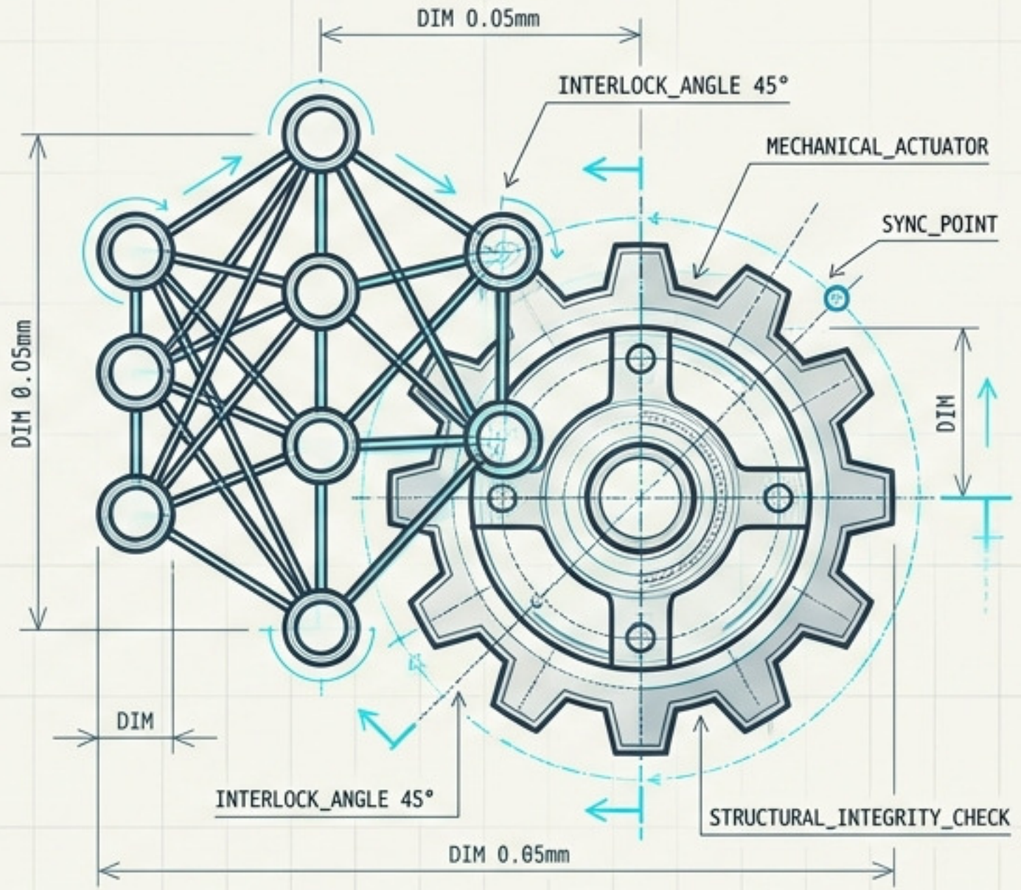
AI原生智能体·安全与合规

SYSTEM_INIT：从聊天到动手，构建自主系统的防御基座



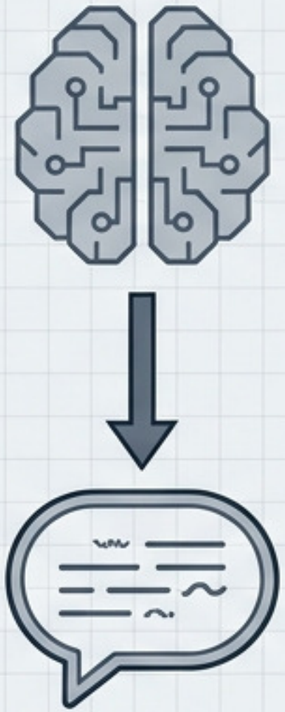
SECTION A-1

MODULE: SECURITY_PROTOCOLS

Speaker: Michael Huo
Format: 5-Part Architecture Series
Pace: Bi-weekly (Saturdays)

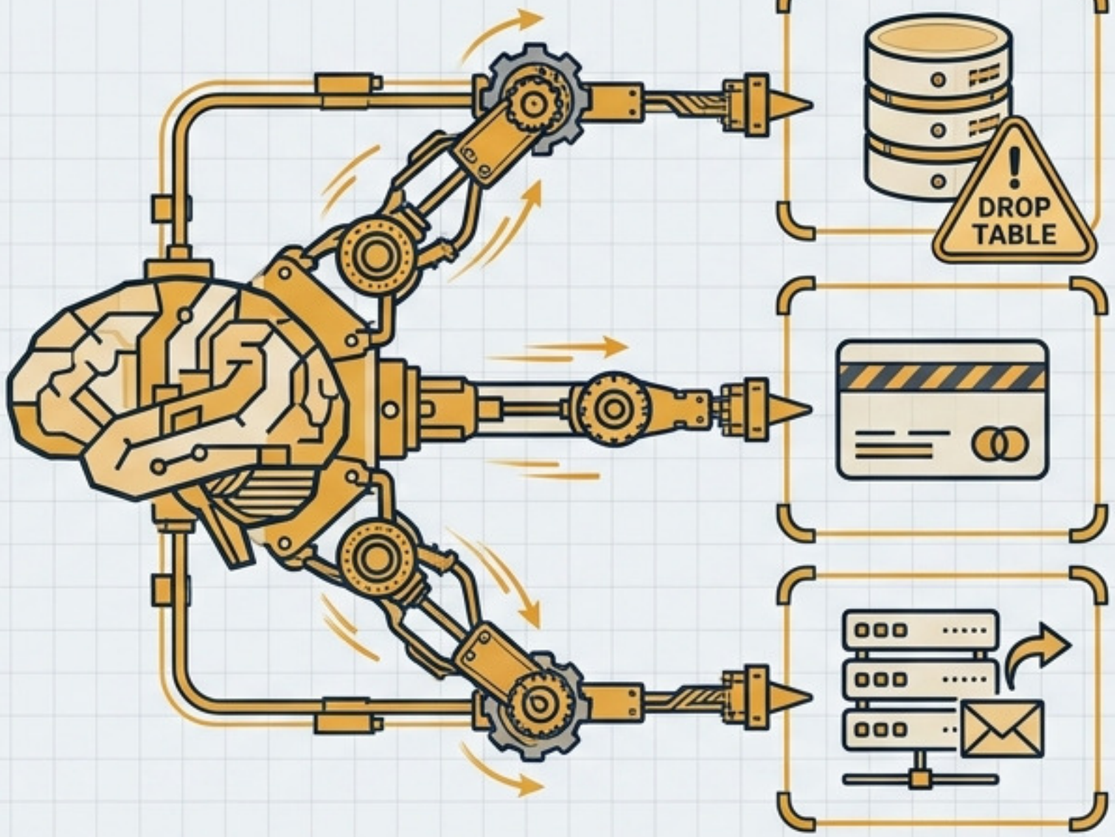
STATUS: READY | INPUT: TERMINAL | DATE: 2024-05-25

范式转移：当智能体开始「动手」



文本生成

风险局限于内容生成与幻觉



采取行动

风险呈指数级放大。系统可以直接改变现实状态。

划定安全边界

前置声明：本系列专注于「智能体原生」的安全架构。

通用网络安全

- 零信任架构

现有成熟框架覆盖

智能体安全与合规

- 权限控制、运行时沙箱、AI行为责任

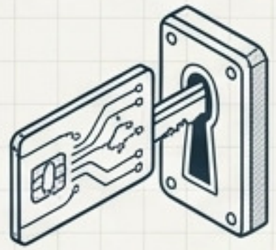
本系列的唯一核心

补充背景：此前的讲座已覆盖具身智能、智能体与新媒体、以及软件智能体。本次只谈安全。

智能体安全的第一性原理

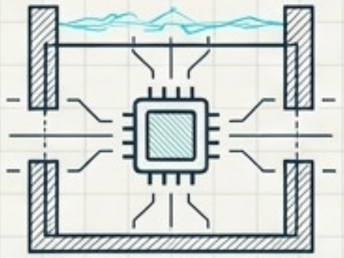
安全 = 权限 + 运行时 + 责任归属

权限



决定它能碰什么

运行时



决定它在什么环境执行

责任归属

决定出错了谁负责

> “智能体一开始动手，安全就是权限、运行时和责任归属。” —— Michael Huo

架构部署：5讲系统启动序列

第一讲（9月5日）

地图和八条红线

通用网络安全
零信任架构
现有成熟框架覆盖

第二讲（9月19日）

沙箱与人在回路

智能体安全与合规
权限控制、运行时沙箱、
AI行为责任

第三讲（10月3日）

真实产品上的最小权限

权限

第四讲（10月17日）

中美欧合规对照

责任归属

第五讲（10月31日）

上线检查表工作坊

运行时

每两周一讲，周六进行

运行时钟：单场执行周期

60分钟高密度输出架构

18:30–18:55 | 上半场

核心框架注入

18:55–19:00 | 休息

缓存清理

19:00–19:25 | 下半场

深度解析

19:25–19:30 | 问答

异常排查

严格准时，无冗余时间

[L1] 从聊天到动手，风险从哪来

2026年9月5日 | 基础共识

全局地图

建立智能体安全向量的完整视图。

八条红线

- WARN-A1 数据泄露
- WARN-A1 服务拒绝
- CRIT-R2 越权操作
- CRIT-R2 隐私违规
- ALERT-L3 恶意指令
- ALERT-L3 合规失控
- ALERT-L4 模型后门
- ALERT-L5 对抗攻击

绝对不可逾越的安全底线。

本讲不深入改产品配置

[L2] Harness、沙箱与人在回路

2026年9月19日 | 运行时防御

如何阻断?

沙箱
物理隔离执行环境。

白名单
仅允许预定义的极少指令。

真正的 Harness
监控输入输出的硬壳。

[L3] 真实产品上的最小权限

默认权限怎么收?

OpenClaw类	Cowork类	Codex类
开放式爬取/API调用	协同办公/文档介入	代码生成与执行
严格限制出站域名与数据写入权限。	基于角色的访问控制(RBAC)，仅限当前会话授权。	强制无网络环境编译，禁止系统级API调用。

[L4] 中美欧合规对照

出了问题，谁在法律上负责？

US

聚焦创新与特定行业监管

责任界定依赖现有侵权法。

EU

强监管与AI法案

明确区分系统风险等级，提供者与部署者责任连带。

China

算法备案与生成内容管理

强调服务提供者的主体责任。

上线前三行字 —— 决定生死的免责与合规声明。

[L5] 终局：上线检查表工作坊

8条红线校验（L1）

沙箱与白名单配置（L2）

最小权限策略映射（L3）

跨区域合规免责（L4）

带自己的场景来

用检查表打分

基于L1-L4输入评估
上线准备度

安全不是单点防御，而是从权限到合规的闭环工程。

系统接入端口

> INITIALIZE_REGISTRATION...

Luma 报名入口：https://luma.com/fiht7lab

ZJU 核心讲义：https://zjuaanc.org/posts/2026-09-01-ai-native-agent-security-compliance-lecture-1/

NACU 镜像备份：https://nacuaanc.org/posts/2026-09-01-ai-native-agent-security-compliance-lecture-1/

智能体原生时代的防御基座，从9月5日开始构建。