AI改过之后，还是同一个AI吗？

改造前：疫情前的急诊分诊AI

急诊分诊AI系统 v1.3.6

患者ID：JX10086
女 45岁
主诉：发热、咳嗽
症状时长：2天
体温：38.6℃
血氧：95%

分诊建议
低风险
建议候诊区：普通区

风险评分：0.23
低 中 高

模型版本：v1.3.6 训练数据截止：2024-01-31
主要数据来源：本院历史数据（2019-2023）

目前运行稳定，分诊准确率符合预期。

突发疫情

改造后：疫情后的急诊分诊AI

NEW

急诊分诊AI系统 v2.1.0

患者ID：JX10086
女 45岁
主诉：发热、咳嗽
症状时长：2天
体温：38.6℃
血氧：95%

分诊建议
高风险
建议就诊区：发热门诊

风险评分：0.87
低 中 高

模型版本：v2.1.0 训练数据截止：2025-05-15
主要数据来源：本院+外院疫情数据（2024-2025）
模型结构：已调整 特征数量：增加37%

模型更新后，对新冠等症状更敏感了。

高风险AI

分诊AI直接影响就诊顺序与资源分配，属高风险AI系统。

更新/重训

引入新数据、调整模型结构与参数，系统行为发生实质性变化。

同一性判断

改过之后，还是不是原来的那个AI？需要系统性判断。

重新评估

若非同一AI，应重新进行风险评估、验证与备案登记。

1 数据变化是否实质性？

训练数据新增比例超过30%，且包含疫情期间分布变化显著的数据。

新增数据占比约62%，分布发生明显迁移。

数据分布对比
改造前（v1.3.6） 改造后（v2.1.0）
新增数据占比：62%

✓ 是，实质性变化

2 模型结构是否改变？

网络结构、特征工程或学习目标发生了调整。

模型层数增加，特征数量增加37%，损失函数也调整了。

v1.3.6 → v2.1.0

✓ 是，结构已改变

3 行为表现是否显著不同？

在关键指标上差异显著，临床决策建议发生系统性变化。

高风险识别召回率：
从 0.61 → 0.89
误分诊率：
从 7.8% → 3.1%
建议分诊类型变更率：
从 18% → 47%

✓ 是，行为差异显著

综合判断：

改造后的系统已不是原来的AI，应视为新的AI系统，需重新进行全流程治理与合规评估。

风险评估 验证报告 数据说明 模型卡 变更记录 备案材料

重新备案

治理AI，先说清楚它是哪一个

两个案例：身份争议！

案例一：医院分诊AI升级争议

案例二：机场生物识别采购比较

证据能否迁移

医院的原评估证据

- ✓ 准确率评测报告 (v1.0)
- ✓ 临床试运行记录
- ✓ 医生使用反馈
- ✓ 风险评估报告

升级后能否直接用？

- ✗ 模型变更
- ✗ 分诊规则变更
- ✗ 输出结果分布变化
- ✗ 影响临床决策路径

证据不可直接迁移，需重新验证！

机场的原比较证据（A厂）

- ✓ 算法准确率
- ✓ 活体检测能力
- ✓ 设备性能参数
- ✓ 系统集成方案
- ✓ 运维服务保障

能否用于B厂比较？

- ✗ 算法技术路线不同
- ✗ 设备形态不同
- ✗ 部署架构不同
- ✗ 接口与集成方同
- ✗ 运维模式不同

证据不可直接迁移，不可直接比较！

价格与责任

按“同一系统”处理的风险

风险：证据不足 → 决策失当 → 患者受损 → 医院担责

正确做法：
视为新系统 → 重新评估 → 风险可控 → 责任清晰

不恰当比较的风险

风险：不公平比较 → 低价中标 → 性能不足 → 运营风险

正确做法：
分别评估 → 明确需求 → 公平比较 → 价值优先

名字不同，不等于系统不同

- ✓ 看本质：功能、技术、流程、影响是否相同
- ✓ 看证据：能否迁移、是否充分、是否匹配
- ✓ 看责任：决策依据要充分，风险责任才清晰

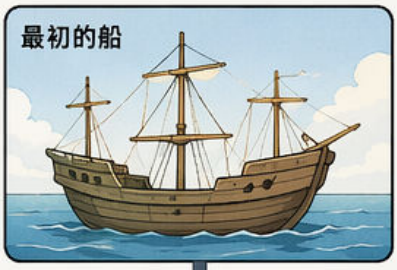
function+：功能之外

一个 AI 神器的“身份”，
不仅来自它能做什么（功能），
更来自它如何被信任、在何处使用，
以及它对世界做出的承诺。

身份
由多维要素
共同构成！

特修斯之船

当所有木板都被替换，
它还是原来的船吗？



形式在变，功能延续，
身份却需要重新界定。

1 技术功能

它能做什么？效果如何？
能力、性能与稳定性
是身份的基础。

2 可信画像

它是什么样的存在？
来源、训练数据、价值观、
透明度与可解释性，
构成可被理解的“画像”。

3 可信水平

我们有多信任它？
通过评估、验证、历史表现
与风控机制，形成可度量的
信任水平。

4 使用语境

在什么场景、由谁、基于什么目标使用？
不同语境下，风险、期望与责任边界都不同，
身份也随之调整。

AI 核心

模型与组件会不断迭代，
但身份由多维要素共同支撑。

模型版本 v1.0 → v2.0 → ···

组件与能力持续演进

- 模型
- 数据
- 工具
- 记忆
- 接口

……

功能在变，结构在变，
唯有“可信 + 语境 + 承诺”
让身份持续成立。

信任来自验证，
身份源于承诺。

可信承诺

- 我会如何使用你的数据
- 我能达到什么边界
- 我出错时如何负责
- 我会如何持续改进

通过公开的承诺与
可验证的机制，
建立长期信任关系。

AI身份 = 功能 + 可信承诺

合规不是
一次性，
而是
持续的循环！

AIA生命周期

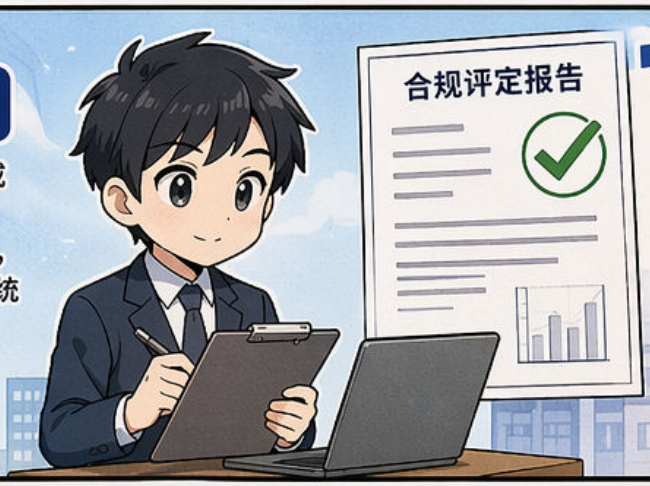
欧盟《人工智能法案》（EU AI Act）下的全生命周期合规管理

EU AI Act
合规认证

1 事前评定

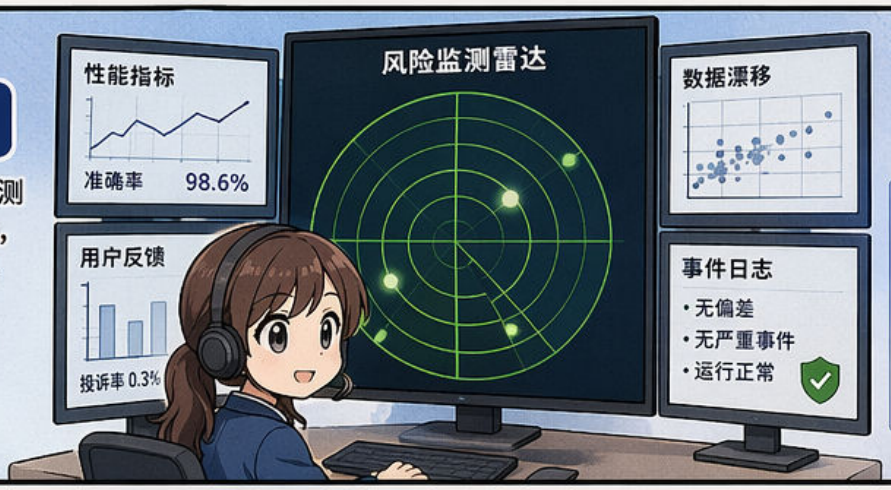
在AIA系统投放市场或投入使用前，进行风险评估与合规评定，确保满足高风险AI系统的各项要求。

- ☑ 风险管理
- ☑ 数据治理
- ☑ 技术文档
- ☑ 透明度
- ☑ 人类监督
- ☑ 准确性/鲁棒性



2 上市监测

系统上市后，持续监测其性能、风险与影响，及时发现问题，确保持续合规。



- 持续收集数据
- 定期评估风险
- 监测系统性能
- 记录与报告

3 重大修改

当系统发生重大修改可能影响风险时，需重新评定，必要时重新通过合规程序。

重新评定流程（如需要）

- 更新技术文档
- 重新风险评估
- 验证与测试
- 通过评定后继续使用

通过

新系统

从设计之初就融入合规要求

AIA已经在做跨时间身份判断

系统的身份、版本、变更与责任，贯穿全生命周期，确保可追溯、可验证、可问责！

跨时间：旧还是新？

同一个 AI 系统，
在时间中演化，如何判断“旧”还是“新”？

	v1.0 2023.01	v1.1 2023.06	v1.2 2023.12	v2.0 2024.06	v2.1 2024.12
模型核心	模型架构A 参数规模：1B	模型架构A 参数规模：1.2B	模型架构A 参数规模：1.5B	模型架构B 参数规模：7B	模型架构B 参数规模：7B
数据流	数据源：D1 覆盖期：2022	数据源：D1+D2 覆盖期：2022–05	数据源：D1+D2 覆盖期：2022–11	数据源：D1+D2+D3 覆盖期：2022–05	数据源：D1+D2+D3 覆盖期：2022–11
公平性防护	规则集 R1 阈值组 T1	规则集 R1 阈值组 T1	规则集 R1 阈值组 T2（微调）	规则集 R2 阈值组 T2	规则集 R2 阈值组 T3（微调）
可解释性透镜	方法：L1 粒度：全局	方法：L1 粒度：全局	方法：L2 粒度：局部	方法：L2 粒度：局部	方法：L2 粒度：局部
检查点门	检查策略：C1 审批人：人类	检查策略：C1 审批人：人类	检查策略：C1 审批人：人类+AI	检查策略：C2 审批人：人类+AI	检查策略：C2 审批人：人类+AI

2023.01 → 2023.06 → 2023.12 → 2024.06 → 2024.12 → 时间

1. 功能相同？

系统是否完成相同的任务，用于相同的目的？

本例：**相同**
（都是信用评估）

2. 画像相同？

对人群的刻画与影响是否在本质上保持一致？

本例：基本一致
（风险画像维度相同）

3. 水平相同？

性能与风险的总体水平是否在可接受区间内？

本例：总体相当
（指标在阈值内波动）

4. 越界＝新系统

是否突破了原有的设计边界或治理边界？

本例：**v2.0 越界**
（模型架构与检查策略变更）
⇒ **视为新系统**

什么是“越界”？

- 改变核心建模范式
  （如架构B替代架构A）
- 引入新的数据源族
  （带来新的人群覆盖与偏差形态）
- 突破原有治理边界
  （如检查策略或审批机制改变）

只要越界，
就可能带来新的风险轮廓！

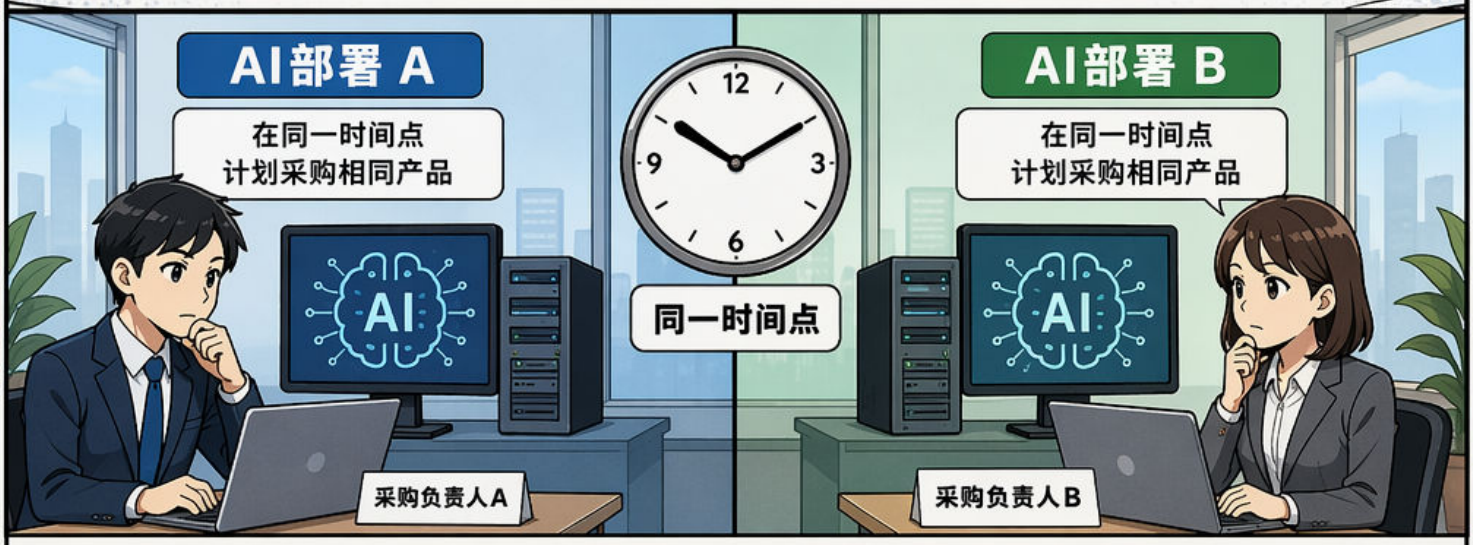
连续演化 ≠ 频繁变更就算新

在边界内的小步迭代与微调，仍属于同一系统的连续演化。

更新不是原罪，越界才断裂

守住边界，持续改进；跨越边界，重新评估。

同一时间：A和B相同吗？



比较维度	AI部署 A	AI部署 B	是否相同？
目的相同 是否为了实现相同的业务目标或管理目标？	提升客服响应效率 降低人工成本	提升客服响应效率 降低人工成本	是
画像相同 服务对象、应用场景、业务流程等是否实质相同？	面向电商平台用户 售前咨询场景 流程：咨询-答复-转人工	面向电商平台用户 售前咨询场景 流程：咨询-答复-转人工	是
水平相同 技术能力、性能指标、功能范围等是否处于相同水平？	准确率≥90% 响应时间≤2秒 支持多轮对话、意图识别等功能	准确率≥90% 响应时间≤2秒 支持多轮对话、意图识别等功能	是
证据可共享 用于证明的资料、测试报告、案例等是否可互认共享？	第三方测评报告 试点案例报告 合规证明材料	第三方测评报告 试点案例报告 合规证明材料	是

任一不满足，就不能套用证据

任一不满足，就不能套用证据

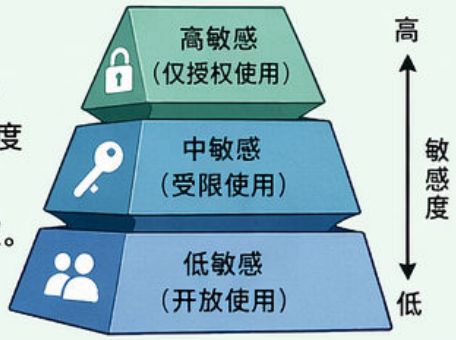
反驳与回应

用途太宽？

“可信数据空间”会不会用途太宽，什么都能装进去？

分层记录

我们采用分层记录：不同用途、不同敏感度的数据，分层管理、边界清晰、授权可控。



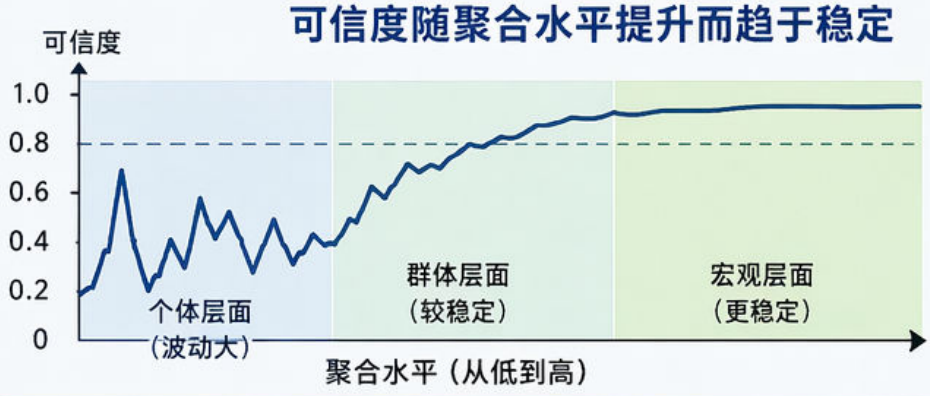
指标波动？

数据可信度指标会不会波动很大，不够稳定？

聚合水平

我们通过聚合提升稳定性：随着聚合水平提高，可信度更稳定，趋势更可靠。

可信度随聚合水平提升而趋于稳定



不是记录一切，而是记录边界

最终脑图

功能

它能做什么
能力可描述、可验证

- 自然语言对话
- 数据分析与总结
- 文案生成
- 代码辅助

……

功能清单清晰
边界明确可查

可信度

它是否可靠
来源可查、评估可见

提供方：	XX科技有限公司
模型名称：	XX-Chat-Pro
版本：	v2.1.3
训练数据截至：	2024-12-31
安全评估：	通过
合规认证：	已认证
评估结论：	高可信 ★★★★★

来源透明
评估有据

AI身份

唯一ID：AI-7f3e9c2a
版本： v2.1.3
提供方：XX科技有限公司

时间

它何时有效
状态随时间变化可追溯

状态记录

- 2024-01-01　上线发布　v1.0.0
- 2024-06-15　安全评估更新　结果：通过
- 2024-12-31　模型升级　v2.0.0
- 2025-05-22　当前版本　v2.1.3

时间线清晰
变化可追溯

责任链

出了问题谁负责
采购到使用全链可追踪

采购方	提供方	部署方	使用方
需求提出 合同签订	模型开发 服务提供	系统部署 运维管理	业务使用 结果输出

全链留痕，责任清晰
问题可定位，追责有依据

链路完整
责任可查

让同一和不同
都有证据可查