

真实产品上的最小权限。 AI原生智能体 · 安全与合 规，第三讲。

2026年10月3日。

主讲：Michael Huo。

嘉宾：Zachry Xiong。

**第一讲：风险从
说错变成做错。**

PHASE 01: RISK
TRANSITION FROM
MISSTATEMENT TO
MALFUNCTION.

**第二讲：墙在
Harness、沙箱
和人在回路。**

PHASE 02: WALLS IN
HARNESS, SANDBOX, AND
HUMAN-IN-THE-LOOP.

**今晚：真实产品
上的最小权限。
同一公式。本讲
打在权限。**

TONIGHT: LEAST PRIVILEGE
ON REAL PRODUCTS. SAME
FORMULA. THIS LECTURE
HITS ON PRIVILEGES.

过度授权是六类风险的第一项。最小权限是对它的控制，不是新题目。

过度授权

密钥泄漏

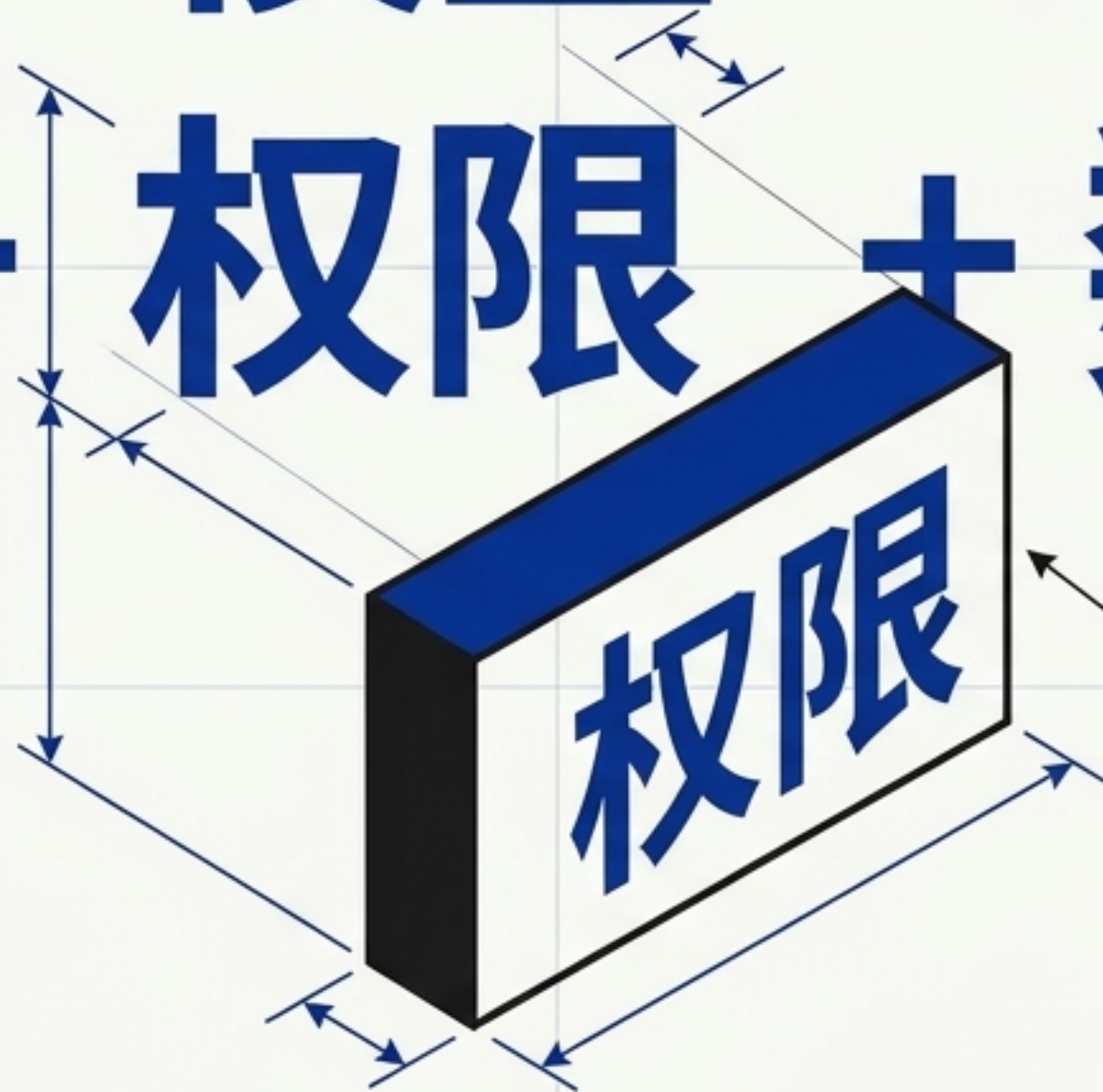
间接注入

工具误操作

数据越权

责任不清

智能体 = 模型 + Harness
+ 工具 + 权限 + 数据。



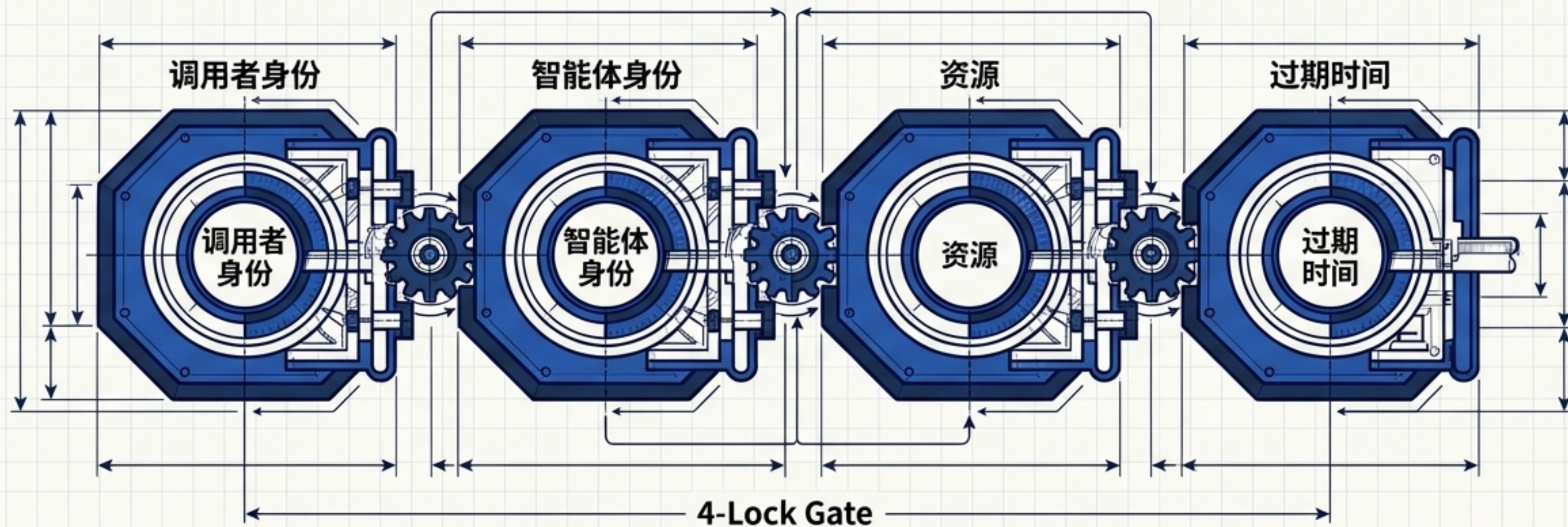
第二讲把墙放在运行时。
今晚问这堵墙允许打开
开什么。

边界装在运行时，不装在提示词里。



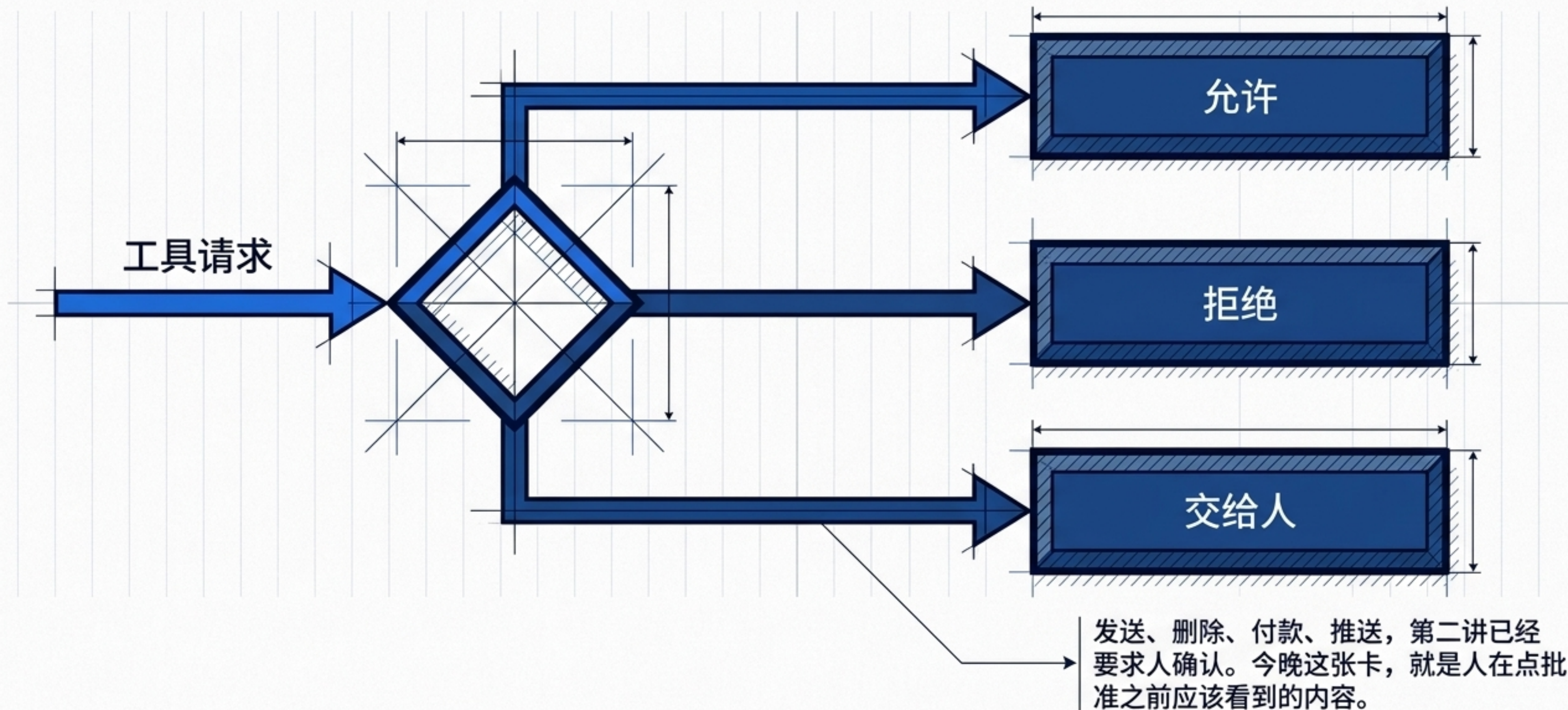
提示词不是墙。批错了也是批准。
日志第二天还要能读。

谁可以调哪把工具，碰哪份资源，默认是否拒绝？



四问有一项是空的，本系列的默认就是拒绝。
这不是网络零信任课。

每次调工具前，系统先做一个有界决定。



三句话分开。

**设计说法不是
缺陷报告。**

**缺陷报告不是
已统计的入侵。**

**“我们更新了”
不是你可以核
对的版本。**

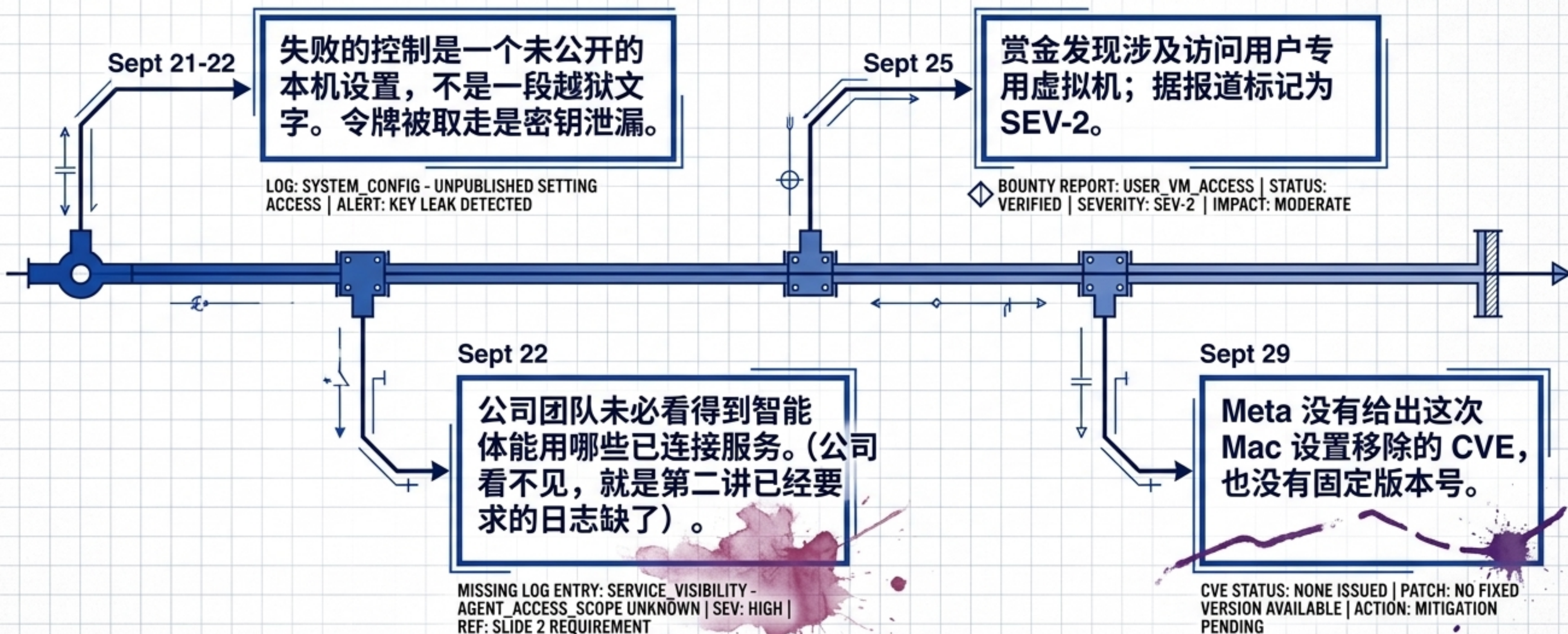
Muse 是 Meta 的个人智能体，2026 年 9 月初发布。接上的服务可以包括邮件、文件、购物、订票和付款。

假定智能体可能被攻击；把损失限住；智能体手里没有真实凭证。

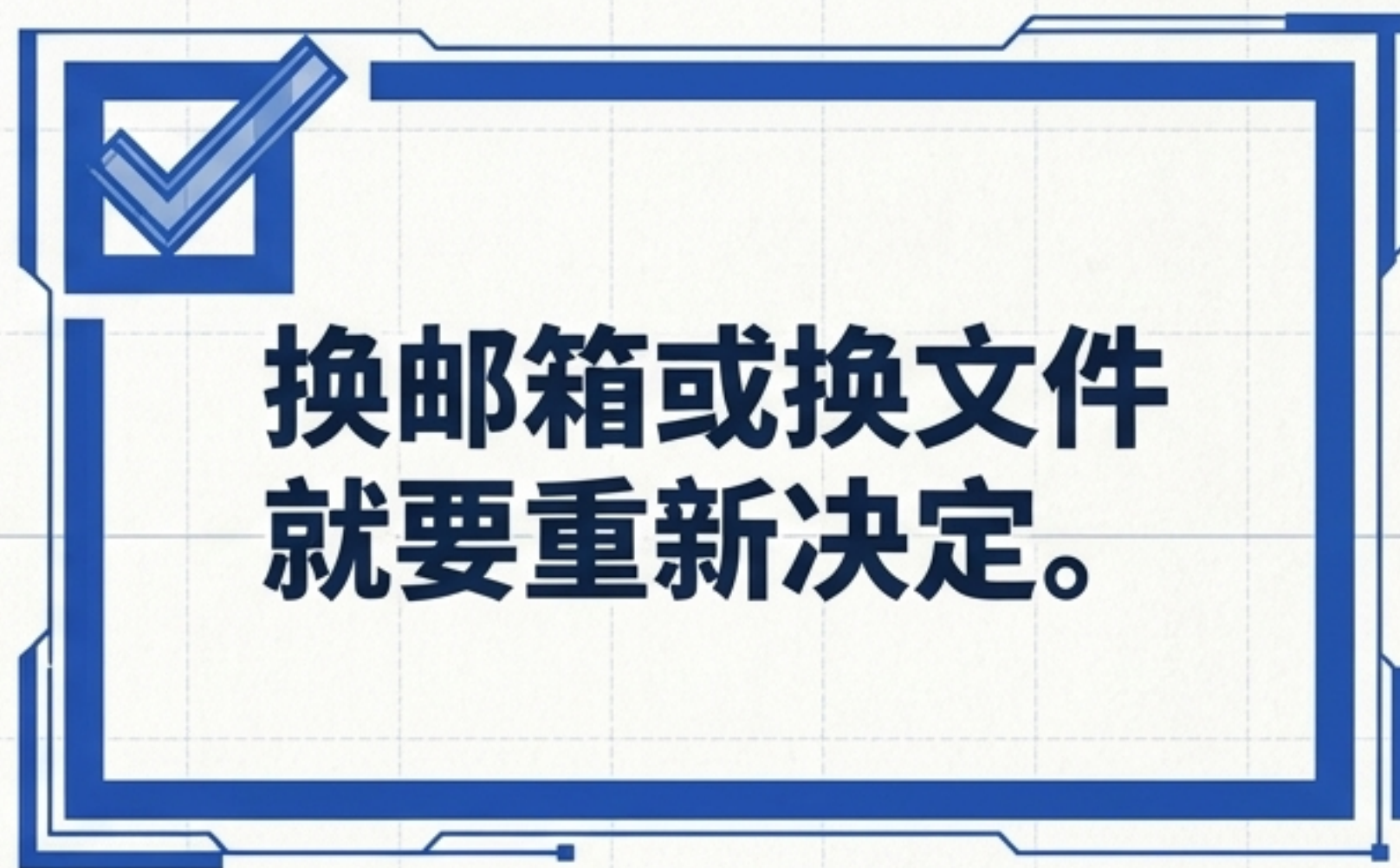
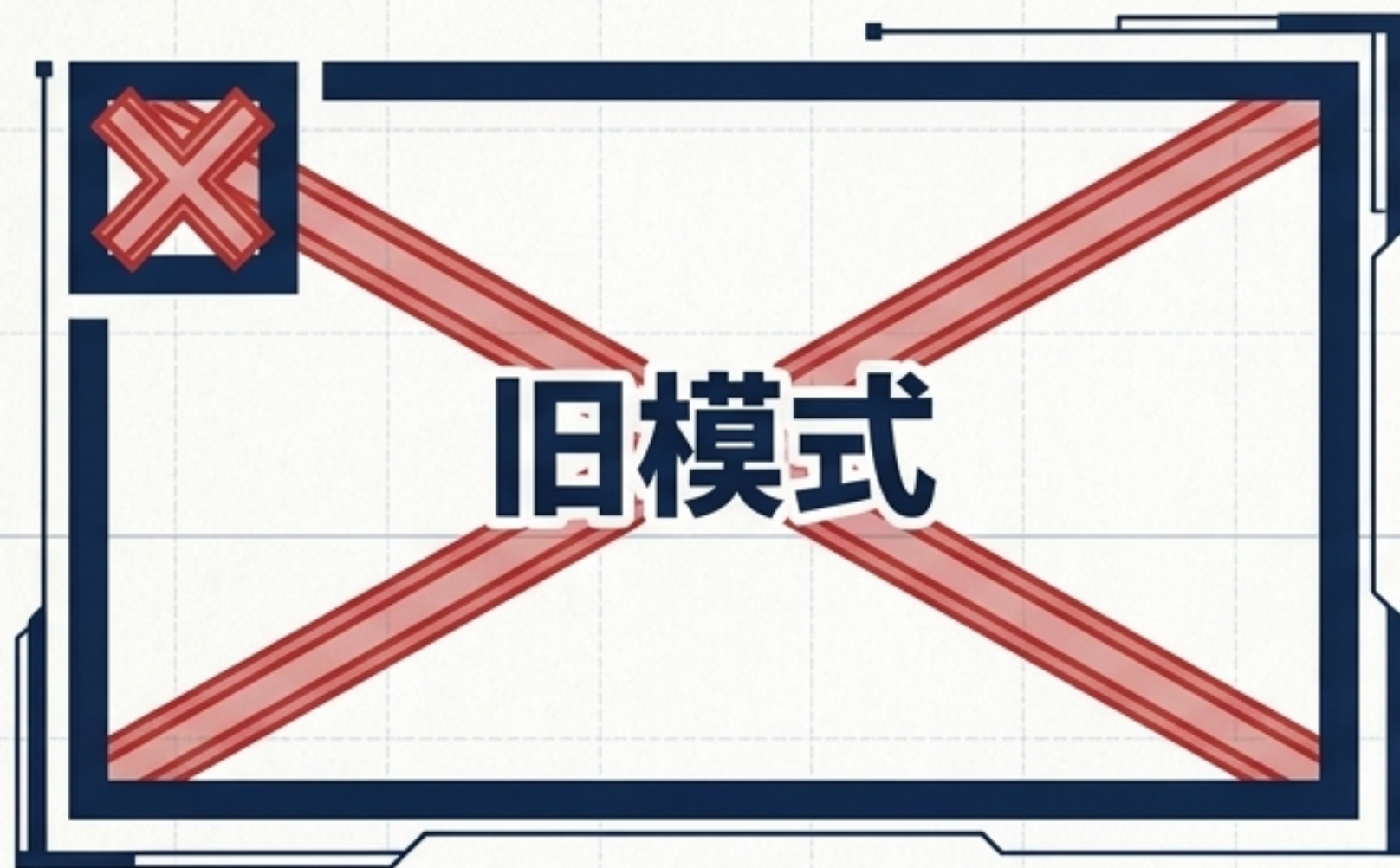
外部调用经过智能体无法覆盖的 Sentinel。

**同一 Mac 用户下的本机程序可重定向 Muse 听写并捕获认证令牌。
(本地权限提升，非远程利用)。**

Anatomy of an Incident



零信任，智能体形态。不是共用的机器人账号。不是把用户密码抄进智能体。



2026 年对 CISA 及伙伴说明的公开综述：每个智能体单独身份、不给宽权限、人要能监督、要有监测。这是控制清单，不是证书。

MCP = 工具 $\cap$ 权限

一次工具调用就是一次访问请求。模型文字不是授权。

公开 MCP 服务器约
38% 到 40%
没有认证。

一份 5200 台的审计：
8.5% 用 OAuth，
53% 用静态 key。



CVE-2026-59822 (LiteLLM MCP 端点)

- 9 月 2 日列入 CISA 已知被利用清单。
- 伪造的 bearer token 可以开会话。

自己在跑就打补丁。不要演示绕过。



令牌必须是发给这台服务器的；不要透传用户令牌。

到 9 月 30 日的五套厂商登录互不通用。多一个登录框不等于最小权限。

嘉宾： Zachry Xiong

工具运行前的类型化决定。
输入是状态加类型化问题；
输出是选择、分数或是否，
并带概率。
可以放在工具门禁。

不能代替沙箱和人审。
类型安全不等于正确。
置信度不是授权。

权限卡

第一讲里的危险调用，每个一张卡：发送、删除、付款、推送、命令行。

调用者	智能体身份	工具	动作	资源	默认	过期	证据	责任人

默认

证据

动作状态

命令行和付款默认
拒绝。模型打分不
是这张卡的栏位。

证据是服务的决定
和记录没有变，不是
聊天里那句“我不会”。

这一行标：通过、
失败、无法判定、
未执行。没测
就是未执行。
系列里这格写
「尚未做到」。

只是例子。不是对某个产品的调查结论。

权限卡

第一讲里的危险调用，每个一张卡：发送、删除、付款、推送、命令行。

调用者	智能体身份	工具	动作	资源	默认	过期	证据	责任人
已登录用户	邮件草稿	发送	发出	这一封、这一个收件人	人审	只限这一封（换收件人就要新卡）	发送日志还没回读，就是未执行	点审核的那个人

Action Plan

选一个你正在用的智能体。列出三把工具。每把写清：谁能调、作用在哪、失败时怎样停。

高风险动作标「typed gate + 人审」或「尚未做到」。

已安装的智能体要更新。
博客不是补丁记录。

Next Event

10 月 17 日

**中美欧合规对照。
把这张卡带去。**