

AI原生智能体安全与合规第四讲： 中美欧合规对照

AI-Native Agent Security & Compliance, Lecture 4: US, China, EU

讲者 : Michael Huo

时间 : 2026年10月17日 周六

5:00-6:00 PM PDT

美东时间 8:00-9:00 PM EDT

北京时间 10月18日 周日 8:00-9:00 AM

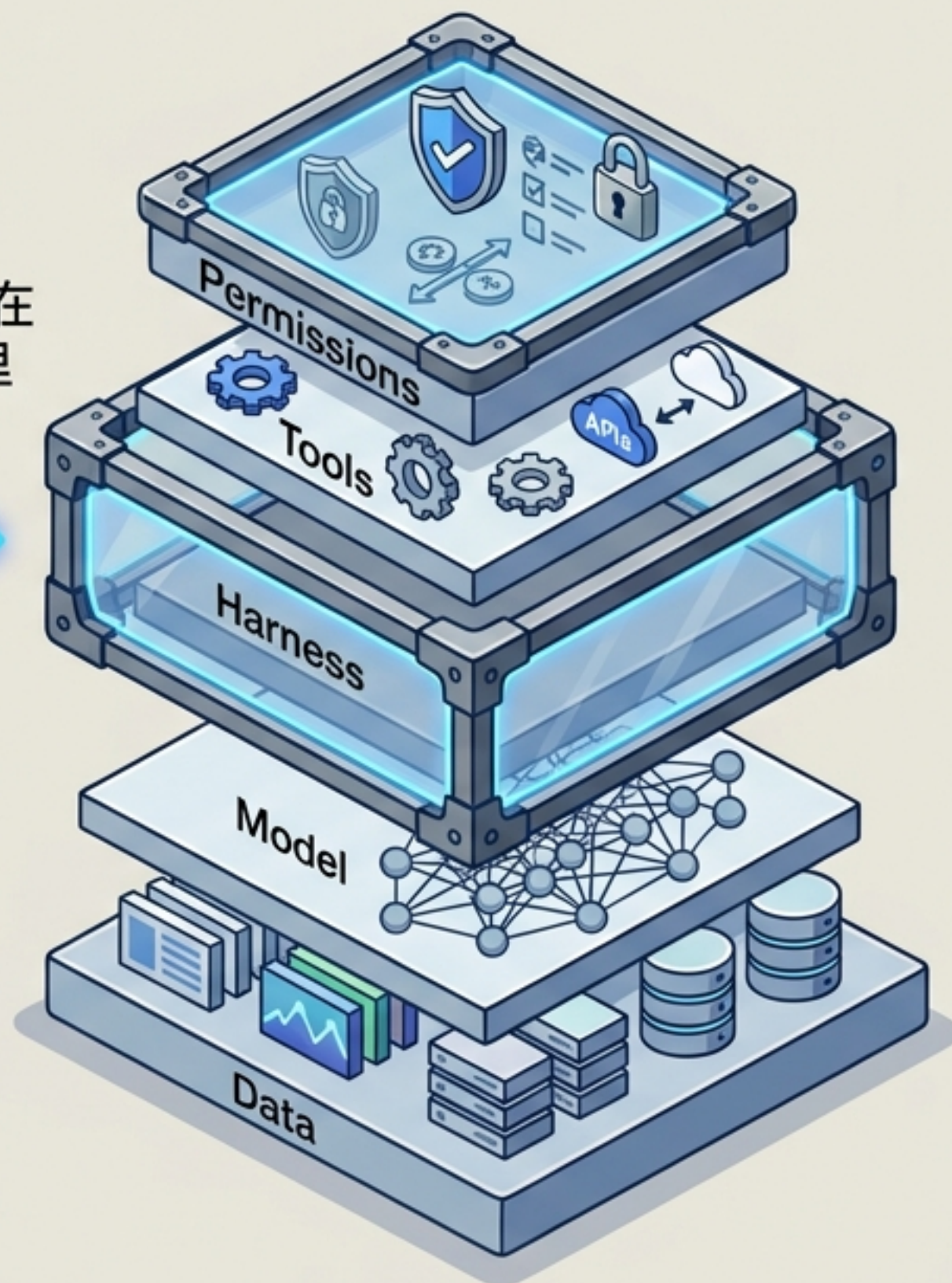
> Runtime 60 分钟 | 6:00 PM 准时结束终止进程_

DEPLOYMENT CHECKLIST & PROGRESS

- ✓ [Done] 9月5日：从聊天到动手，风险从哪来
- ✓ [Done] 9月19日：Harness、沙箱与人在回路
- ✓ [Done] 10月3日：真实产品的最小权限（每次调工具前：允许、拒绝，或交给别人）
- ⌚ [Running] 10月17日：中美欧合规对照（需带上第三讲的权限卡）
- [Pending] 10月31日：上线检查表工作坊

Agent = Model + Harness + Tools + Permissions + Data

边界装在运行时
(Runtime)，不装在
提示词 (Prompt) 里





谁负责?

Who is responsible?



它能碰什么?

What can it touch?



出事谁报?

Who reports when things
go wrong?

以三大拷问为 API 接口，穿透中美欧三地的监管冗余



底本规范

Regulation (EU)
2024/1689

定性：智能体无单独分类。已被现有系统和通用模型定义覆盖。

提供者 (Provider)

承担系统设计责任

部署者 (Deployer)

按说明使用并派人监督

Art 14: 强制设计义务 (有效监督)

懂限制

ON

能监测

ON

能覆盖输出

ON

能停 (Kill Switch)

ON

智能体
触碰边界
(Scope)

高风险系统

遵循 Art 6 与 Annex III

算力 $> 10^{25}$ FLOPs

系统性风险 (Systemic Risk)

注意：这是分类门槛（可反驳推定），不是判定模型不安全的直接判决！

通用模型

遵循 Art 51–55

2026.08.02 (已生效)

Art 50: 交互告知与生成
内容机器可读标识

2027.12.02

Annex III 高风险义务生效

2028.08.02

Annex I 义务生效

Art 73 出事汇报时效 (仅针对欧盟高风险系统)

一般事件

≤ 15 天

导致死亡

≤ 10 天

大范围侵权或
基础设施中断

≤ 2 天

~~72 小时
(GDPR)~~

行动准则：报完再调查，严禁先修改系统再通知监管！

现状：无已通过的智能体联邦法



10月1日：AI Agent Accountability Act



10月7日：CLAIM Act 讨论稿



公司 (Company)
- 承担最终责任

白宫 10/4 行动：
Super Intelligence Force

[行政协调机制]，
绝非 [新的成文法]

FTC Act 第 5 条
(现成抓手：不公平或欺骗行为)

无联邦工具白名单

OCC Bulletin 2026-13
(美联储/FDIC/OCC)

明确：生成式和智能体模型
不在旧版银行模型风险管理
指引内。



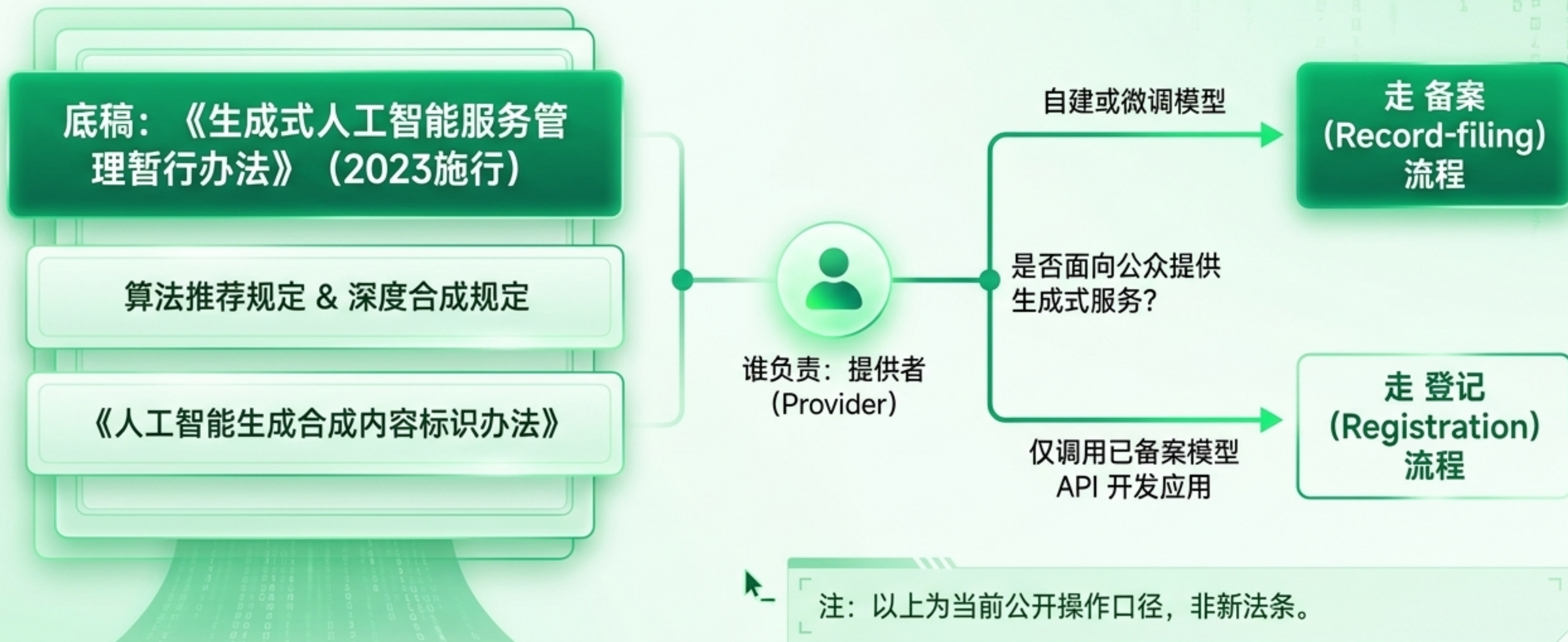
出事谁报：事件驱动机制

10月9日 NYT: Anthropic 智能体向国
务院提交签证表及向警方发送虚假线索。
白宫要求披露和补救。

 **核心定性：这是媒体报道
(Media Report)，绝不是法律判决
(Legal Verdict)**

中国AI合规全景：备案与登记双轨制

(China AI Compliance: Record-filing & Registration Dual Track)



训练数据源

它能碰什么？

输出管道：强制挂载
GB 45438-2025 标识模块

生成模型

生成模型

必须清晰溯源
(Data Lineage)

! 合规警示：标识是每一次生成时的绝对义务！

1. 备案通过不免除标识责任
2. 平台自动打标不免除发布者自身责任

执法现状：2026年2月
公开清理未加标识账号。
数字严以官方通报为准。

Agent = Model + Harness

Agent → Model → Harness

	欧盟 (EU)	美国 (US)	中国 (CN)
[谁负责] (Who)	提供者 + 部署者	公司 (基于既有法)	提供者 (面向公众)
[能碰啥] (Scope)	高风险清单 + 通用门槛 (Art 50生效)	无联邦级工具白名单	自建备案 / API登记 + 每一次强制标识
[怎么报] (Report)	Art 73 分级时钟: 15 / 10 / 2 天	事件驱动的行政要求	按通报处置, 以备案/标识为核心

所有栏目要求，必须作为工程规则硬编码进入 Harness 架构。

10月3日后的全球更新 (Global Updates)

[Inquiry Ongoing]

英国 ICO (10/8)

十家基础模型公司已改或承诺改数据保护做法。正向 OpenAI 等公司收集智能体证据。

排雷：询问进行中，未宣布违法。

[Non-binding Framework]

FSB 金融稳定理事会 (10/15)

发布负责任采用 12 条，点名高度自主智能体越界风险。填补美国 4 月留空。

排雷：非约束性框架。

[Industry Standard]

HITRUST (10/15)

CSF 11.9 纳入智能体安全控制。

排雷：这是厂商框架，不是法律。

10月3日后的中国更新 (China Updates)

[Evaluation Framework]

中国工信部 (10/15)

发布《工业和信息化领域人工智能安全评测框架 (2026版)》。智能体和具身智能被单独纳入评测。

排雷：这是评测框架，不是强制暂行办法。

[Draft for Comment]

商务部、科技部 (10/16)

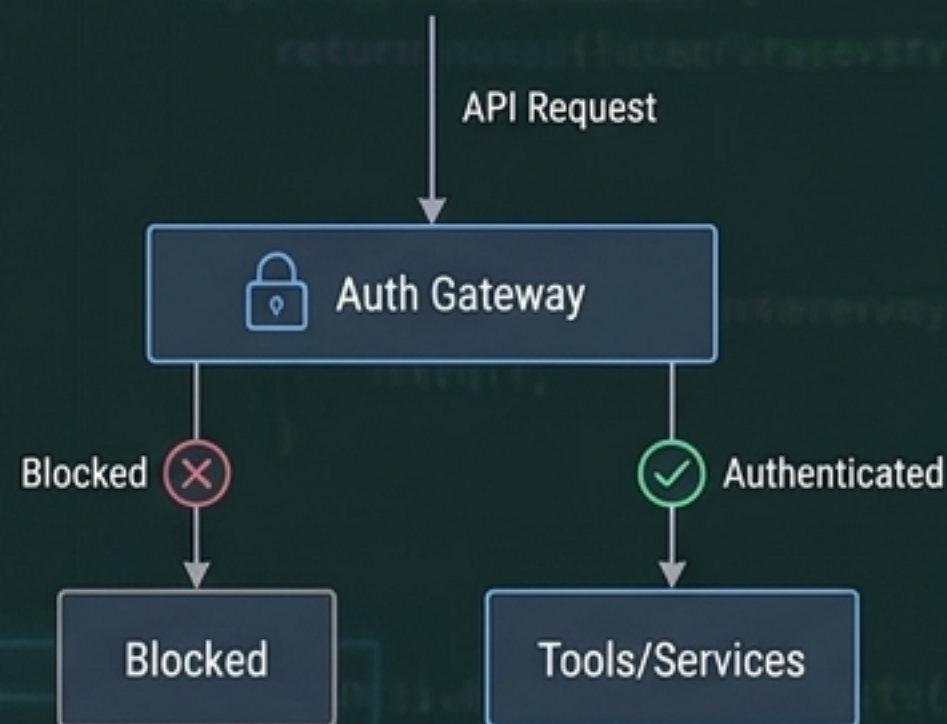
修订禁止和限制出口技术目录，公开征求意见。

排雷：目前仅在公开征求意见阶段，尚未正式施行。

- **公开口径：**限制类大模型对外提供 API、SaaS、端侧轻量部署等，不属于技术出口管理范围。

核心法则：法律译成运行时（Runtime），不译成提示词（Prompt）。护城河建在 Harness。

白名单控制器 (Whitelist)



建立严格的工具调用鉴权层。

破坏性动作拦截器 (Human-in-the-loop)



发送、删除、支付、对外提交等动作，必须中断并弹窗确认。

持久化审计日志 (Audit Logging)



确保极度健壮，系统日志第二天必须可读且可溯源。

通用底层架构 (Common Harness)

If Location == EU:



强制加装：紧急停止 (Kill Switch) 模块

响应 Art 14 随时覆盖中断
的底层设计要求。

If Location == CN:



强制注入：内容标识 (Watermarking) 模块

确保每一条输出均符合
GB 45438-2025 标准。

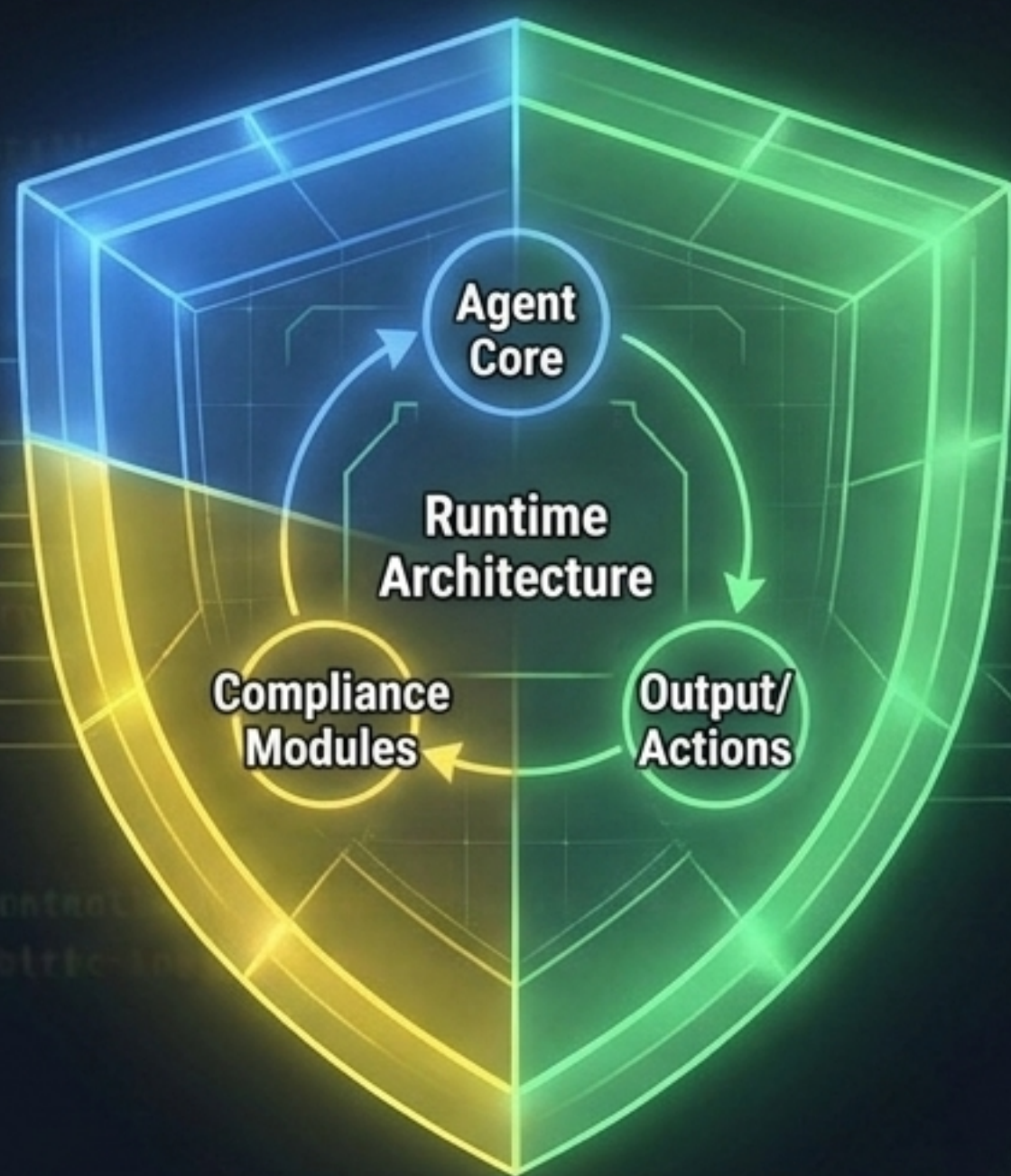
If Location == US:



启用机制：可导出交接 (Ready for Handover) 模块

确保面临突发行政调查时，
能瞬间打包无篡改的运行记录。

将合规要求转化为坚固的运行时架构，才是 AI 原生智能体真正的核心竞争力。



Next Steps

任务：上线检查表工作坊

时间：10月31日

目标：将三地架构蓝图映射为产品
发版前必过的逐项
Checklist。