From Agent OS to Constitutional Runtime — 速读版

Why long-running AI work needs a governance layer, not a smarter agent.

> *5-minute read. 完整版见正文。*

> *本文中的 “Constitutional Runtime” 与 Anthropic 的 “Constitutional AI” 不是同一概念。后者关注模型训练阶段的价值对齐方法;本文聚焦运行时层面的 mission 治理结构。*

从 2025 年起,AI agent 系统不再只是”更会聊天的模型”。OpenClaw、Hermes、各种 Agent OS 已经把长期记忆、工具调用、跨渠道入口、后台任务全部串起来了。

但当我把这类系统真正用于跨周持续的工作时,问题反复浮现——而且每一次浮现的位置都不一样,却又看上去都像同一种病。

八个缺陷,一个结构

短任务里看不出来。任务一旦变长、变厚、变多分支,以下八种结构性缺陷会反复出现:

1. 任务归属漂移。 一个原本明确的任务,经过几十轮 context compaction 后,agent 已经在围绕另一件事工作,但它依然认为自己”在继续推进任务”。Chroma 在 18 个前沿模型上的实证:每一个都随上下文长度增长性能下降。

2. Support 偷渡成 ownership。 一句”顺便看一下 X”的辅助请求,三轮之后变成了系统认定的新主线。Compaction 会保留”做了什么”,却丢掉”它在任务结构里是什么身份”。

3. 记忆 ≠ 证据。 系统记得某件事曾经被讨论,不代表那件事仍然有效,更不代表它有资格改写当前状态。截至 2025 年 5 月,已有 116 起法庭案例记录律师把 LLM 生成的判例引用当成”已被记住即已被证实”提交,法庭核查时发现这些判例不存在。

4. 人格偷渡成法源。 给一个 agent 写一段 backstory,它就开始按 persona 一致的方式行动——arxiv 2601.10102 把这命名为 **role identity bias**,“即使与显式给出的目标、指令或激励相冲突”。CrewAI Issue #2838 里用户在 backstory 写明”MUST NEVER perform the tasks themselves”,runtime 仍然让 manager 亲自接管所有任务。

5. 委托 ≠ 授权。 Handoff 让 agent 进入流程,不等于授予了它任务主权。CrewAI Issue #4783 显示,其 hierarchical process 中 manager 从不真正 delegate,“effectively making the hierarchical process behave like a sequential process”;Issue #2054 显示 manager 会静默继承其他 agent 的工具权限。

6. 全局对话 ≠ 组织。 把所有 agent 放进同一个频道,后续 summary 会把不确定线索净化成事实,把审查意见误读成命令,把建议悄悄升级成主权决定。Anthropic 自家的 multi-agent research system 因此明确选择 orchestrator-workers 而不是 group chat。

7. 工具调用 ≠ 能力治理。 Endpoint 在线,不代表 agent 有权使用。OWASP 把此类风险列为 LLM06 “Excessive Agency”,MCPTox benchmark 显示 o1-mini 在 MCP tool poisoning 攻击下成功率 72.8%——失败原因不是模型不安全,而是 “existing safety alignment simply isn’t designed to catch malicious actions that use legitimate tools for unauthorized operations.”

8. 知道继续,不知道停止。 证据不足继续生成,权限不清默认推进,closure 不足就用一句”done”收尾。

Anthropic 自己 2026 年 2 月的数据:Claude Code 在最复杂任务上主动 clarify 的比例 16.4%。同一篇 11 月的工程博客承认反向失败:“declare the job done” 当工作并未真正完成。

一句话诊断

这八个不是互不相关的 bug。它们是同一个结构问题——

single-agent OS 把太多系统职责压进了一个连续对话体。

对话承担任务归属,记忆承担证据判断,persona 承担权限解释,工具列表承担能力治理,handoff 承担组织关系,“继续生成”承担任务推进,一句 “done” 承担 closure。短任务里压缩能 work;长任务里,所有边界开始互相污染。

真正缺的不是更聪明的 agent

更强的模型、更大的 context window、更厚的 memory、更丰富的 tool list——确实可以缓解一些表面症状。但它们不能自动解决任务归属、权限边界、证据等级、审计权和合法结案。

因为这些不是智能问题,**是组织问题**。

我说的 **Constitutional Runtime** 不是要重新发明一个 agent OS。OpenClaw、Hermes、CrewAI 这一层完全可以继续做执行平面。真正缺的是它们之上的一层组织结构——一个把 mission、authority、support、evidence、state delta、closure 当成一等 runtime object 的治理层。

下一篇会进入这一层的具体接口规范:mission envelope schema、authority relation schema、evidence item schema、state-delta policy、closure checklist、receipt ledger schema。

完整版正文展开了这八个缺陷与其公开实证。

完整版见 [Full version]; references list 在完整版底部。