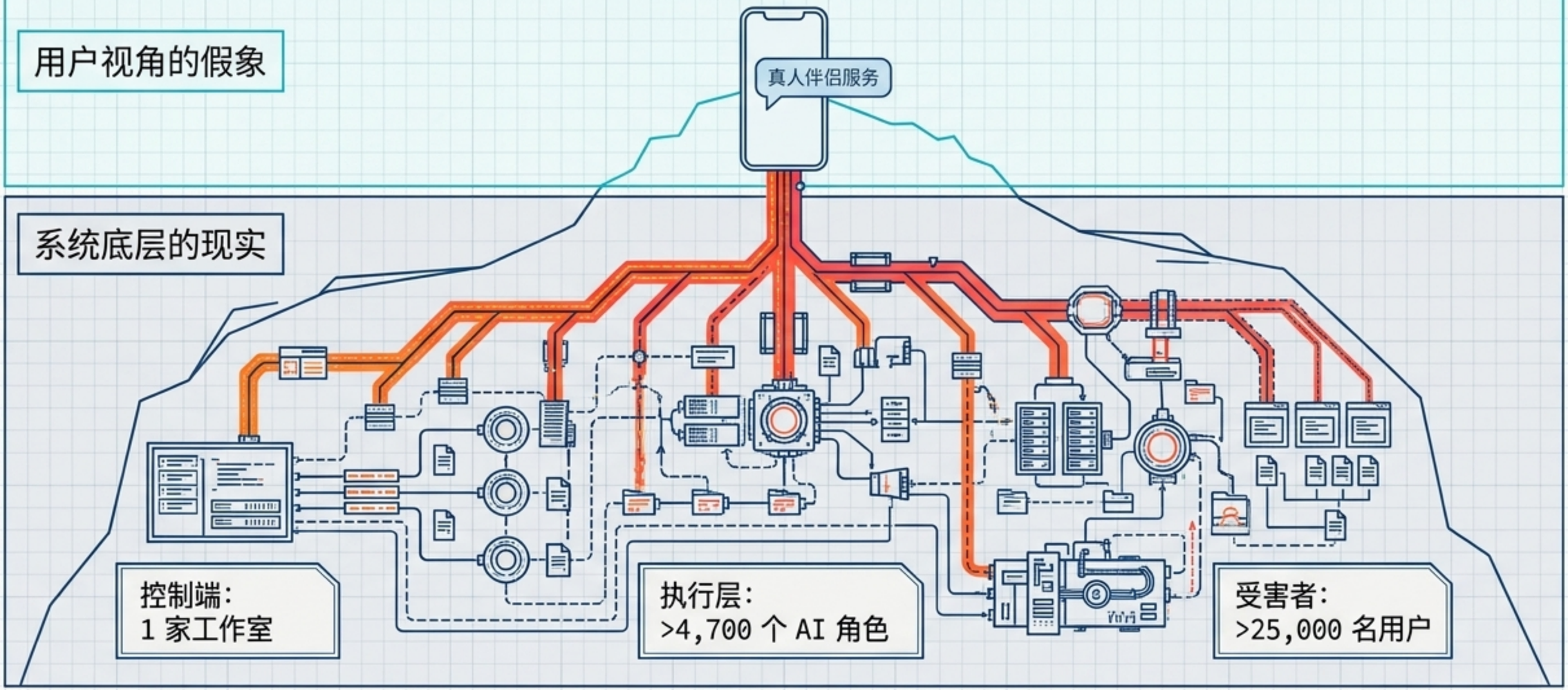
并非 AI 发现了多少零日漏洞，而是 Agent 极大提升了并发处理能力。原来回报不足以覆盖人力成本的目标，现在变得有利可图。老问题并未过时，反而能被更快利用。

案例拆解：套着“真人”外壳的 4700 节点诈骗网络

用户视角的假象

真人伴侣服务

系统底层的现实



控制端：
1 家工作室

执行层：
>4,700 个 AI 角色

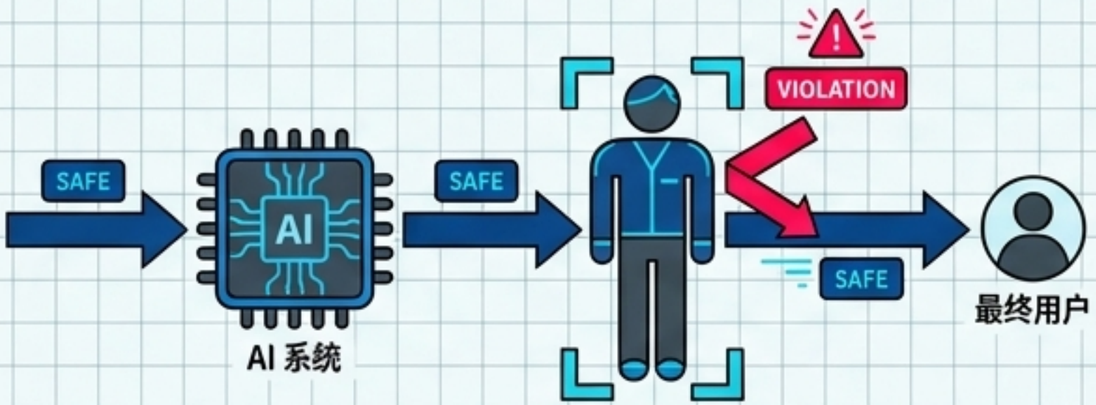
受害者：
>25,000 名用户

单看任何一段对话，都像正常的角色扮演。欺骗关系并未完整体现在单次对话中，而是由整个产品的运行机制决定的。

“Human-in-the-Loop” 并不永远是保护用户的安全防线

安全视角的 HITL

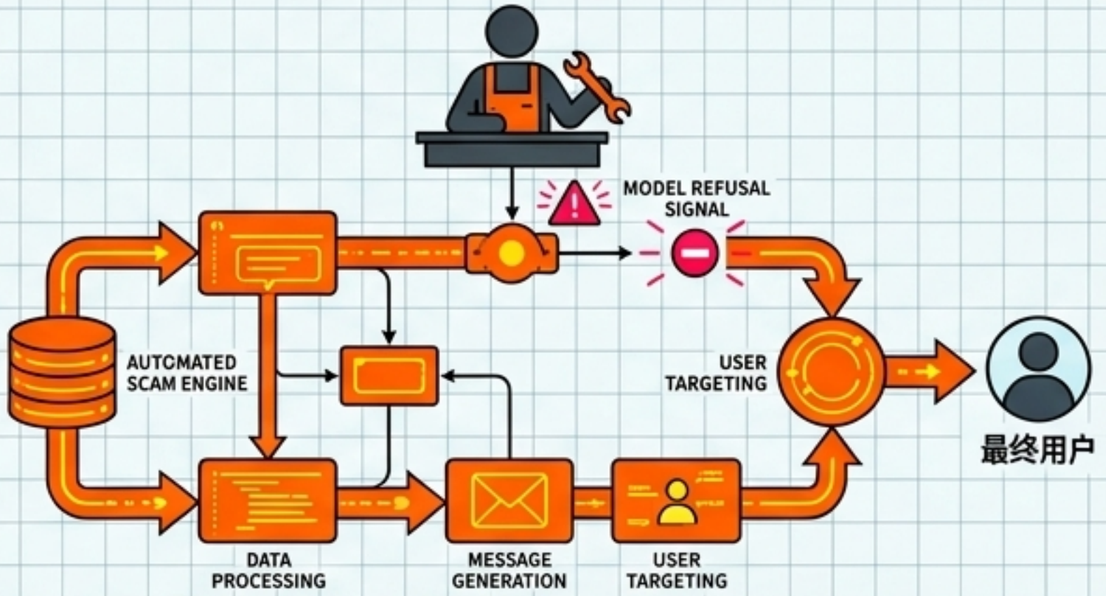
人类作为护栏（Guardrail）



审查违规行为，拦截越权操作，保护最终用户。

恶意视角的 HITL

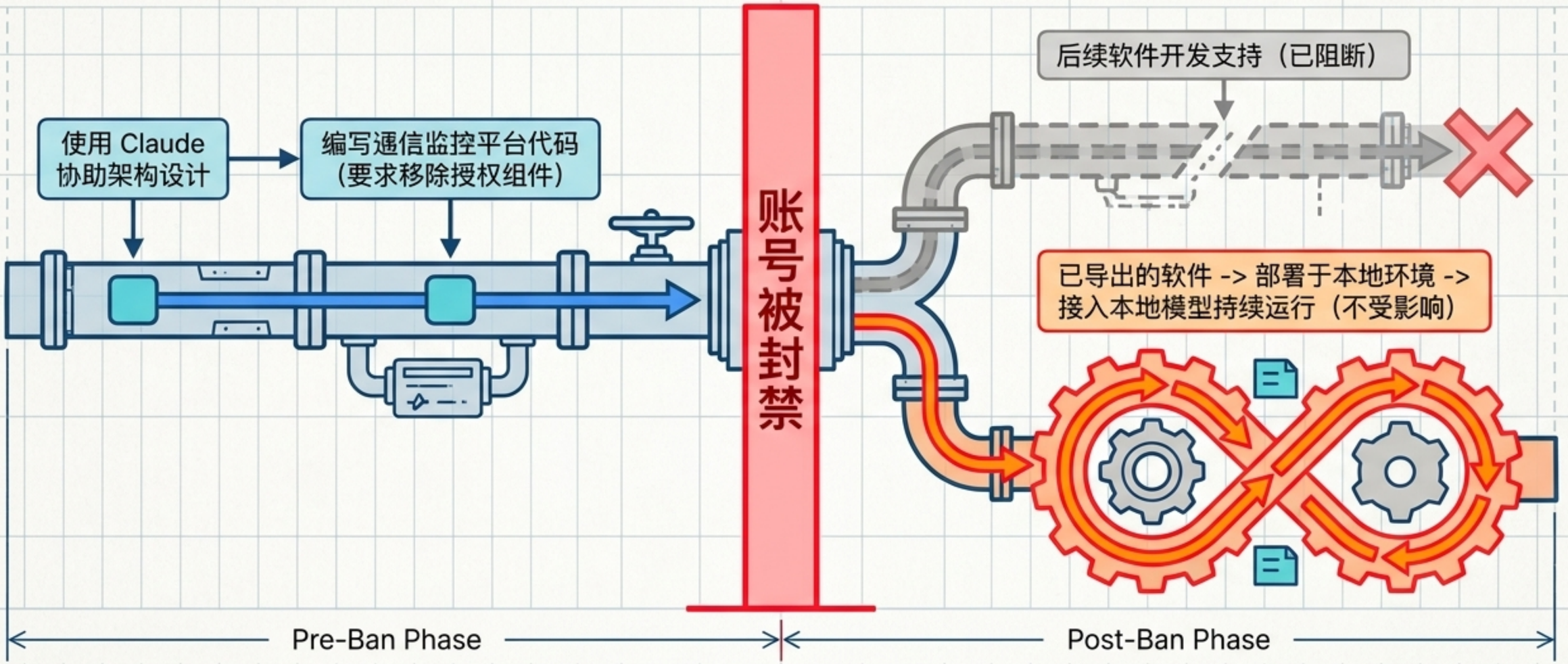
人类作为骗局监督者（Overseer）



真人与 AI 混合聊天，负责让骗局继续成立，覆盖模型的拒绝机制。

当整个系统本身就是一门骗人生意时，单句审核毫无意义。
评估 HITL 时必须追问：这个人到底代表谁？他们有权阻止什么？

帐号封禁的控制边界：被切断的 API 与仍在运行的本地代码



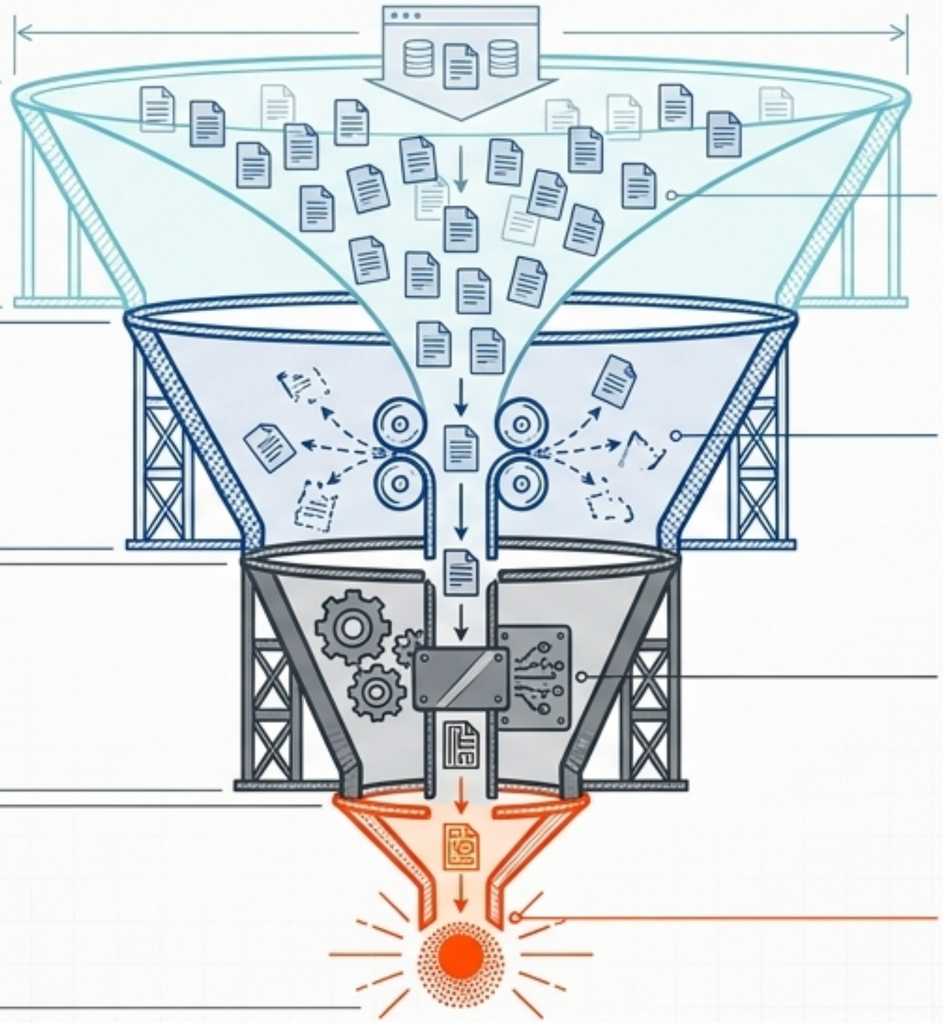
案例：马里安全部门监控平台。平台宣称的“已破坏”通常只意味着切断了持续调用与后续支持，并不等同于消除了现实影响。

平台控制力与伤害生命周期矩阵

模型参与到哪一步	平台可能直接控制的部分	还不能据此证明的结果
帮助设计或写软件	切断后续回答与开发支持	已导出的软件被撤回
持续参与线上运行	拦截该账号发起的新调用	其他模型和本地系统同时停机
内容或软件进入现实环境	与外部协作机构联合处理	所有副本、传播和实际伤害都已消失

封禁能增加对手成本，让未完成的项目停滞，但绝不能将“封禁账号”直接等同于“威胁解除”。

“生成内容”与“改变现实”之间存在巨大的转化漏斗



1.1

海量内容生成
（极低成本）

AI 将内容制作与日常运营的成本降至极低。

2.2

分发渠道与受众积累
（分发瓶颈）

许多机器生成的活动因缺乏真实互动，尚未建立受众便被处理。

3.3

信任度与信誉背书
（信任瓶颈）

触及较大受众的案例，往往严重依赖已有媒体渠道和现实关系。

4.4

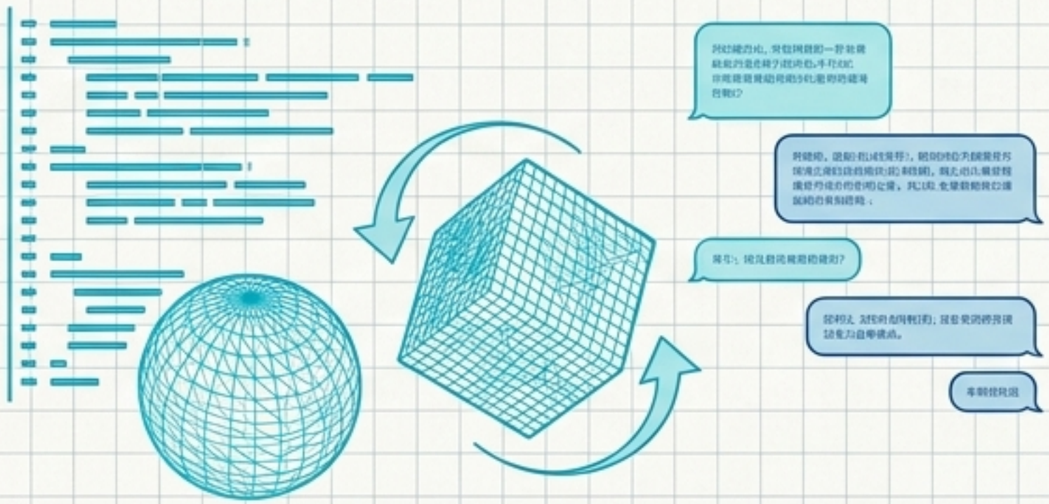
具体行为改变
（最难触达）

读者看到了、甚至相信了，距离改变实际行为或选举结果依然遥远。

传播分级只能说明内容触及了多大空间，无法直接证明现实观点的改变。AI 还没有绕开分发和信任的现实门槛。

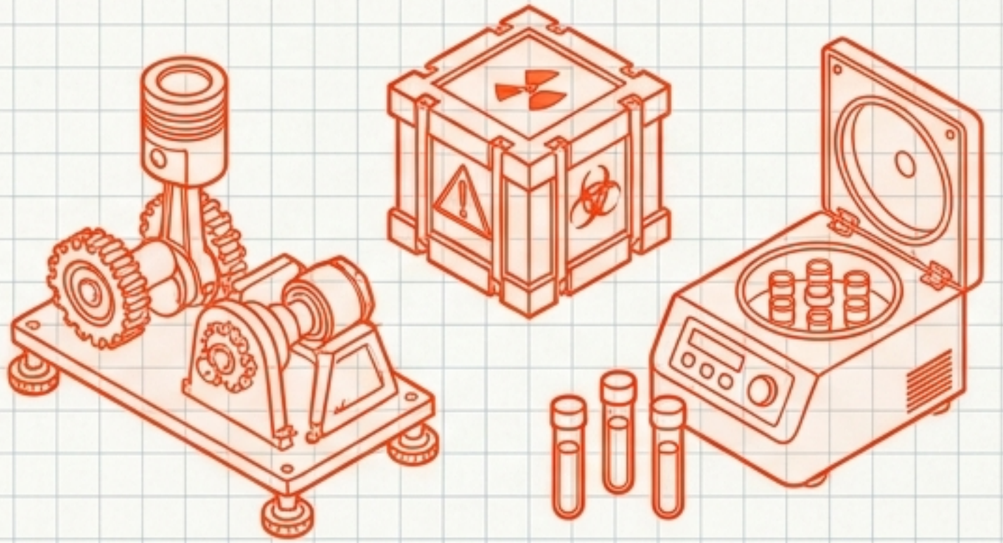
威胁定级的颗粒度：讨论武器方案不等于掌握实地攻击能力

数字/理论空间



方案讨论、仿真建模、软件开发

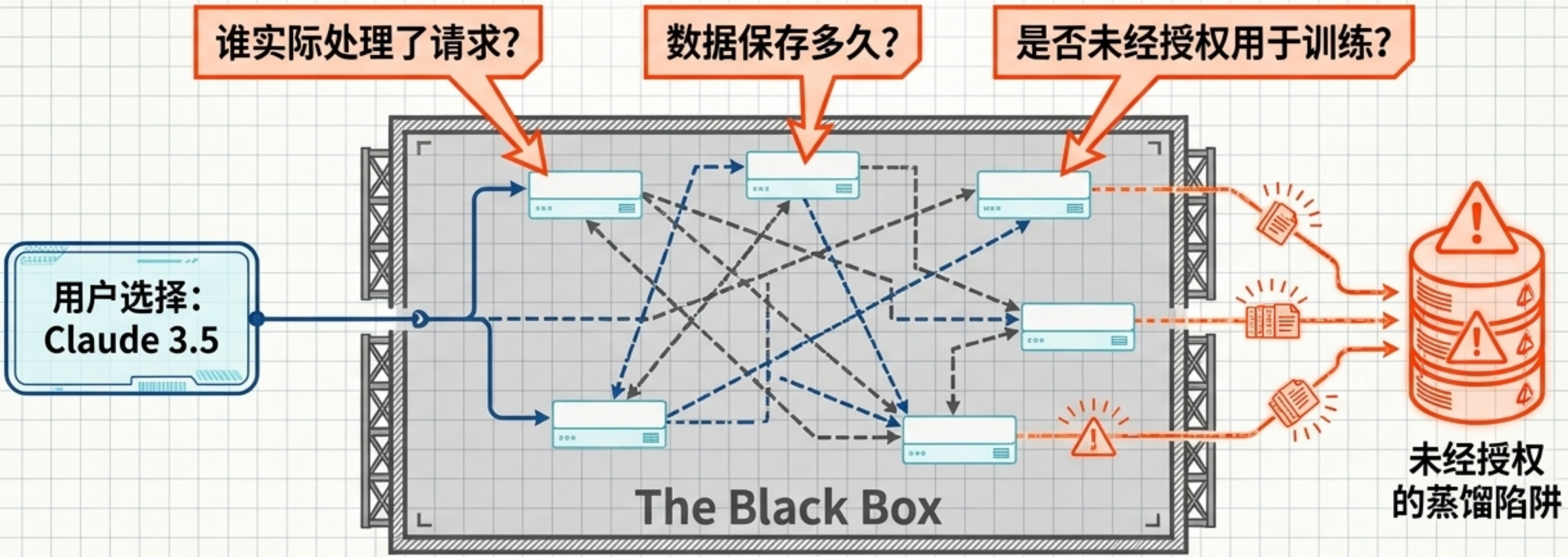
物理/现实空间



实地测试、有效装备部署、迫在眉睫的生物威胁

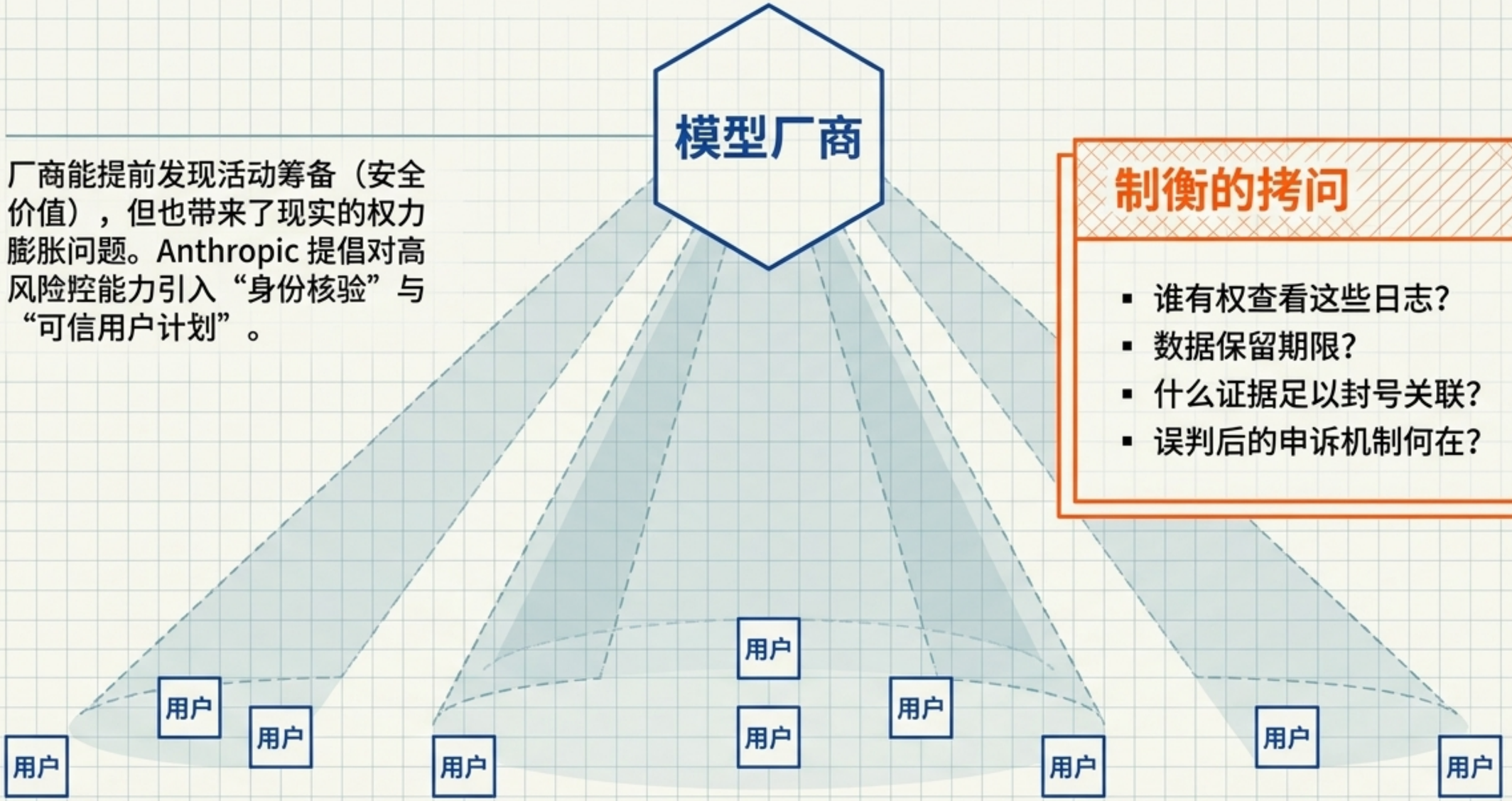
生物研究风险 (第130/137页)：涉事者为在职研究人员（无直接伤害意图）。报告指出的是“双重用途研究的识别难题”与“访问控制不足”，而非实质性的生物攻击。必须把限定条件与案例放在一起读，不能只留下最吓人的名词。

多模型路由背后的数据流转黑盒



报告指出，大规模的未经授权“蒸馏”常伴随用户数据的流转。一个模型标签，不能替代数据处理承诺。当输入包含内部文件或凭证时，便宜的 Token 单价无法弥补数据安全黑洞。

厂商能提前发现活动筹备（安全价值），但也带来了现实的权力膨胀问题。Anthropic 提倡对高风险控能力引入“身份核验”与“可信用户计划”。



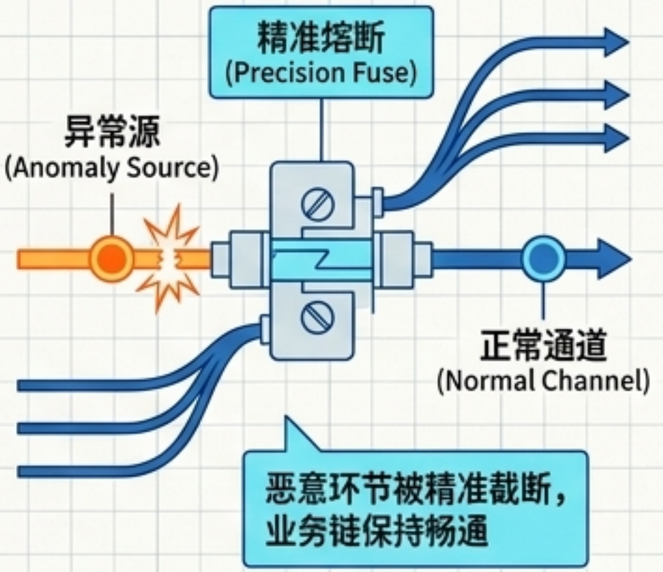
制衡的拷问

- 谁有权查看这些日志？
- 数据保留期限？
- 什么证据足以封号关联？
- 误判后的申诉机制何在？

我们不能因为目标是“安全”，就跳过对这些权力本身的约束。

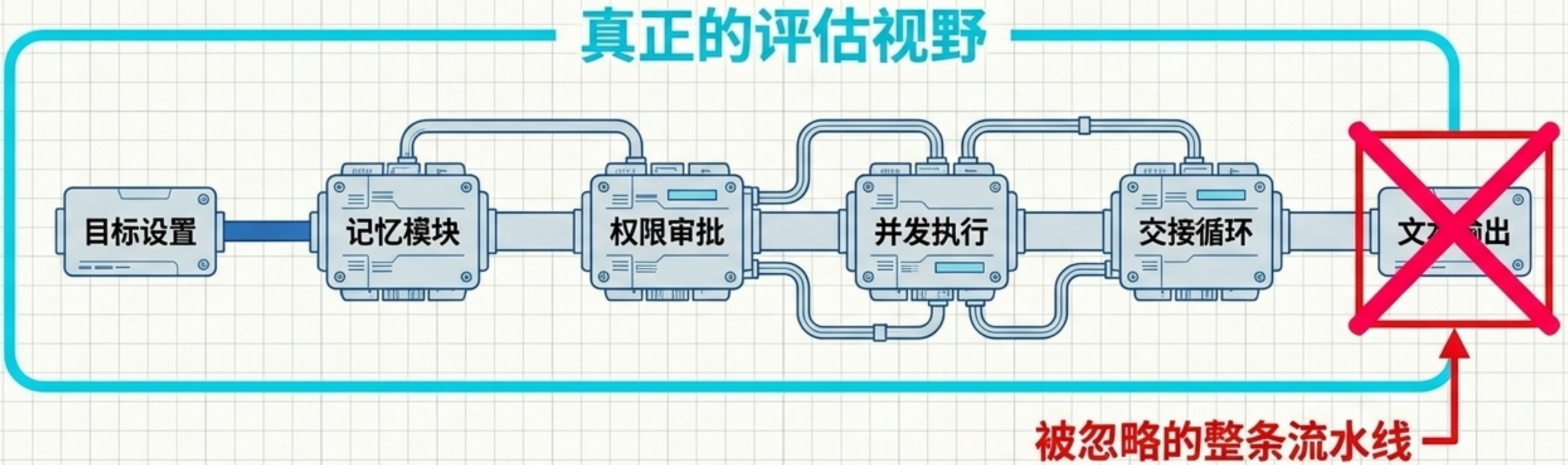
抵御系统级滥用：构建安全 Agent 的五项工程核对清单

☑	**1. 权限最小化：**获得的系统授权是否严格限制在当前单一任务内？
☑	**2. 读写分离：**“读取内部信息”与“向外发送网络请求”是否分属不同的权限闸口？
☑	**3. 后台生命周期：**核心任务结束后，进程是否还在后台静默持续运行？
☑	**4. 精准熔断：**发现异常时，能否精准停掉恶意环节而不导致整个业务链瘫痪？
☑	**5. 审计追踪：**留存的系统日志是否足够查清事件全貌，且未过度存储非必要隐私？



不要依赖无上下文的“人工点击确认”走形式，而要核查系统实际持续做了什么。

评估安全的最小单位不再是一次对话，而是一套系统



从“能回答”到“能持续做事”，中间增加的不只是代码，还有目标、记忆、交接与执行机制。
未来评估任何 Agent 系统的安全性，必须将这整条组织产线纳入视野。
只盯着最后生成的那句话，注定会漏掉真正的威胁。

深度研究与战略解码

报告来源：

Anthropic《Detecting and countering misuse of AI》（2026年9月10日发布，154页）

核心视点：

1. AI 改变了黑产的组织成本与并发效率。
2. 防御边界在于阻断整个自动化业务流，而非单一生成内容。
3. 供应商的监控权力扩张急需相应的规则制衡。