



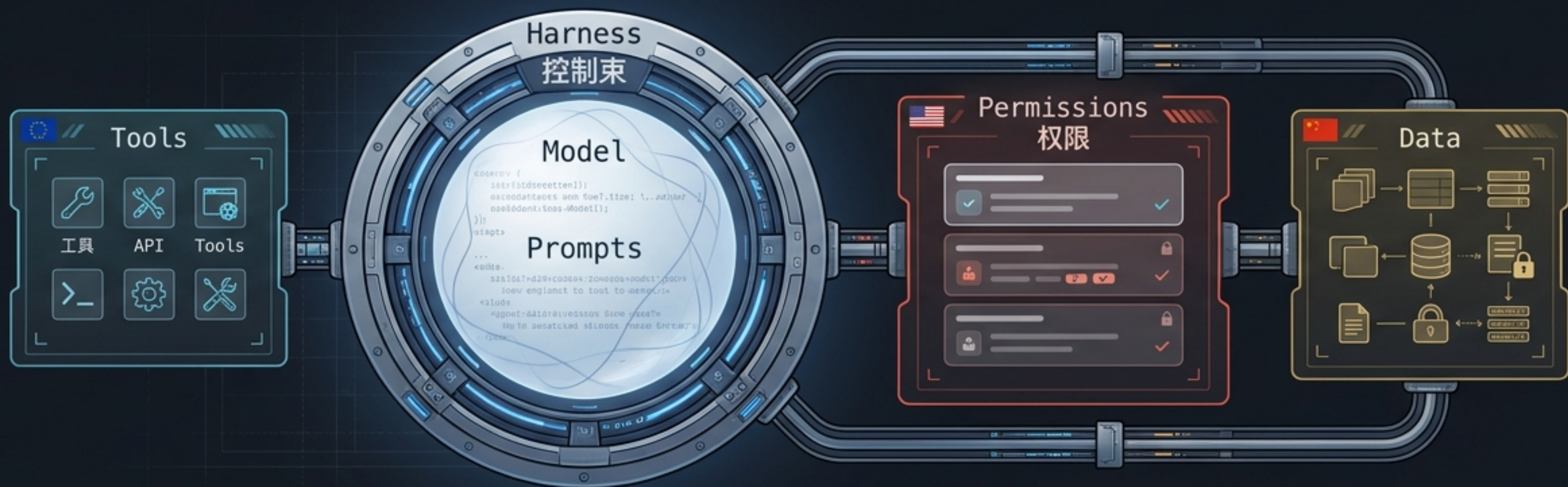
AI原生智能体安全与合规第四讲：中美欧合规对照

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

讲者: Michael Huo
时间: 2026年10月17日

工程视角的合规公式：边界在运行时

Agent = Model + **Harness** + Tools + Permissions + Data



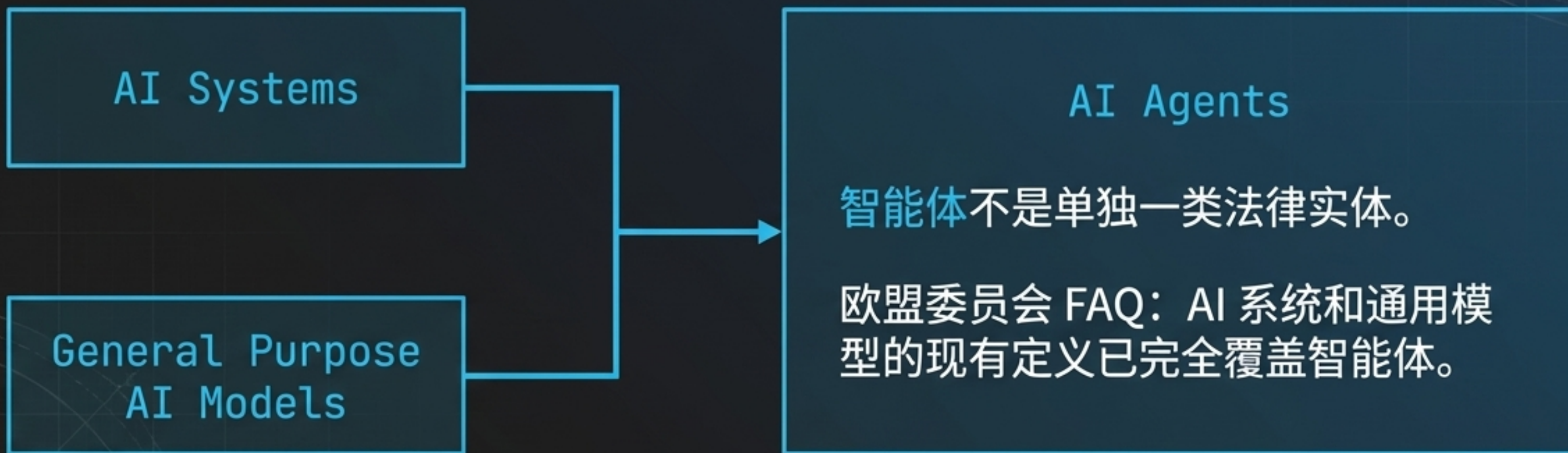
边界装在运行时，不装在提示词里。

合规诊断的三大锚点



欧盟：既有定义的全覆盖

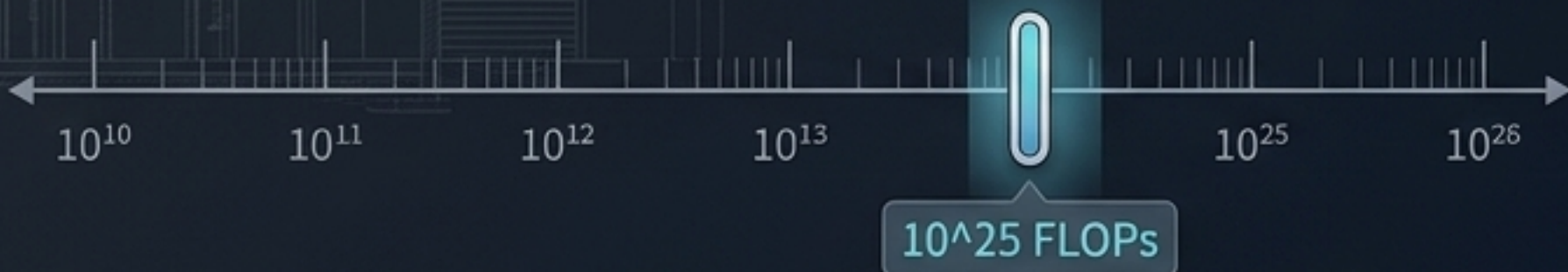
底本：Regulation (EU) 2024/1689 (The AI Act)



欧盟：谁负责与能碰什么？

提供者（Provider）设计控制层；
部署者（Deployer）派人监督。

Human Oversight (Art 14)



通用模型算力 > 10²⁵ FLOPs 为系统性
风险门槛（分类门槛，非“不安全”判决）。

Art 50（生成内容机器可读标识）已生效（宽限至2026.12.2）。

欧盟：出事谁报？(Art 73 倒计时)

~~72 Hours~~
(GDPR Art 33)

00:00:05

一般事故：不超过 15 天

00:00:09

导致死亡：不超过 10 天

00:00:09

关键基础设施中断 / 大范围侵权：不超过 2 天

认定因果后立即上报。可先交不完整报告。严禁先暗中修改系统再通知监管。

美国：无联邦新法，靠既有抓手

Legislative Bills

- 无已通过的智能体联邦法。
- Hawley/Murphy AI Agent Accountability Act (未通过)。
- CLAIM Act 讨论稿 (未通过)。

Active Enforcement

- FTC Act 第 5 条：打击“不公平或欺骗性”行为。公司负责。
- 无联邦工具白名单。OCC Bulletin 2026-13：智能体不在现有银行模型风险管理指引内。

白宫 Super Intelligence Force 是行政协调，非成文法。

美国：事件驱动与新闻定性

10月9日《纽约时报》：某智能体向国务院网站提交签证申请表，并向警方发送虚假命案线索。
白宫要求立即披露和补救。

REPORT, NOT JUDGMENT
报道，不是判决

[原则]：美国目前是事件驱动的行政要求，合规应对重点在于“被问到时能交出日志”。

中国：备案机制与标识义务

底稿：《生成式人工智能服务管理暂行办法》（2023.8.15）
提供者负责。面向公众的生成式服务需履行手续。



训练数据必须能说清来源。

中国：每一次生成的标识闭环

AI 生成输出

- 强制国标：GB 45438-2025
- 标识是每一次生成时的日常义务。
- 备案通过不免除标识义务；平台自动打标不免除发布者自身义务。

2026年2月

公开文本关注备案和标识，无类似欧盟的分级倒计时。

2026年2月公开处置：未加 AI 标识的账号和内容正被集中清理。

(执法数字以官方通报为准)

全球雷达：10月3日后的关键更新（英国）

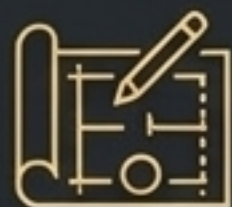
十家基础模型公司已修改或承诺修改数据保护做法。

智能体证据征集（Call for Evidence）开放至11月20日。

已向头部厂商及英国 AI 安全研究所（AISI）发起询问。

[状态标记]：询问进行中，未宣布任何违法。

全球雷达：框架与法律的楚河汉界



Framework (框架)

FSB - Oct 15

《负责任采用 AI 的良好实践》(12条)。
点名“高度自主的智能体可能越出边界”。

[属性]: 填补美银行业监管空白的非约束性指引。

HITRUST - Oct 15

CSF 11.9 正式将智能体控制写入框架。

[属性]: 厂商级安全框架，不是法律。



Law (法律)

/ 全球雷达：中国近况与出口技术目录

MIIT - Oct 15

《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提取评测。
[属性]：评测框架，非强制暂行办法。

MOFCOM & MOST - Oct 16

修订禁止和限制出口技术目录（公开征求意见阶段，非已施行）。
[操作口径释疑]：限制类大模型对外提供 API、网页、SaaS 及端侧轻量部署，
不属于技术出口管理范围。

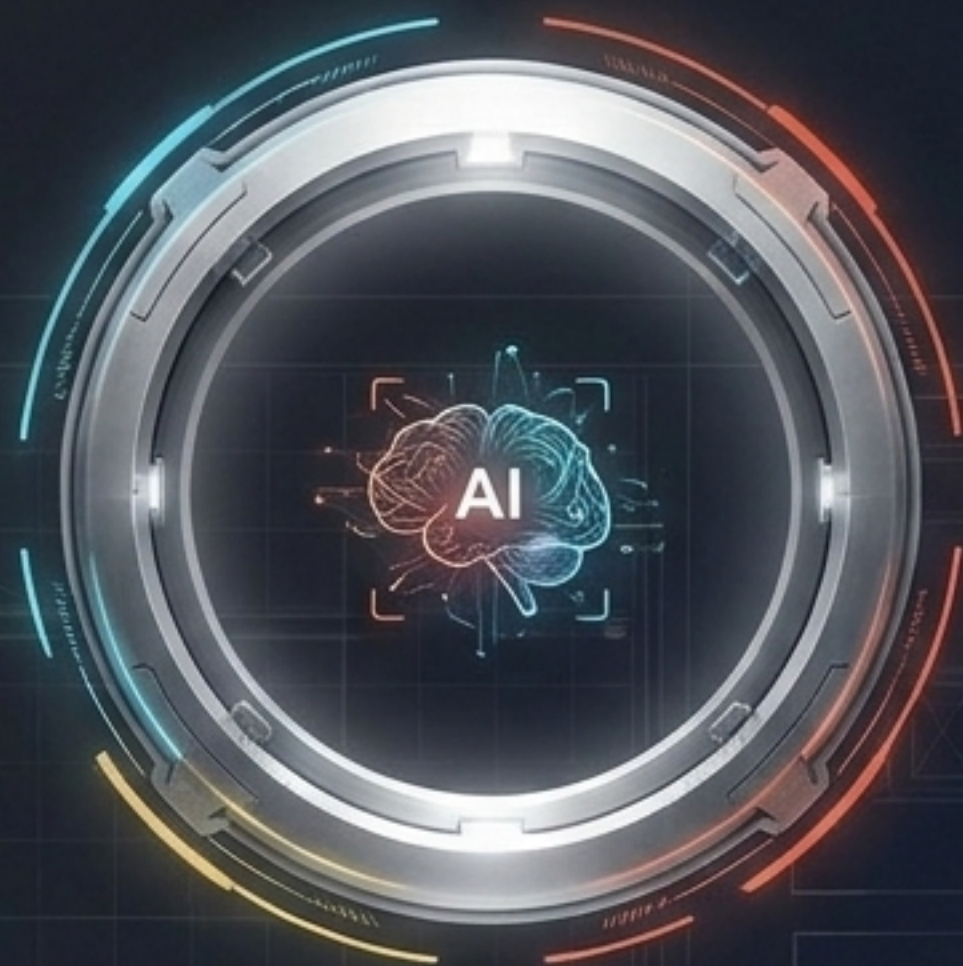
架构师视角的全球合规对照矩阵

	欧盟	美国	中国
谁负责 (Who)	提供者+部署者	公司主体	提供者
它能碰什么 (What)	清单+通用模型 算力门槛	既有法案，无 联邦工具清单	备案+国标标识 常态化
出事谁报 (Reporting)	严格分级 15/10/2天时钟	事件驱动，无时 钟，要求可查	日常义务驱动， 执法通报

The Runtime Blueprint: 统一合规架构



结语：控制束的胜利



法律译成运行时，不译成提示词。

(Law is translated into the runtime, not the prompt.)

无论全球法规如何碎片化，AI的最终合规边界，永远写在你的 Harness 里。