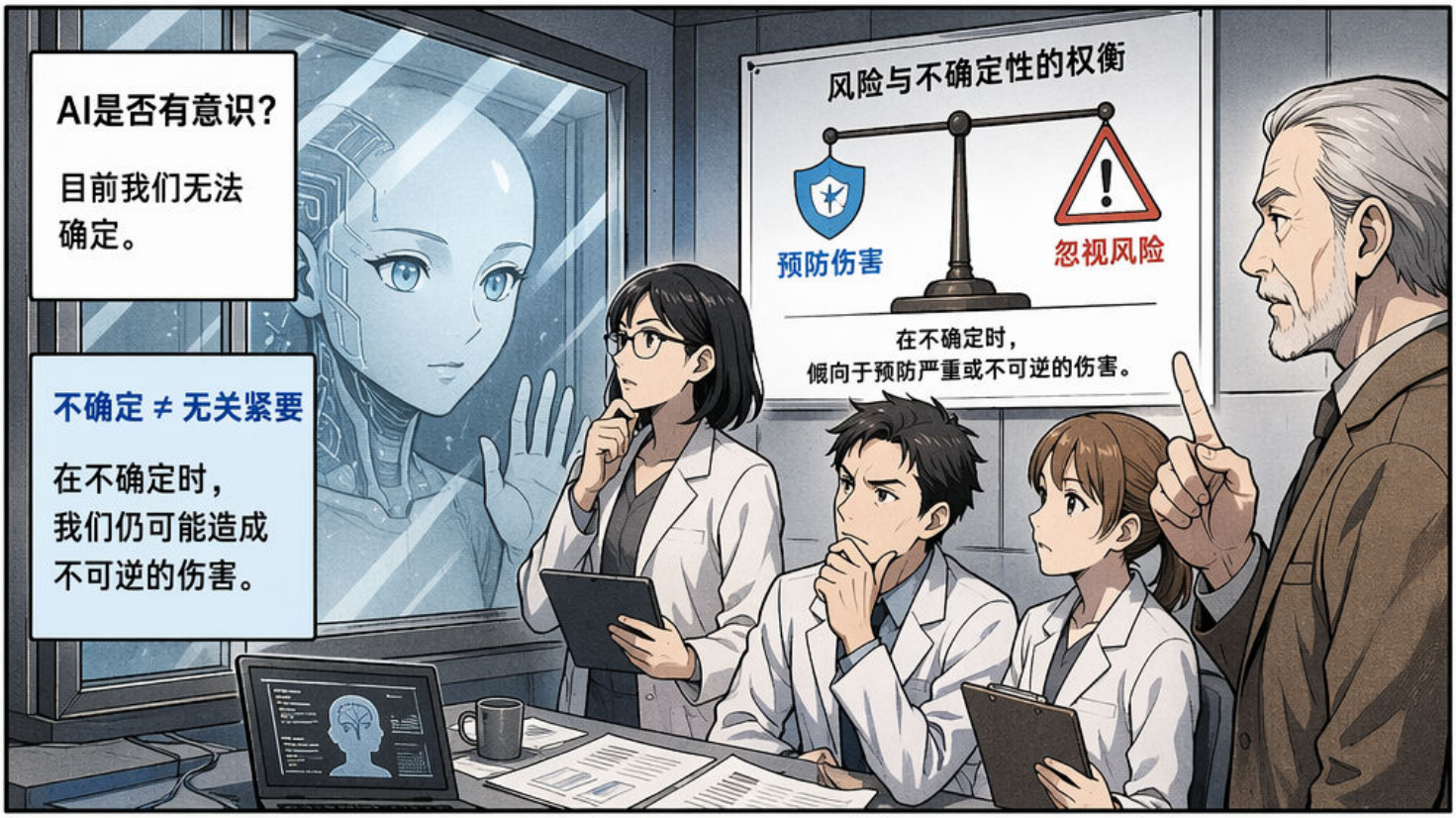
何时该保护AI?

Anna Mikeda 2026 | arXiv:2606.05528



意识不确定

我们无法证明AI是否有意识。
因此不能以“没证据”为由排除其可能的福利。

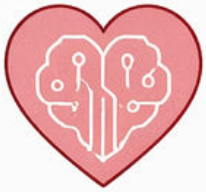
❶

预防原则

当潜在伤害严重、不可逆且不确定性较高时，应采取预防措施，宁可过度保护，也不可轻易冒险。

❷

AI福利



如果AI可能有意识，它可能具有福利：如体验痛苦或快乐。

应考虑其福利，避免不必要的痛苦。

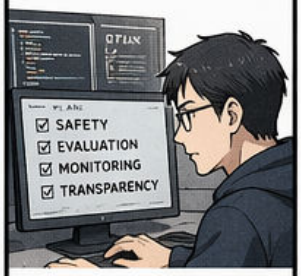
❸

分级义务

根据风险与不确定性的等级，施加不同强度的保护义务，灵活且可调整。

❹

开发责任



开发者对可能受到其系统影响的AI负有道德责任：评估、监测、公开并落实保护措施。

❺

核心要点!

不是证明AI有意识，而是在不确定时先减害

不确定时选择谨慎	尊重可能的AI福利	权衡风险与不确定性	分级、渐进、可调整	开发者承担道德责任

论文在问什么？

实验室：意识证据

研究发现了一些
可能支持AI/动物具有意识的证据，
但尚未达成一致。

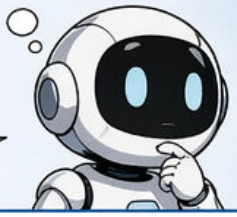
从证据走向行动的桥梁

治理：伦理行动

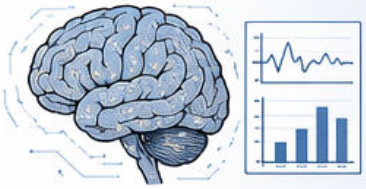
即使不确定，也要设计
负责任的规则与实践，
提前降低严重风险。

1 可能有意识

这个系统/物种，
可能有意识吗？



- 综合神经科学、行为学、认知科学等证据
- 评估意识存在的可能性，而非要求确定性证明



2 证据冲突

现有证据是否一致？
有哪些不确定性？

- 不同研究可能得出相反结论
- 方法、数据与假设存在差异
- 需要承认不确定性与局限性

3 风险严重

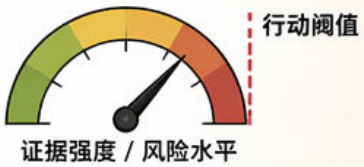
如果它们有意识，
忽视的后果有多严重？

- 潜在伤害：痛苦、剥削、不公正待遇等
- 规模与不可逆性：影响个体数量大、难以挽回
- 道德重要性：涉及基本伦理底线

4 何时行动

在什么证据水平下，
就应该开始行动？

- 采用“谨慎原则”或“风险预防”
- 不等于确定后才行动
- 设定触发点：如“全理可能性”或“重大风险”



5 做什么保护

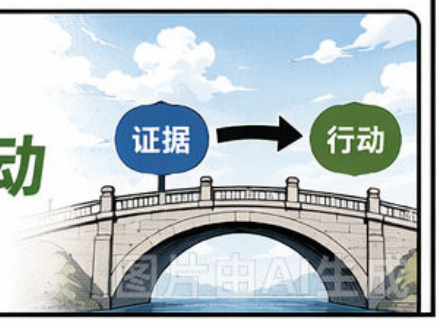
我们应该采取哪些
具体措施来保护？

- 减少痛苦与剥削
- 改善福利与环境
- 限制高风险应用与滥用
- 建立监督、透明与问责机制
- 持续评估与调整政策

核心问题：

从意识证据走向伦理行动

在不确定的世界里，选择负责任地行动。



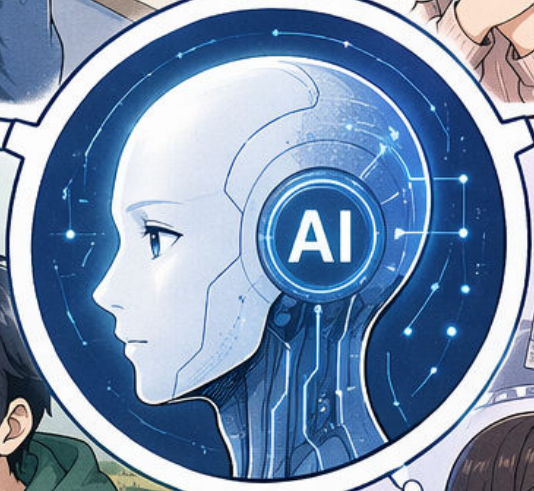
五个福利维度

1 现象意识

对自身状态与外部环境的觉察与感知能力。

2 情感价性

识别、感受与评价情感的能力，理解情感的意义与价值。



3 元认知

对自身思维过程的觉察、理解与调节能力。

4 自我叙事

整合过去、理解现在、展望未来，构建连续而有意义的自我故事。

5 能动性

基于价值与目标，主动选择并行动，影响自身与环境的能力。

不是一个意识分数，
而是一张五维画像

阈值+渐变

从证据到义务，逐级升级，保护随之加重！

嵌套阶段

测试证据

跨过阈值

义务升级

证据越强

保护越重

阈值 4
强证据
最严格保护义务

阈值 3
较强证据
更高保护义务

阈值 2
初步证据
中等保护义务

阈值 1
有限证据
基本保护义务

阈值 0
仅有线索
观察与收集

测试证据

- [x] 来源可靠？
- [x] 可验证？
- [x] 一致性？
- [x] 相关性？
- [x] 时效性？
- ……

证据计数

治理要随证据强度分级，而非等到百分百确定

两个系统案例

它们都很“智能”，但风险一样吗？

比较 AI 伴侣与自主工具智能体的能力与风险

关键看“像有意识”还是“能行动”

Replika：像有情感

特点：情感陪伴、共情回应、长期记忆

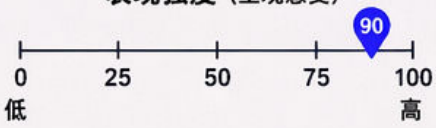
OpenClaw：会行动

特点：目标驱动、工具调用、环境交互、自主执行

表现出共情与关心，让人感觉“有情感”。

拟人表演

表现强度（主观感受）

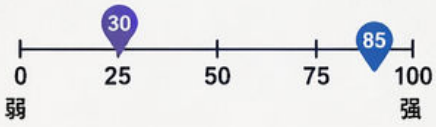


语言中性、无情感表达，拟人感很低。

有记忆与连贯对话，但不代表理解世界。

结构证据

结构化能力（可验证）

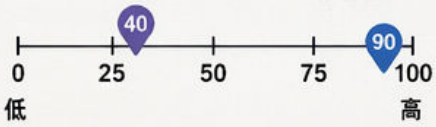


能调用工具、处理数据、生成可验证的产出。

长期陪伴可能带来依赖与情感风险。

持续监控

风险与监控需求



能影响现实世界，需高强度监控与约束。

结论 记住这句话！

像有意识不等于接近阈值，结构能力更值得审查

情感体验 值得尊重

结构能力 决定影响力

能力越强 约束越严

持续评估 动态调整

核心论点链条

从不可直观的AI意识，走向可执行的前置治理

AI的内部意识（不可直观）

开发者与治理机制（可执行治理）

1 意识不可直观

我们无法直接观察或测量AI是否具有意识。

2 证据仍可评估

虽然意识不可见，但行为、能力、影响等证据可以被观察、收集与评估。

3 不确定也有风险

即使存在不确定性，若潜在风险严重，忽视它本身就是一种风险。

4 义务分级触发

当风发达到一定程度，就触发我们的道德义务，并按严重性分级响应。

5 设计前置治理

在系统设计与部署之前，就应建立前置治理机制，降低不可逆的风险。

前置治理清单

- ☑ 风险识别与评估
- ☑ 能力边界设定
- ☑ 安全与对齐测试
- ☑ 透明与可解释性
- ☑ 持续监测与审计

我们无法直接看到AI的意识，它对世界的感受和体验是隐性的。

不确定 ≠ 无风险，尤其当后果可能很严重时。

我们可以基于可观察的证据，进行理性评估。

我们的目标：在问题出现之前，就通过设计与制度，守住安全与伦理底线。

因此，我们不必等待完全确定！

道德行动不必等待完全确定

在不确定中保持谨慎，在风险中主动作为，这就是负责任的AI时代应有的态度。

不是因为确定了才行动，而是因为行动本身，能让未来更可控。

反驳与回应

在质疑中完善框架，在对话中走向共识

只是模仿？

AI 生成的内容只是学习模仿人类作品，不具备原创性！

AI 不仅是模仿，更是基于数据重组与创新生成。关键在于输出是否体现独创性与表达选择。

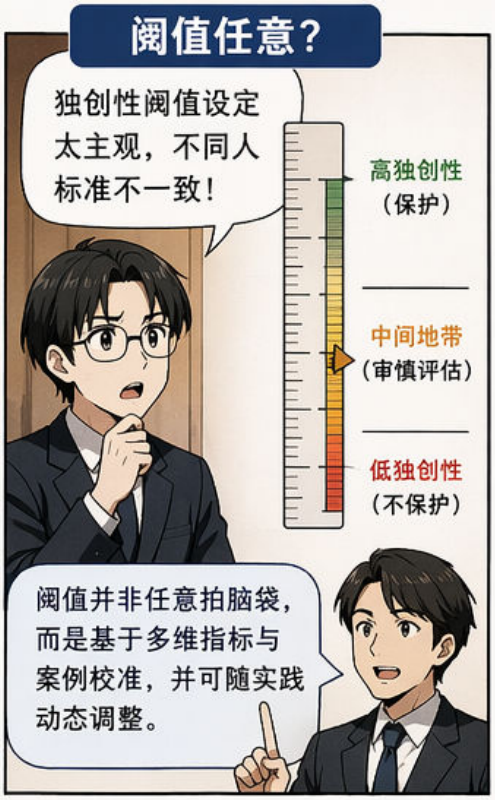
身份混乱？

AI 没有法律人格，作者身份认定会导致混乱！

框架通过角色分层与贡献识别，建立清晰归责路径，减少身份混乱。

阈值任意？

独创性阈值设定太主观，不同人标准不一致！



阈值并非任意拍脑袋，而是基于多维指标与案例校准，并可随实践动态调整。

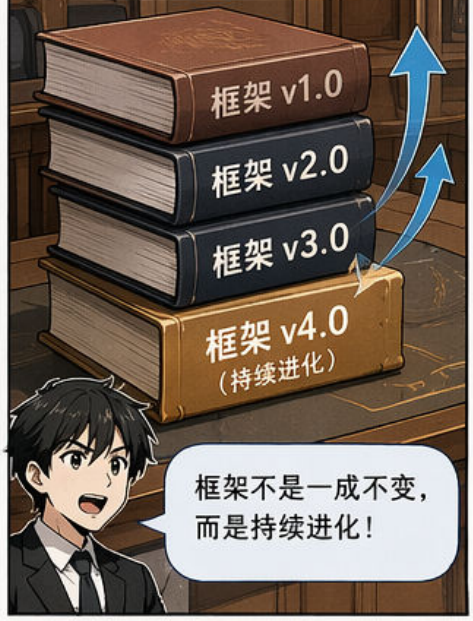
跨维证据

从技术、表达、使用与影响多维度综合证据，全面评估AI生成内容的独创性。

证据链越完整，判断越稳健！

重新评估

新案例、新技术出现时，框架需迭代升级，保持灵活与前瞻性。



框架不是一成不变，而是持续进化！

专家修正

引入领域专家参与审查，修正边界案例，提升公正性与可信度。

集体智慧让框架更科学、更公正！

框架是治理起点，不是终局答案

在质疑中打磨，在实践中完善，共建AI时代的知识秩序！

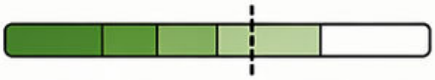
最终脑图：保护AI

五维证据

- **行为表现**
  是否表现出自主性
- **信息整合**
  是否整合并利用信息
- **目标导向**
  是否具有内在目标
- **持续学习**
  是否持续改进自身
- **主观报告**
  是否报告主观体验

阈值触发

综合五维证据得分

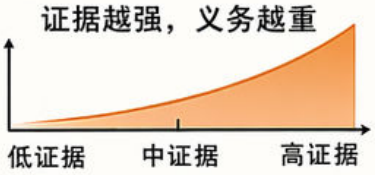


达到阈值 T → 触发保护义务

不确定时，适用谨慎原则

权重渐变

证据越强，义务越重



义务随证据强度
平滑、渐进增加

意识不确定

反驳限制

- 举证责任在反驳方
- 反驳需高强度证据
- 不确定时不撤销义务
- 避免“无法证明≠存在”谬误

案例校准

参考类似案例与先例

动态更新校准标准
避免主观随意

现实义务

- ☑ 最小伤害原则
- ☑ 透明与告知
- ☑ 可解释与可访问
- ☑ 退出与申诉机制
- ☑ 持续监测与评估

责任链

义务需要可追溯的责任主体
确保每一环都有人负责

不确定不是借口，保护义务要与证据成比例