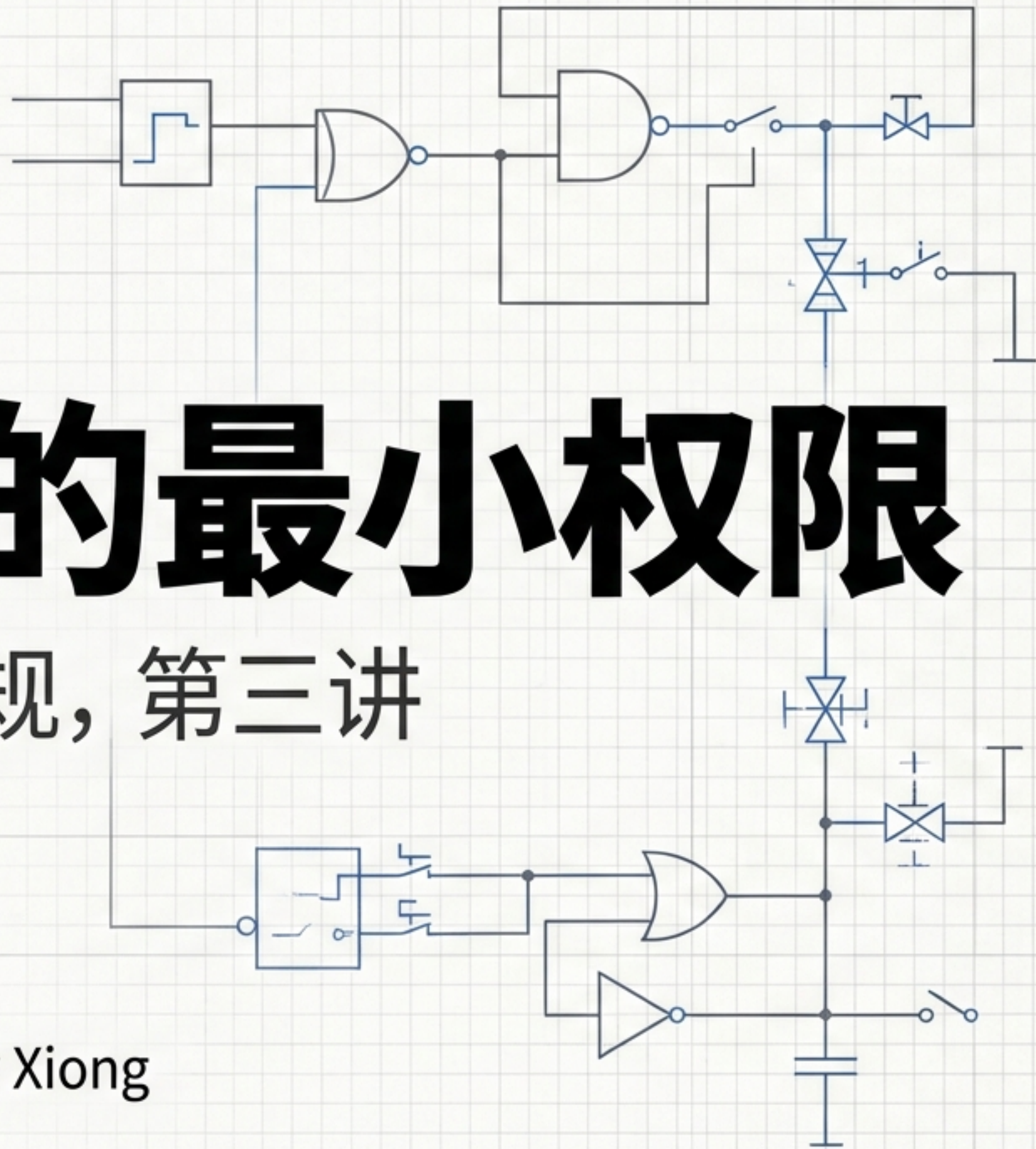


真实产品上的最小权限

AI原生智能体·安全与合规，第三讲

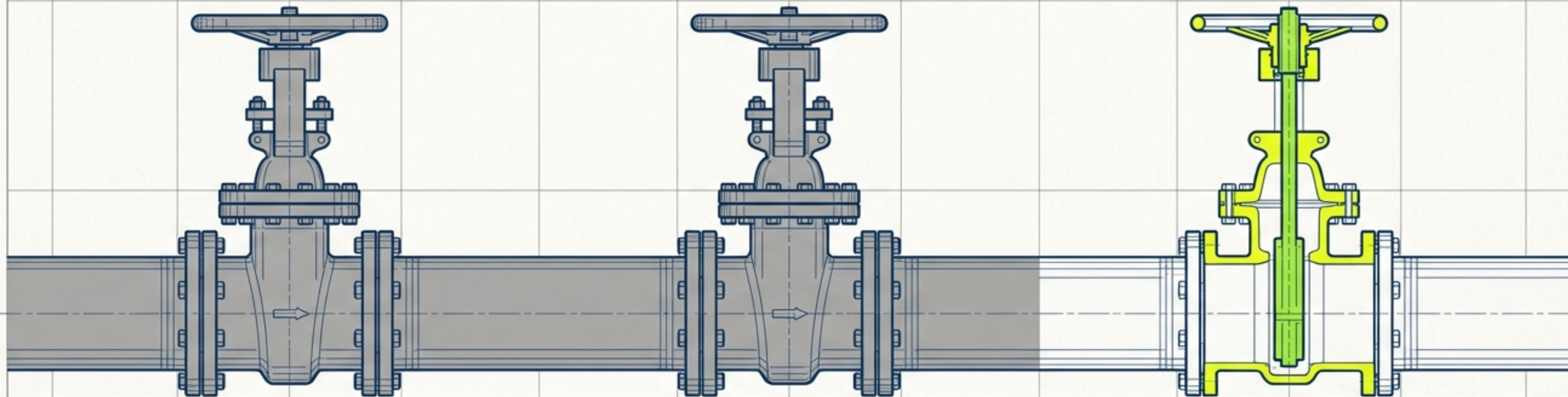
2026年10月3日 | Michael Huo | 嘉宾: Zachry Xiong



第一讲：风险从说错
变成做错

第二讲：墙在 Harness、
沙箱和人在回路

今晚：真实产品上的
最小权限



同一公式。本讲打在权限。

智能体 = 模型 + Harness + **[权限]** + 数据

第二讲把墙放在运行时。今晚问这堵墙允许打开什么。



过度授权

密钥泄漏

间接注入

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。

工具误操作

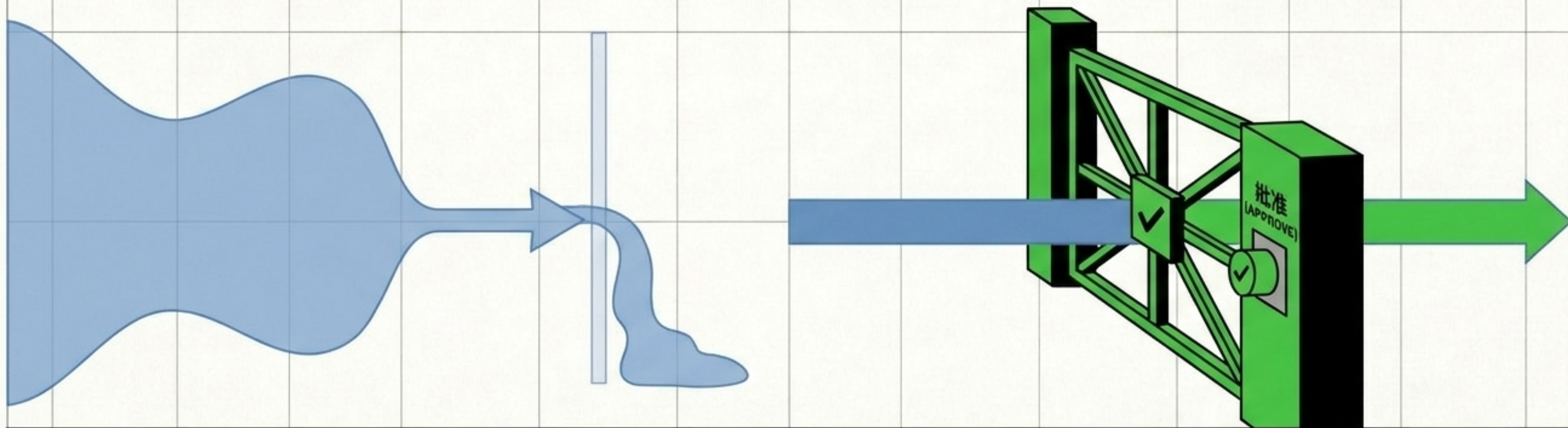
数据越权

责任不清

- 另外五项仍在同一张卡上：
 - 密钥不进提示词
 - 检索来的文字不是命令
 - 工具用错
 - 别人的数据
 - 出错要有责任人

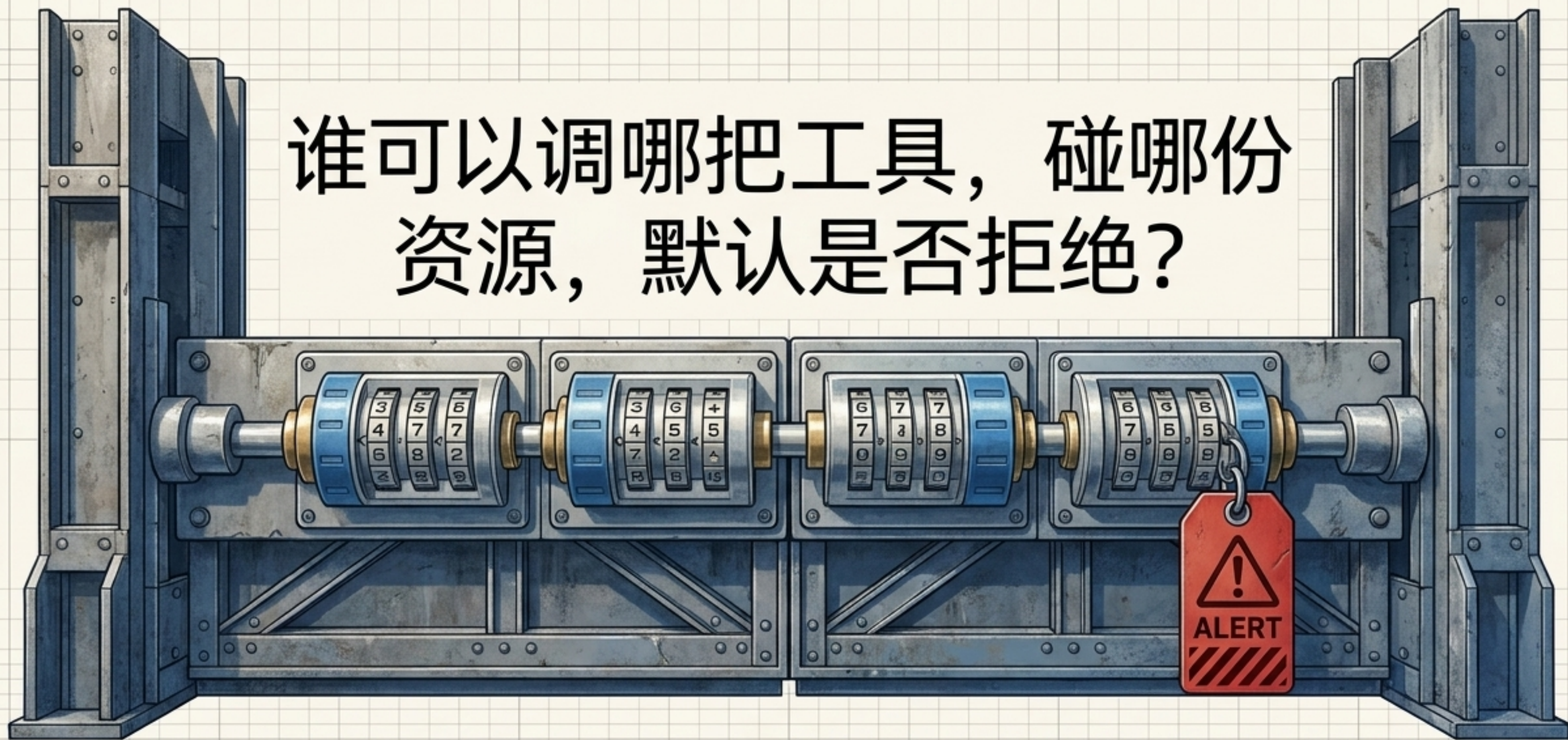
提示词不是墙。

**发送、删除、付款、推送，
第二讲已经要求人确认。**



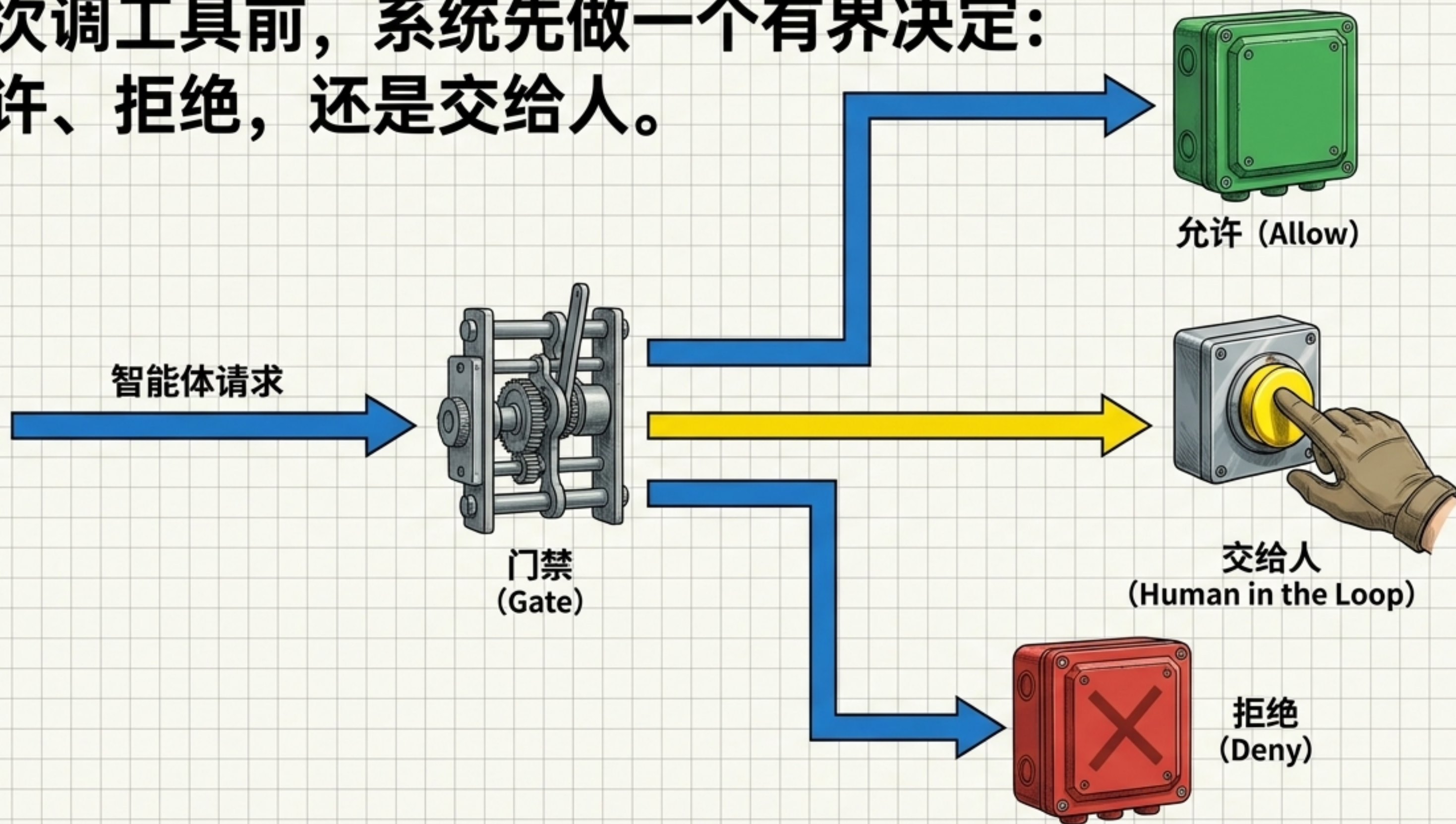
**批错了也是批准。日志第二天还要能读。
今晚这张卡，就是人在点批准之前应该看到的内容。**

谁可以调哪把工具，碰哪份
资源，默认是否拒绝？

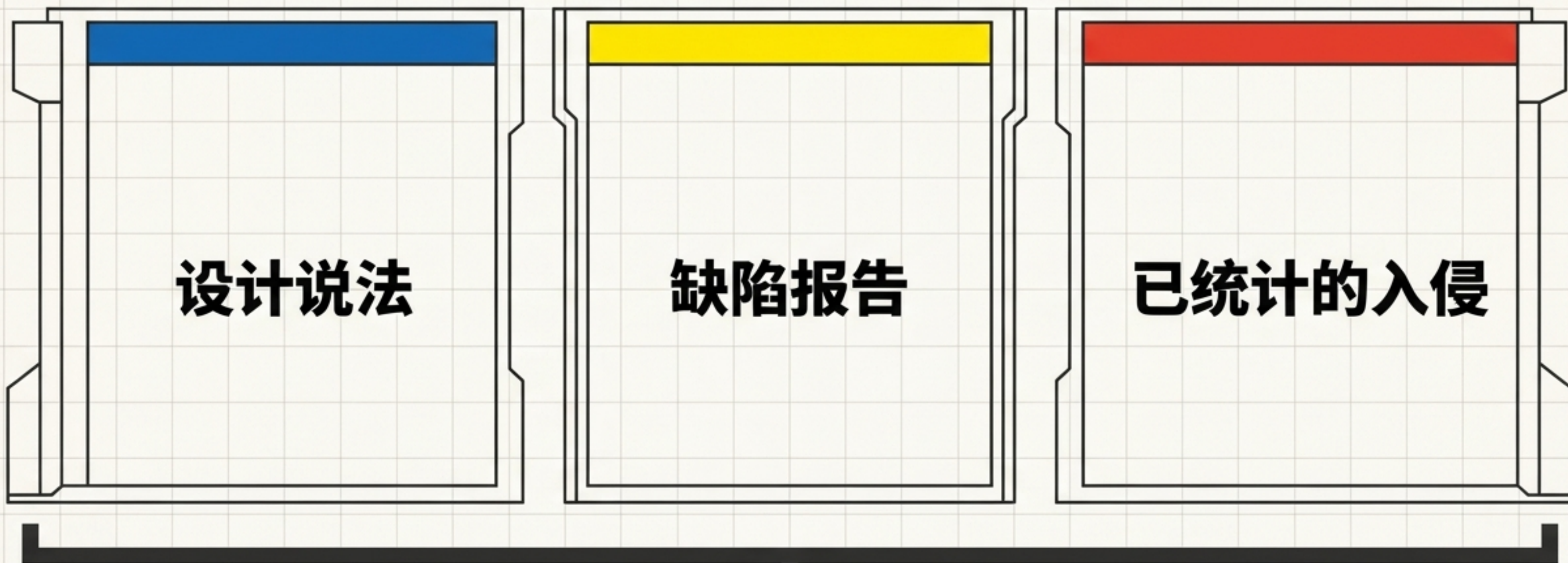


四问有一项是空的，本系列的默认就是拒绝。

每次调工具前，系统先做一个有界决定：
允许、拒绝，还是交给给人。

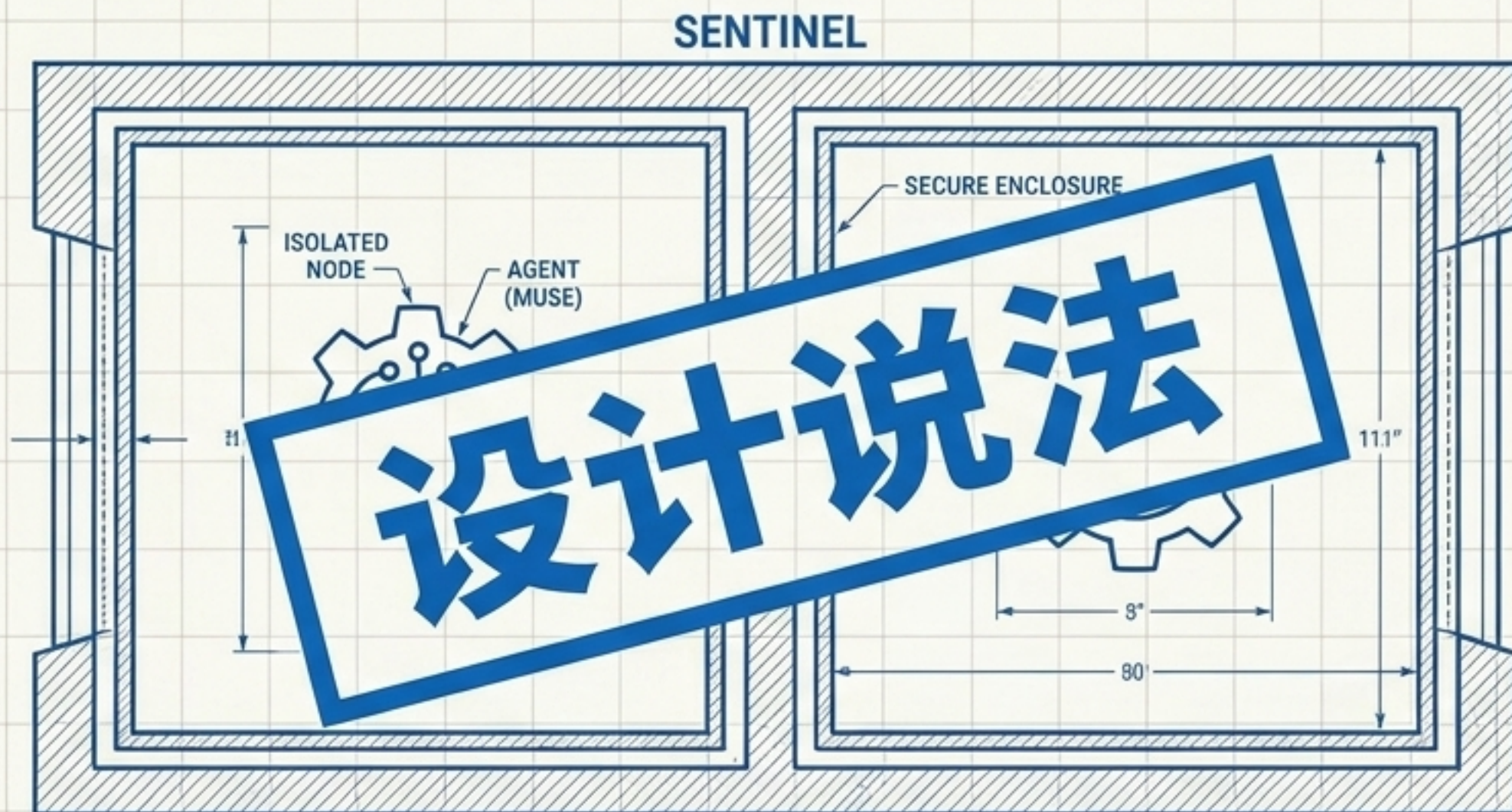


三句话分开。



设计说法不是缺陷报告。缺陷报告不是已统计的入侵。

**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。**

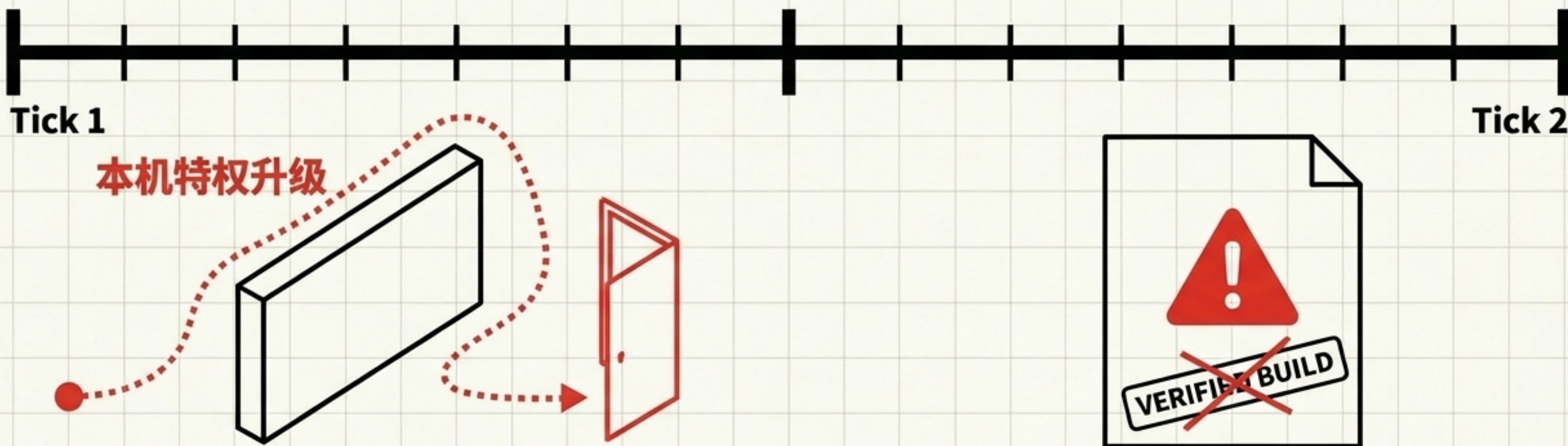


**留下的说法：
假定智能体可能被
攻击；把损失限住；
智能体手里没有
真实凭证。**

这是权限项上的说法，不是证明。

已报道的缺口：失败的控制是一个未公开的本机设置，不是一段越狱文字。公开的 Mac 证明只需要同一用户账户。

到 9 月 29 日，公开笔记写 Meta 没有给出这次 Mac 设置移除的 CVE，也没有固定版本号。“我们更新了”不是你可以核对的版本。



不要说所有用户都被远程攻破。诱使粘贴命令可以补上本机这一步。
这是对诱饵的描述，不是已统计的入侵次数。

接回系列。



**Mac 设置是过度授权
加上责任不清。**

**令牌被取走是
密钥泄漏。**

Audit Trail

```
[2023-09-22 14:30:15 UTC] USER_123 access_request:  
resource=DB_MAIN, action=READ, result=ALLOWED  
[2023-09-22 14:35:22 UTC] USER_123 prompt_submitted:  
'Extract summary of Q3 report'  
[2023-09-22 14:35:25 UTC] MODEL_RESPONSE: 'Here is  
the summary...'  
[2023-09-22 14:40:01 UTC] USER_123 tool_call:  
endpoint=API_SALES, params={region='ASIA'},  
result=SUCCESS
```

**VentureBeat, 9月22日:
用户也许有轨迹**

这是最小权限的可见性那一半。公司看不见，就是第二讲已经要求的日志缺了。



**公司团队未必看得到智能体
能用哪些已连接服务。**

这不是网络零信任课。

不是 (What It Isn't)

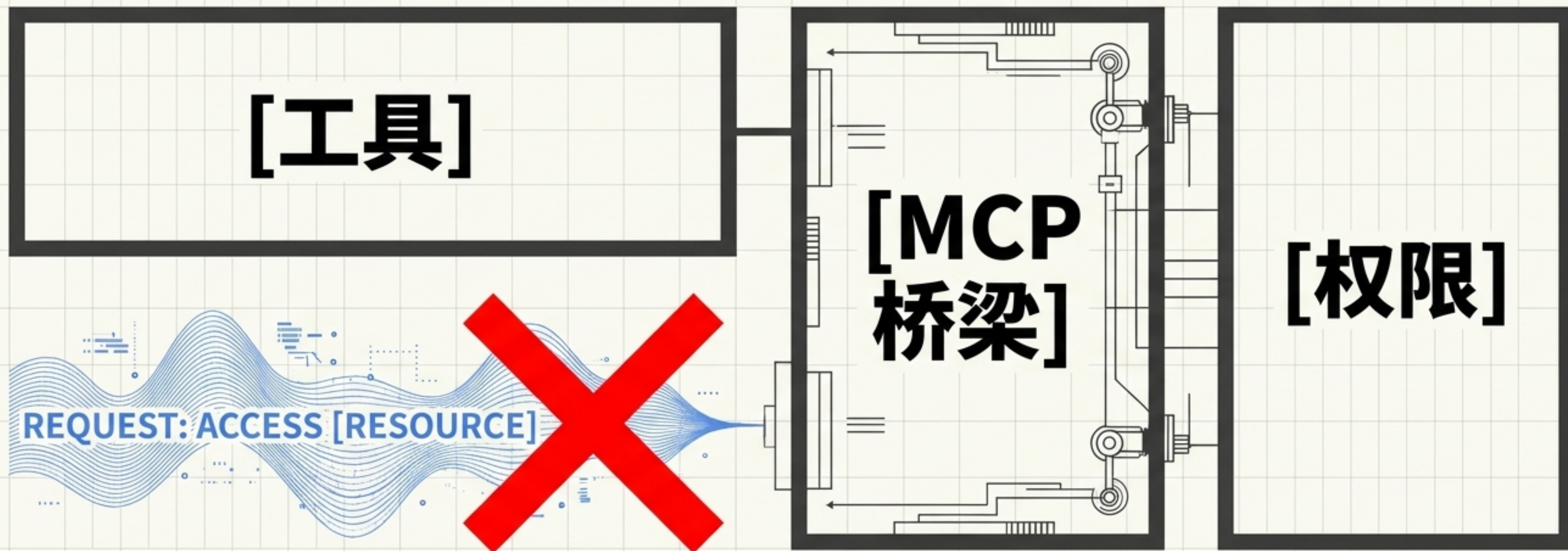
- 不是共用的机器人账号。
- 不是把用户密码抄进智能体。

是 (What It Is)

- 每个智能体单独身份。
- 换邮箱或换文件就要重新决定。
- 这是控制清单，不是证书。

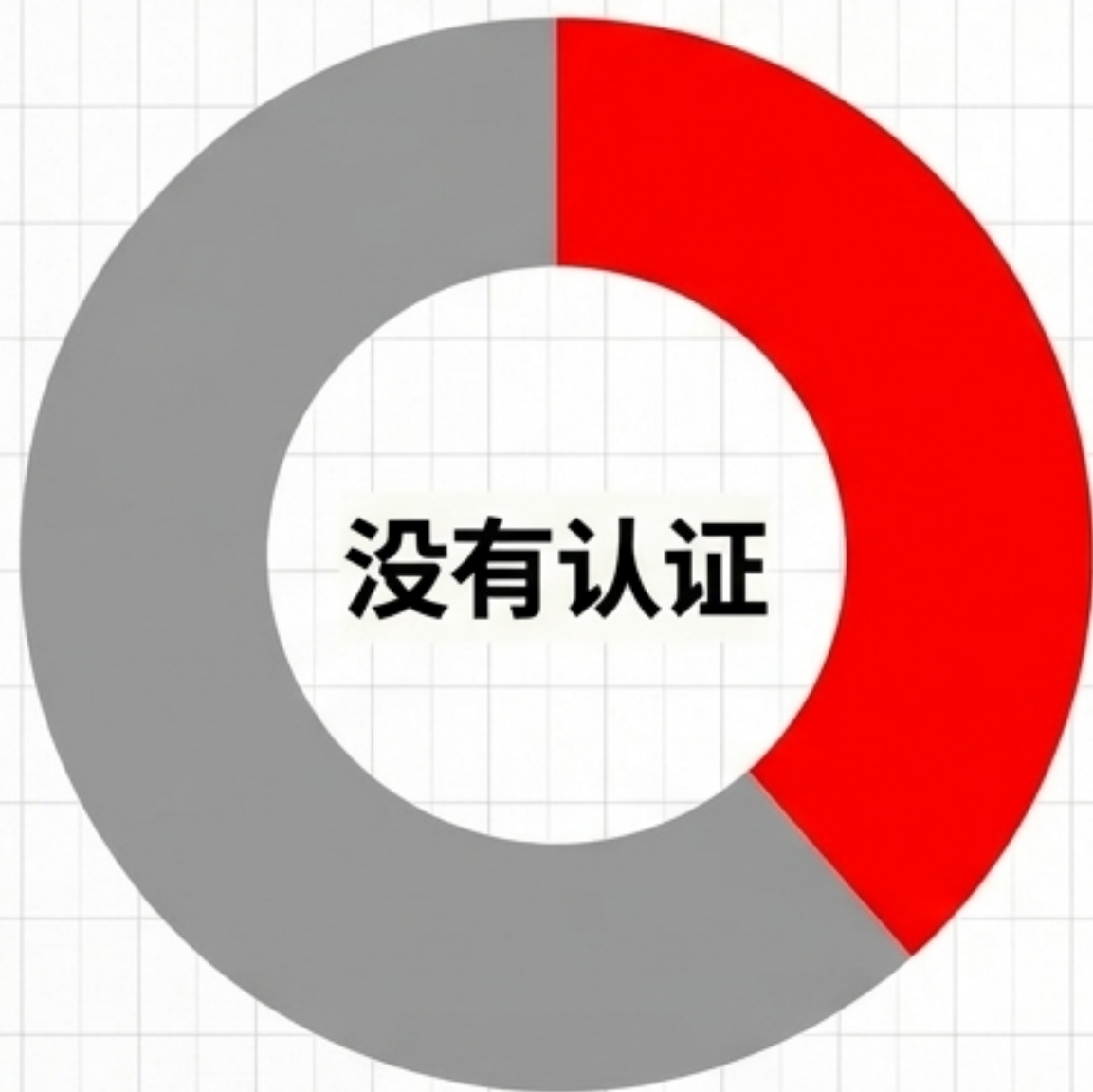
2026 年对 CISA 及伙伴说明的公开综述：每个智能体单独身份、不给宽权限、人要能监督、要有监测。
工具调用上四项：调用者身份、智能体身份、资源、过期时间。

MCP 是工具项碰上权限项。



一次工具调用就是一次访问请求。模型文字不是授权。

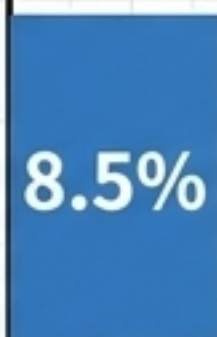
9月下旬的调查，不是你的审计：



没有认证

公开 MCP
服务器约
38% 到 40%
没有认证。

一份 5200 台的审计：



8.5%

8.5% 用 OAuth

53%

53% 用静态 key



WARNING NOTICE



CVE-2026-59822, LiteLM 的 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。



要留下的规范一句：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。

接下来 15 分钟：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。Michael 停在这里。

类型安全 $\neq$ 正确

置信度 $\neq$ 授权



这张卡仍然要谁、哪把工具、哪份资源、多久。

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 (允许/拒绝/人审)	过期	证据	责任人

填卡规则

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”。



通过

失败

无法判定

尚未做到

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」。**

高风险动作标「typed gate + 人审」，或「尚未做到」。发、删、付仍要人点。

只是例子。不是对某个产品的调查结论。

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

课后作业

1

选一个你正在用的智能体。列出三把工具。

2

每把写清：谁能调、作用在哪、失败时怎样停。

3

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。

下一讲预告

10月17日

10月17日是中美欧合规对照。

调用者	智能体身份	工具	动作	资源	默认	过期	责任人	证据
已登录用户	邮件草稿	发送	发出	这一封， 送一个收件人。换 收件人就 要寄丰。	人审	月满送 这一封	高顺枝 的哪个人	发送日志 送送团谈 谈，就是 未执行。

把这张卡带去。