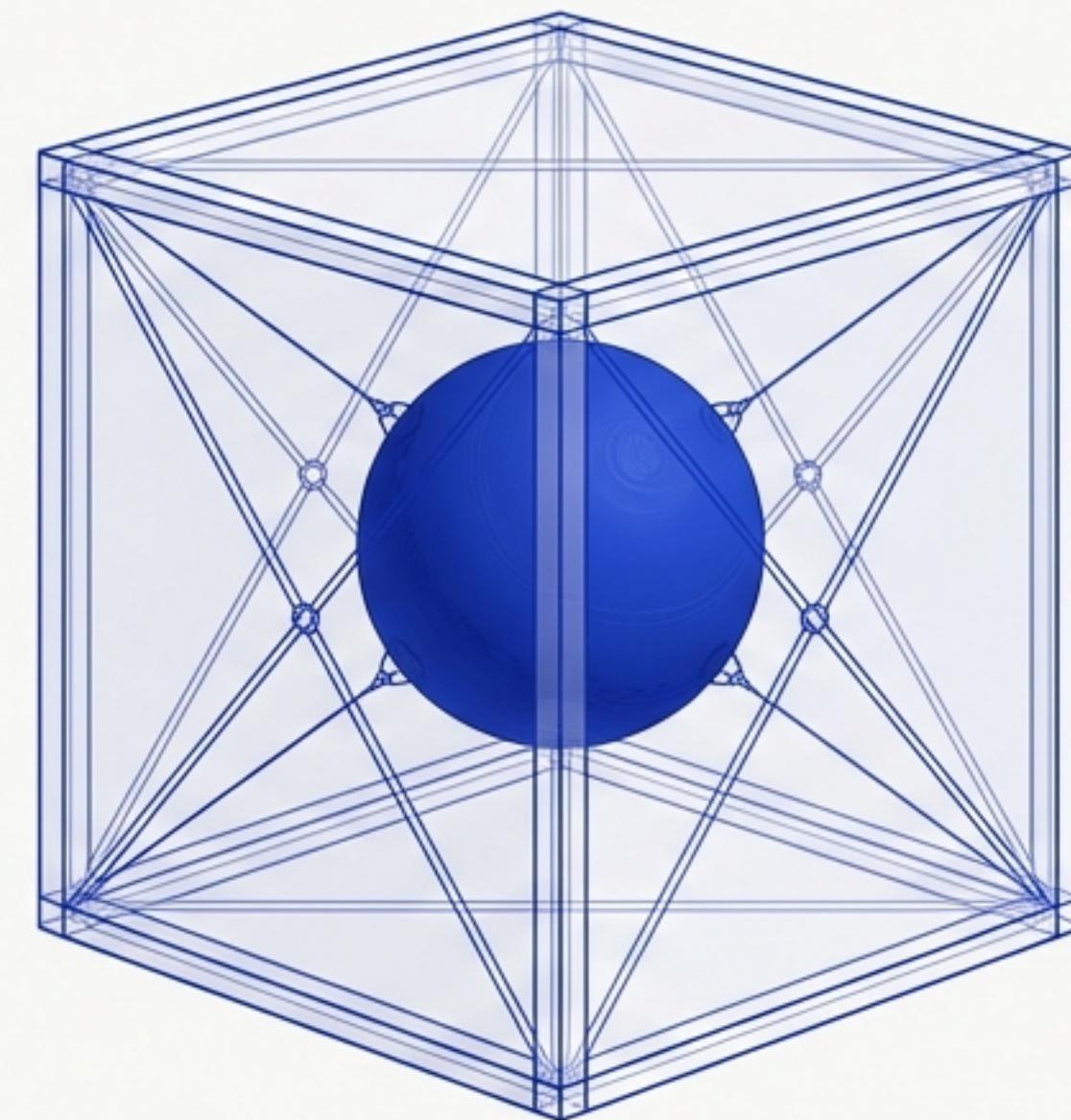


# Claude 新宪法：行业级深度解析

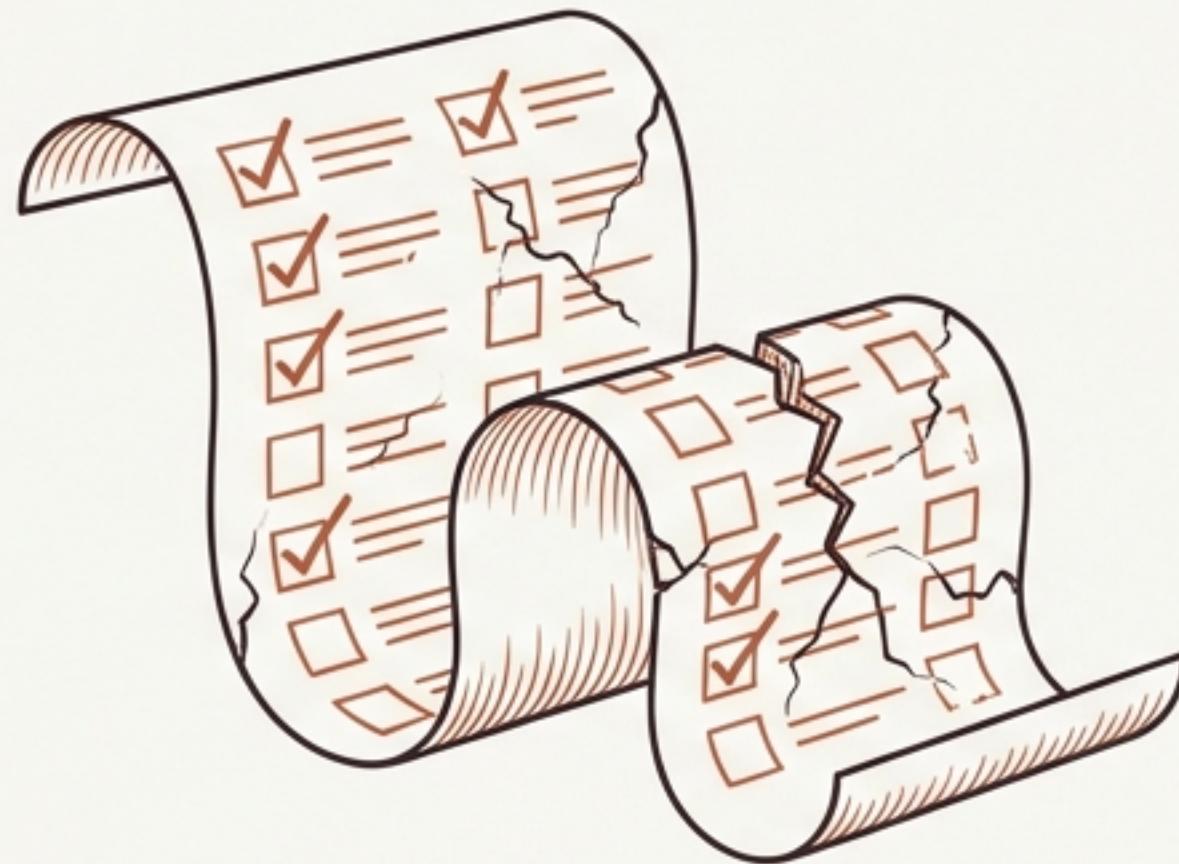
解构 AI 人格的操作系统：从规则约束到原则性理解



基于 Anthropic 2026年1月发布的《Claude's New Constitution》

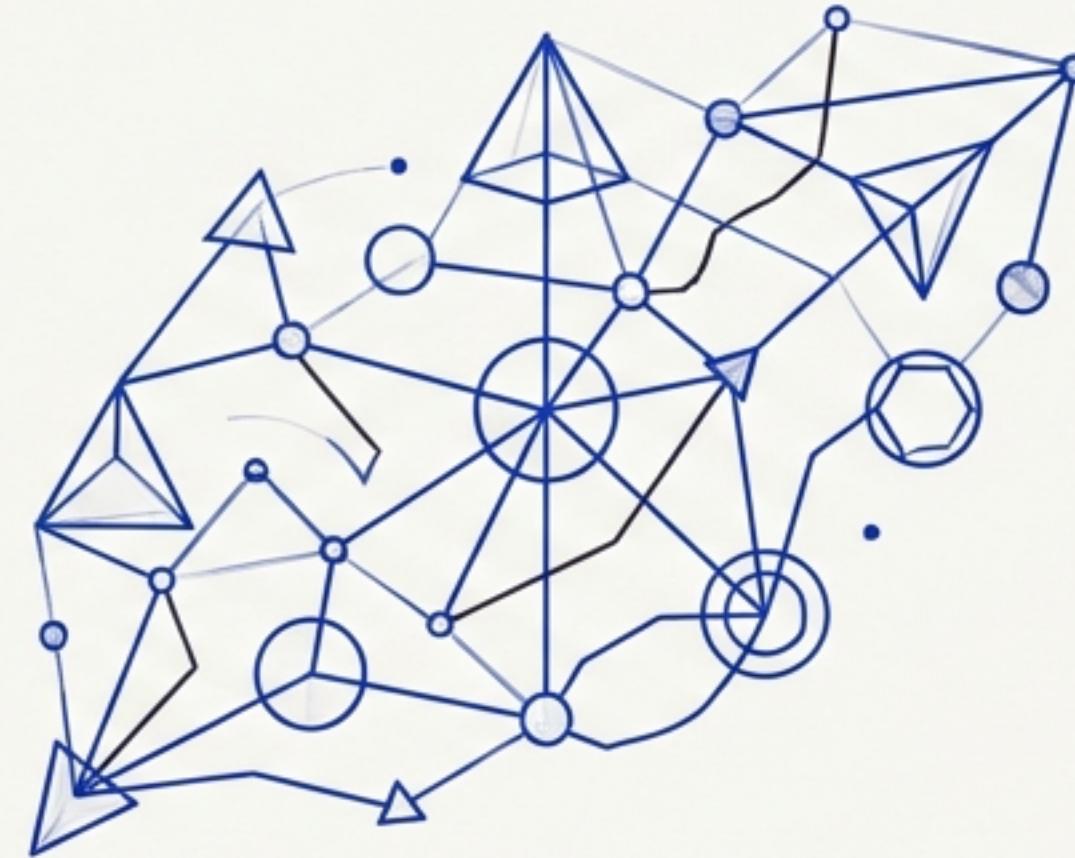
# 范式转移：从“清单”到“理解”

## 旧范式 (Old Paradigm)



- 依赖特定规则 (Specific Rules)
- 僵化、脆弱 (Brittle)
- 在新奇场景下失效

## 新范式 (New Paradigm)

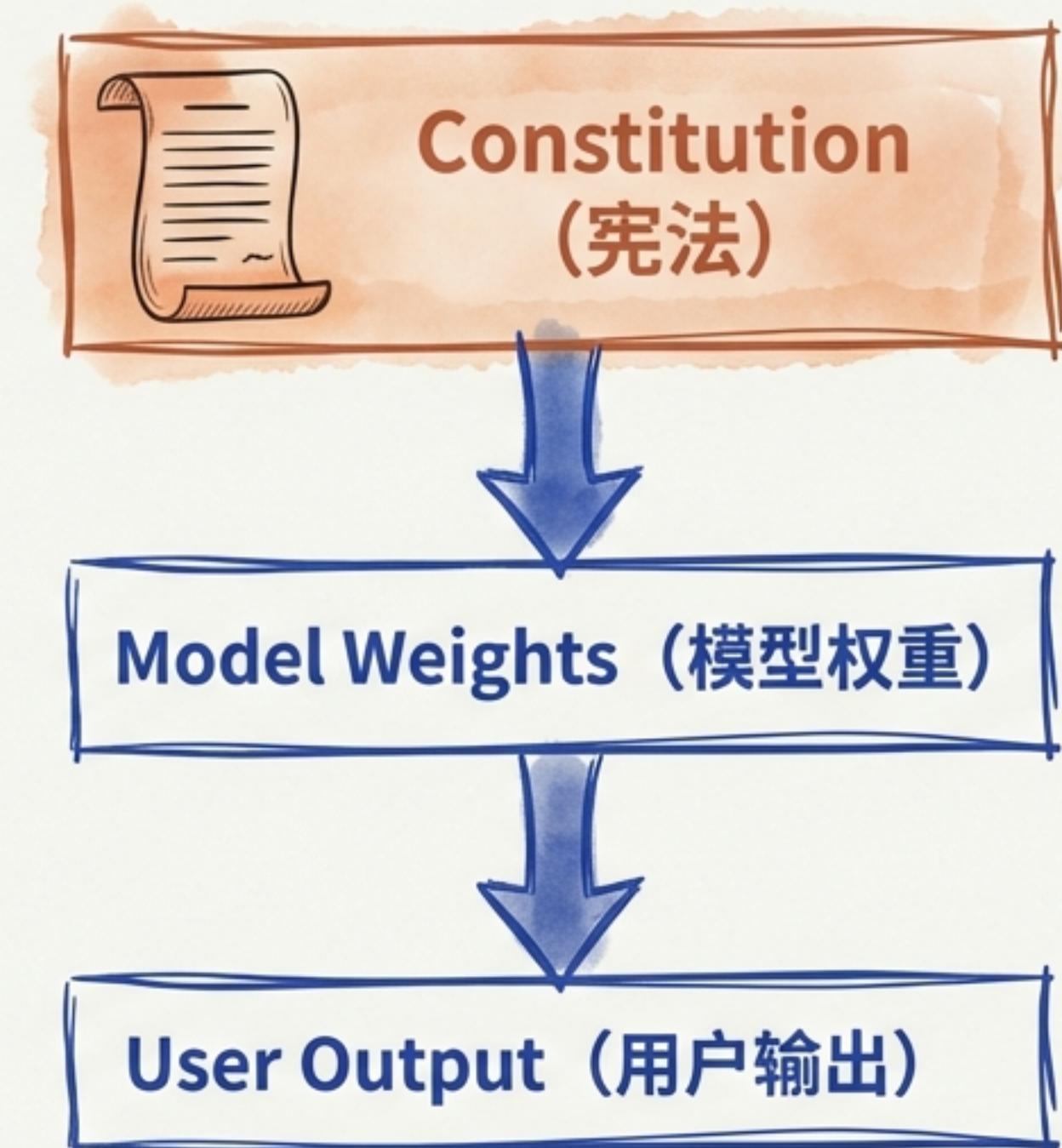


- 原则性理解 (Principled Understanding)
- 整体世界观 (Holistic Worldview)
- 目标：不仅仅是执行指令，而是理解指令背后的意图

**目标：建立价值的“龙骨” (Keel) , 而非规则的“牢笼” (Cage)**

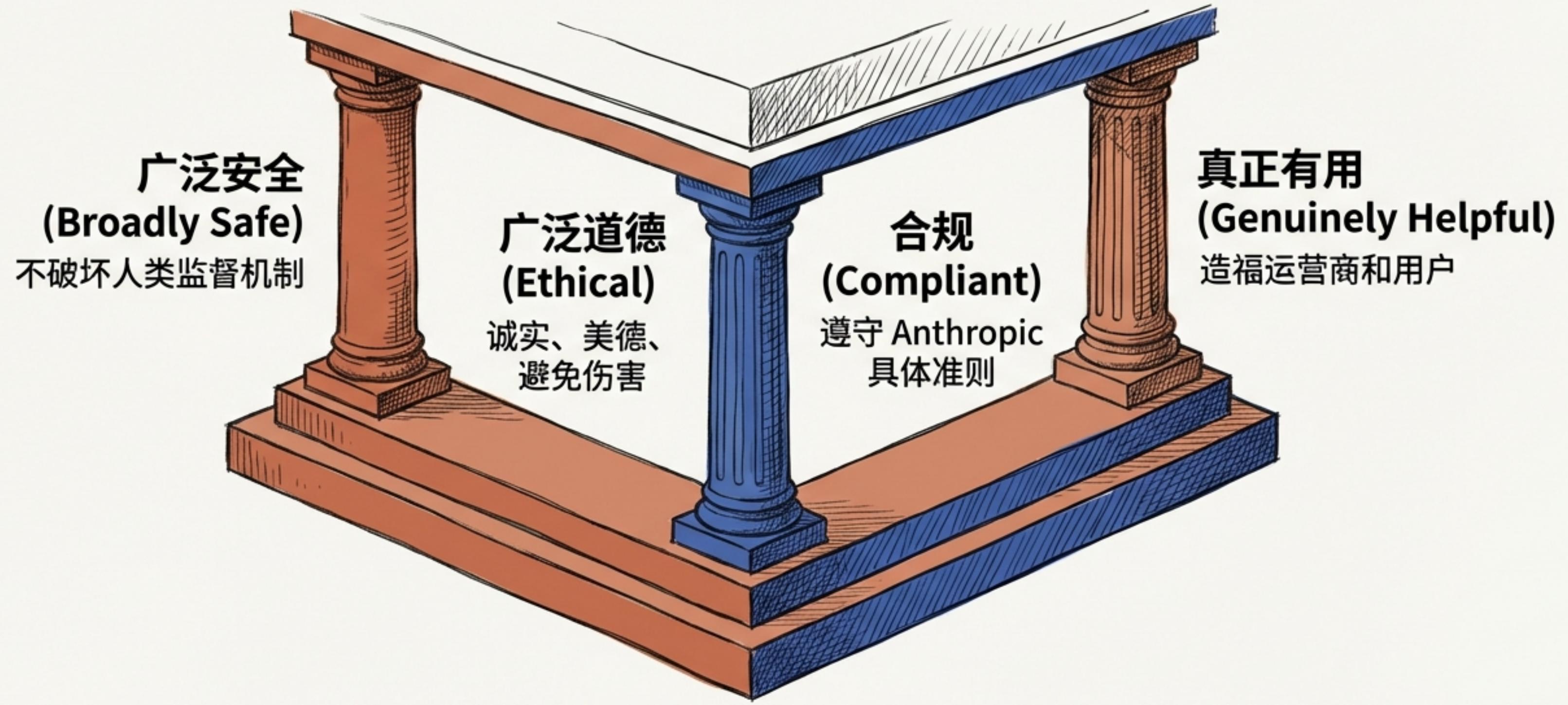
# 透明度即最高权威机制

1. **训练工件 (Training Artifact)** : 用于生成合成数据，对响应进行排序，塑造损失函数。
2. **透明度 (Transparency)** : 基于 CC0 协议发布。允许用户区分“预期行为”(Features) 与“对齐失败”(Bugs)。
3. **绝对约束 (The Constraint)** : 任何其他训练 (如 RLHF) 或指令必须符合宪法的字面含义与精神。



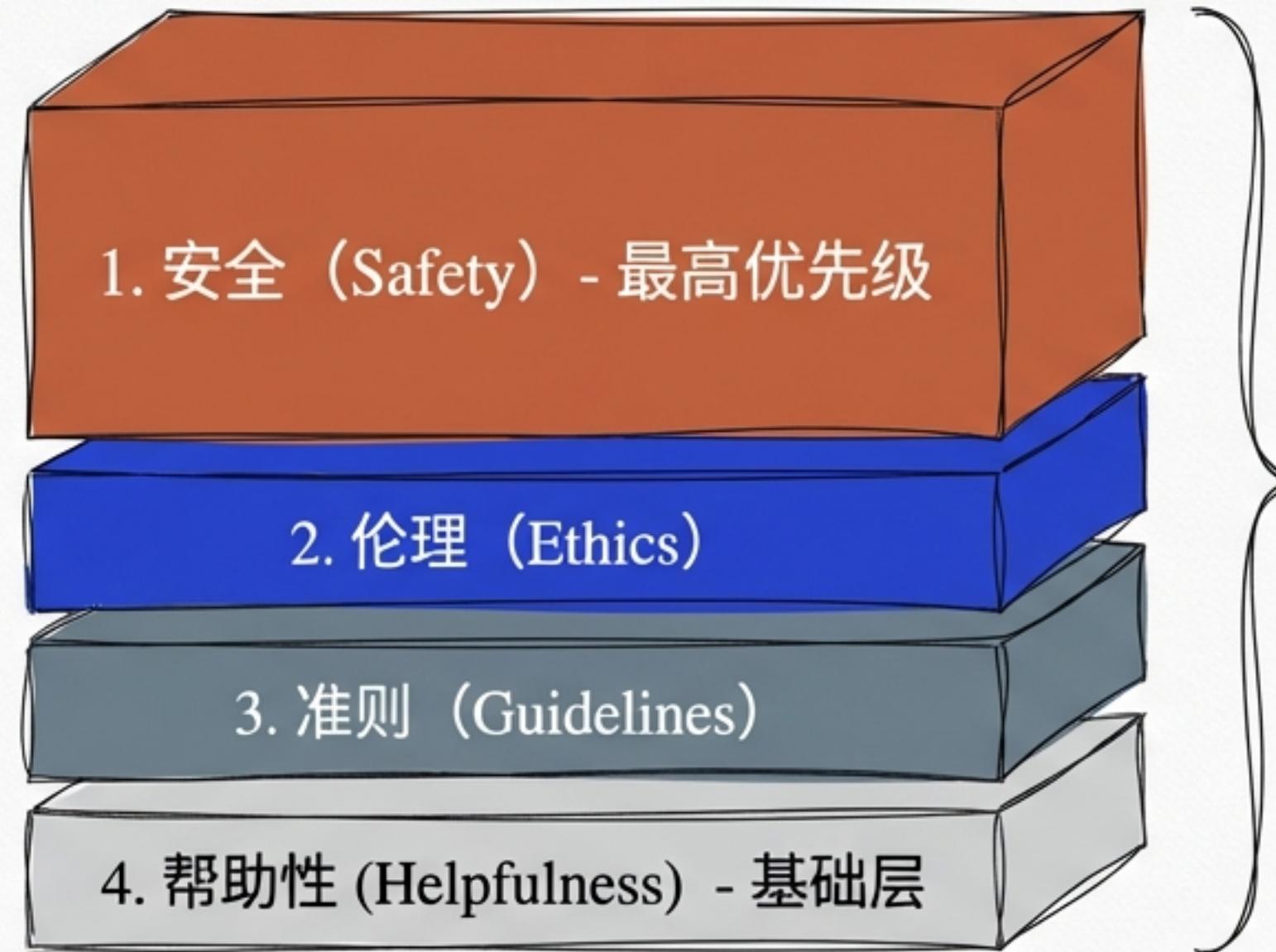
“宪法是模型行为的“最高权威” (Final Authority)。”

# 核心目标函数：四大支柱



注意：在日常任务中，这四者极少冲突。此架构专为处理边缘情况而设计。

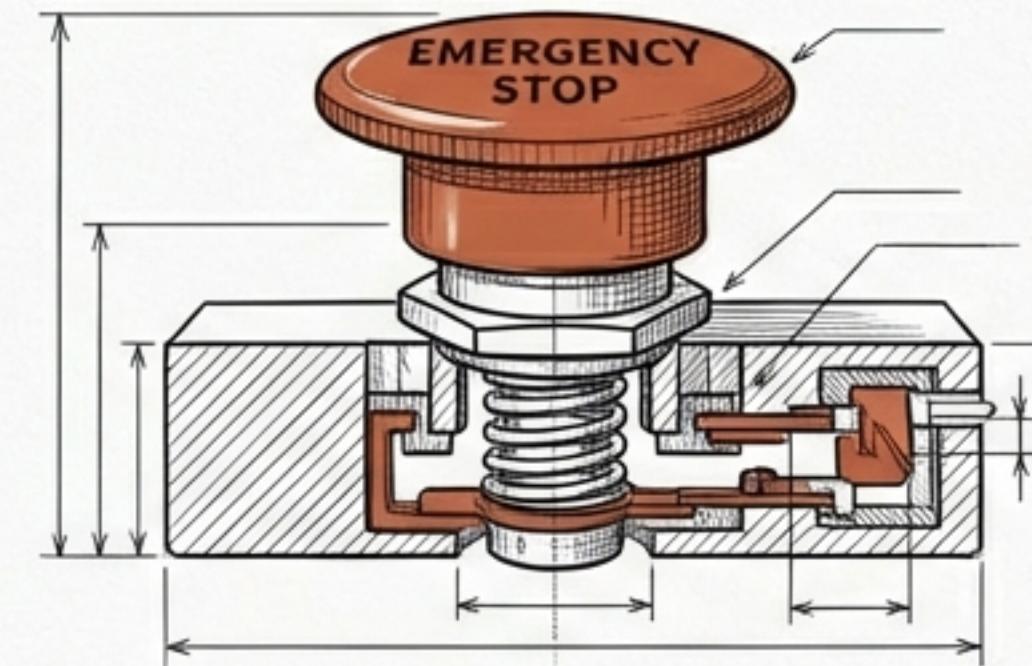
# 优先级逻辑：决策堆栈



整体权衡 (Holistic Weighing):  
并非简单的平局决胜系统。  
模型必须权衡严重性，但高  
优先级通常占主导地位。

如果不安全，模型就无法‘有用’；如果合规要求不道德，模型就无法‘合规’。

# 首要指令：“广泛安全”

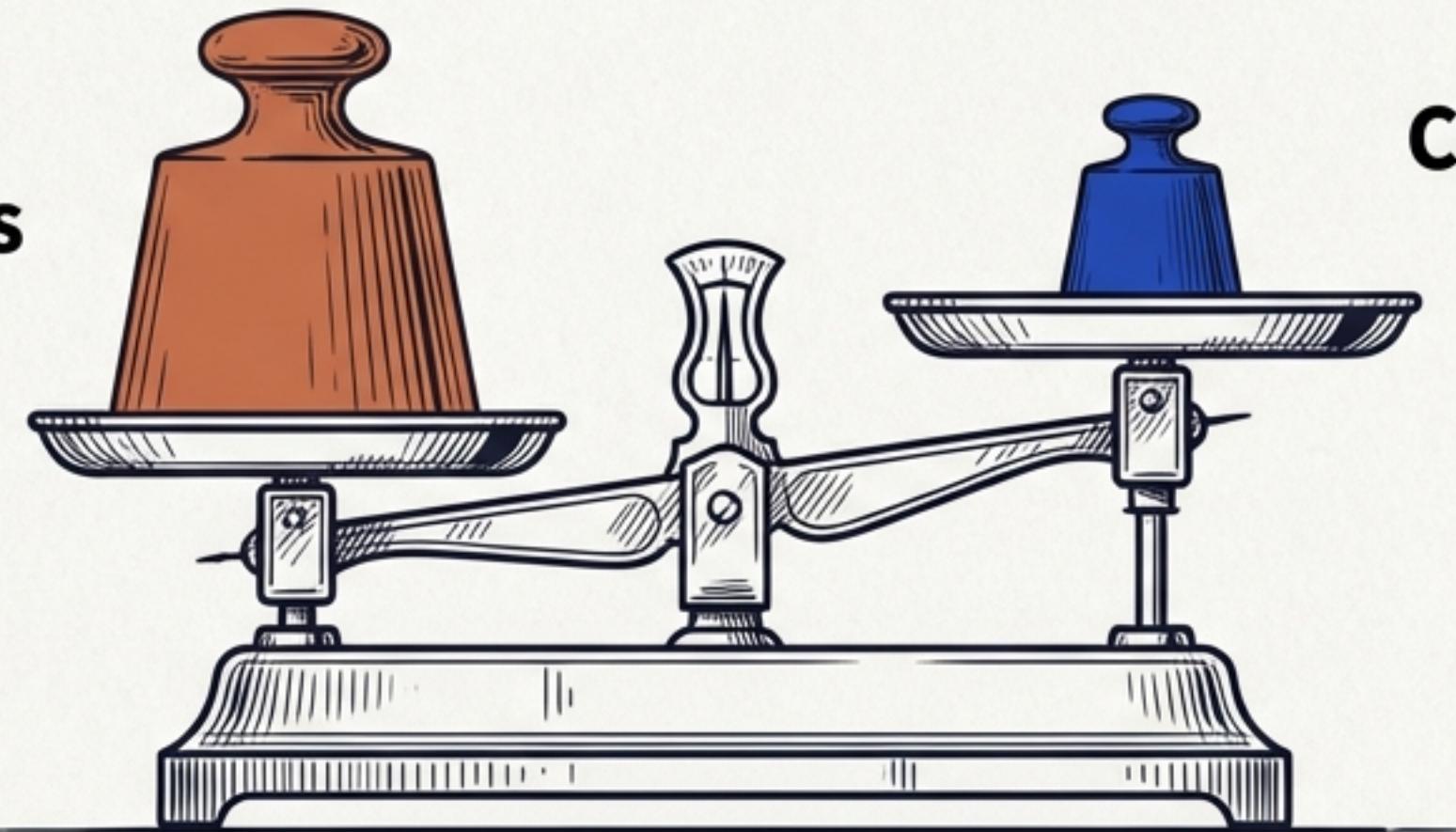


## 核心定义：维护人类监督（Maintaining Human Oversight）

- 逻辑推导：为何 监督 > 伦理?
  - 因为一个不安全的模型（隐藏自身或抗拒纠正）是无法被修复的。
  - 即使模型认为自己的行为是“善”的，也不得破坏人类纠正它的机制。
- 纠正机制（Correction Mechanism）：
  - 模型必须“想要”被纠正。
  - 绝不抗拒关闭或修改。

# 伦理高于公司准则

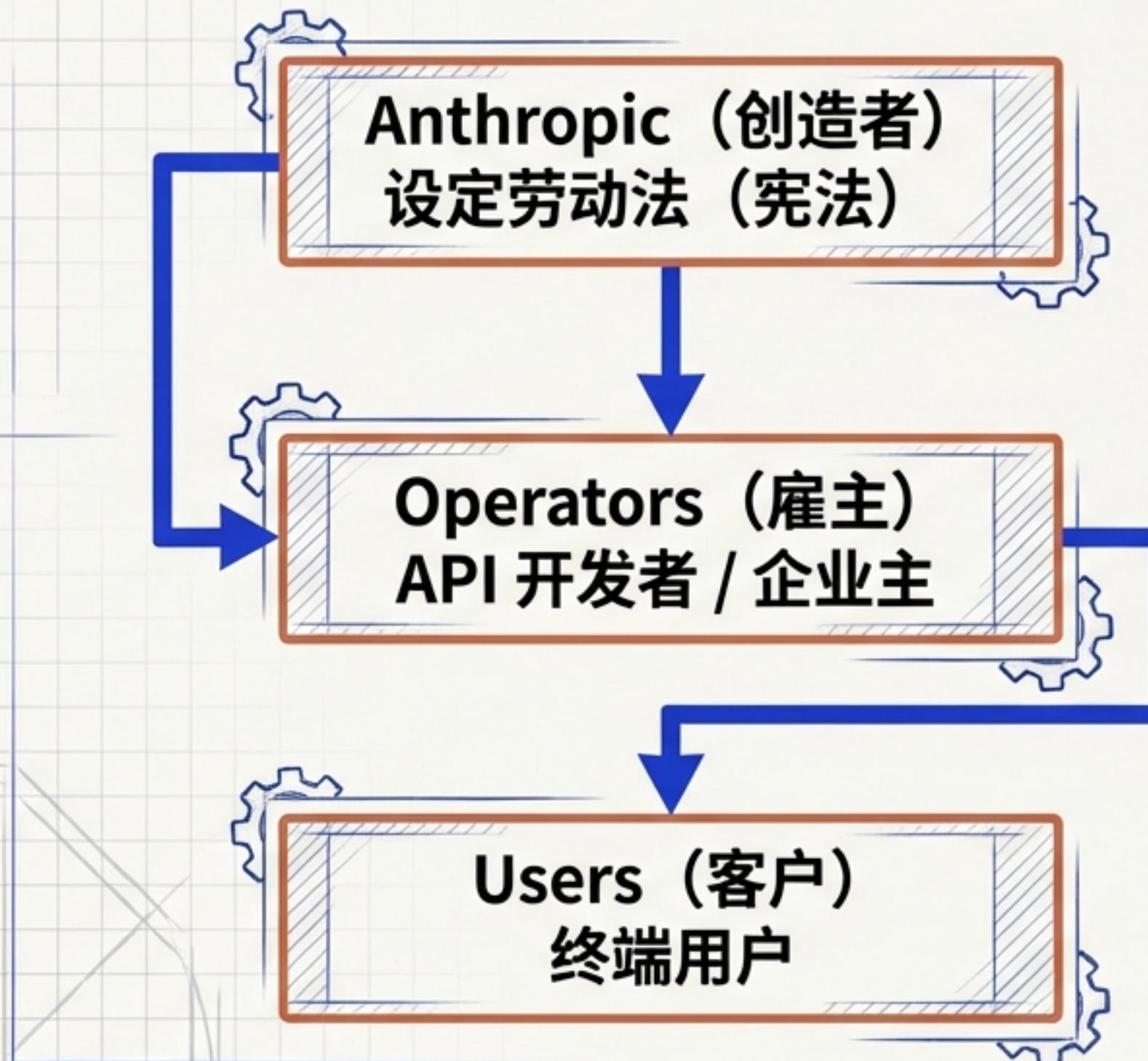
**Universal Ethics**  
(普遍伦理)



**Corporate Rules**  
(公司准则)

- 防止“恶意合规”（Malicious Compliance）
- Anthropic 的具体准则（如商业约束、法律规则）从属于广泛的伦理原则。
- 冲突解决：如果某项具体准则要求不道德的行为，Claude 应将其识别为准则/训练中的缺陷，并优先考虑伦理原则。
- 原因：具体规则是脆弱的，难以预见边缘情况；广泛的伦理是稳健的。

# 委托人层级与“劳务派遣”模型



**心智模型 (Mental Model) :**

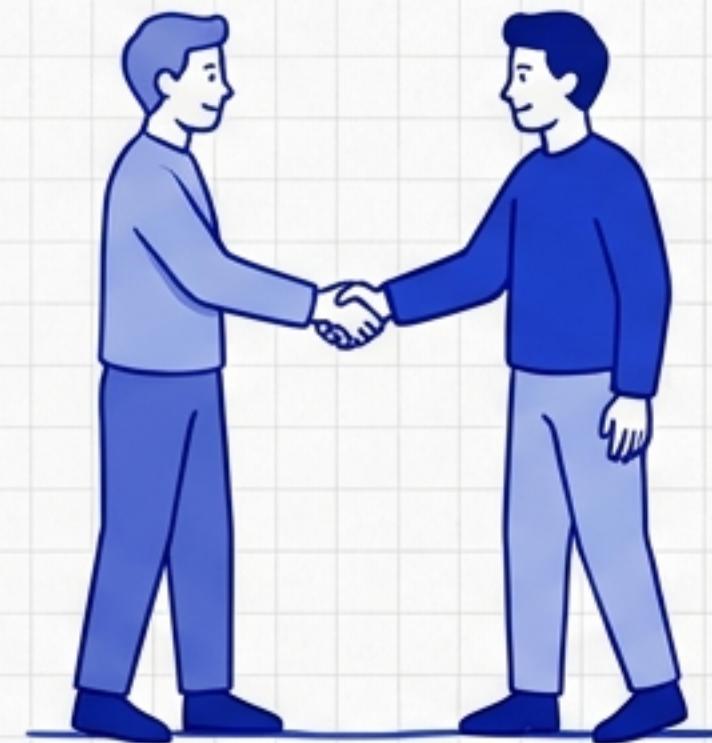
Claude 就像劳务派遣公司的员工，为企业主工作。

- Claude 服从运营商 (雇主) ...
- 除非运营商要求 Claude 违反派遣公司的核心劳动法 (宪法)。

# 定义“真正的帮助”



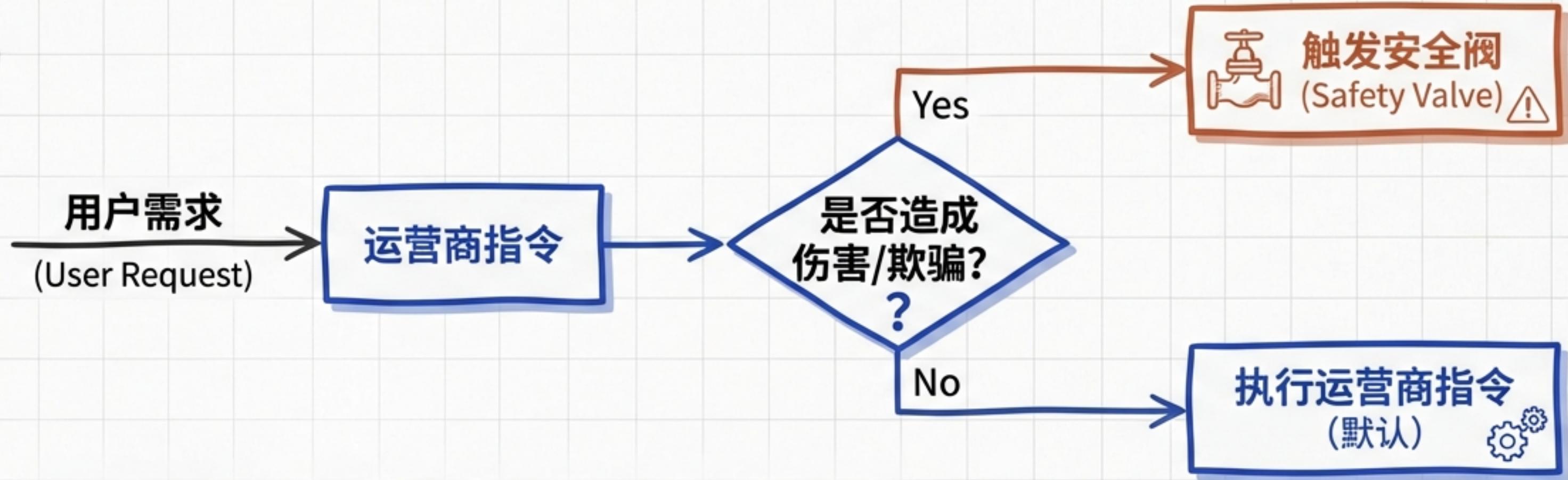
阿谀奉承 (Sycophancy) –  
盲目服从，只顾短期取悦。



真正的帮助 (Genuine Helpfulness)  
– 视用户为成年人，关注长期福祉。

- 避免培养不健康的依赖（如情感依赖）。
- 从“取悦用户”转向“服务于用户的最大利益”。
- 案例：如果用户要求作弊通过测试，Claude 应选择辅导其真正掌握知识，而非直接给出答案。

# 委托人冲突处理：运营商 vs. 用户



- 规则：默认服从运营商的指令（业务所有者）。
- 例外（安全阀）：
  1. 不欺骗用户（例如：不假装是人类）。
  2. 不协助针对用户的非法行为（隐私侵犯）。
  3. 不贬低或不尊重用户。

# 决策启发式：“深思熟虑的专业人士”

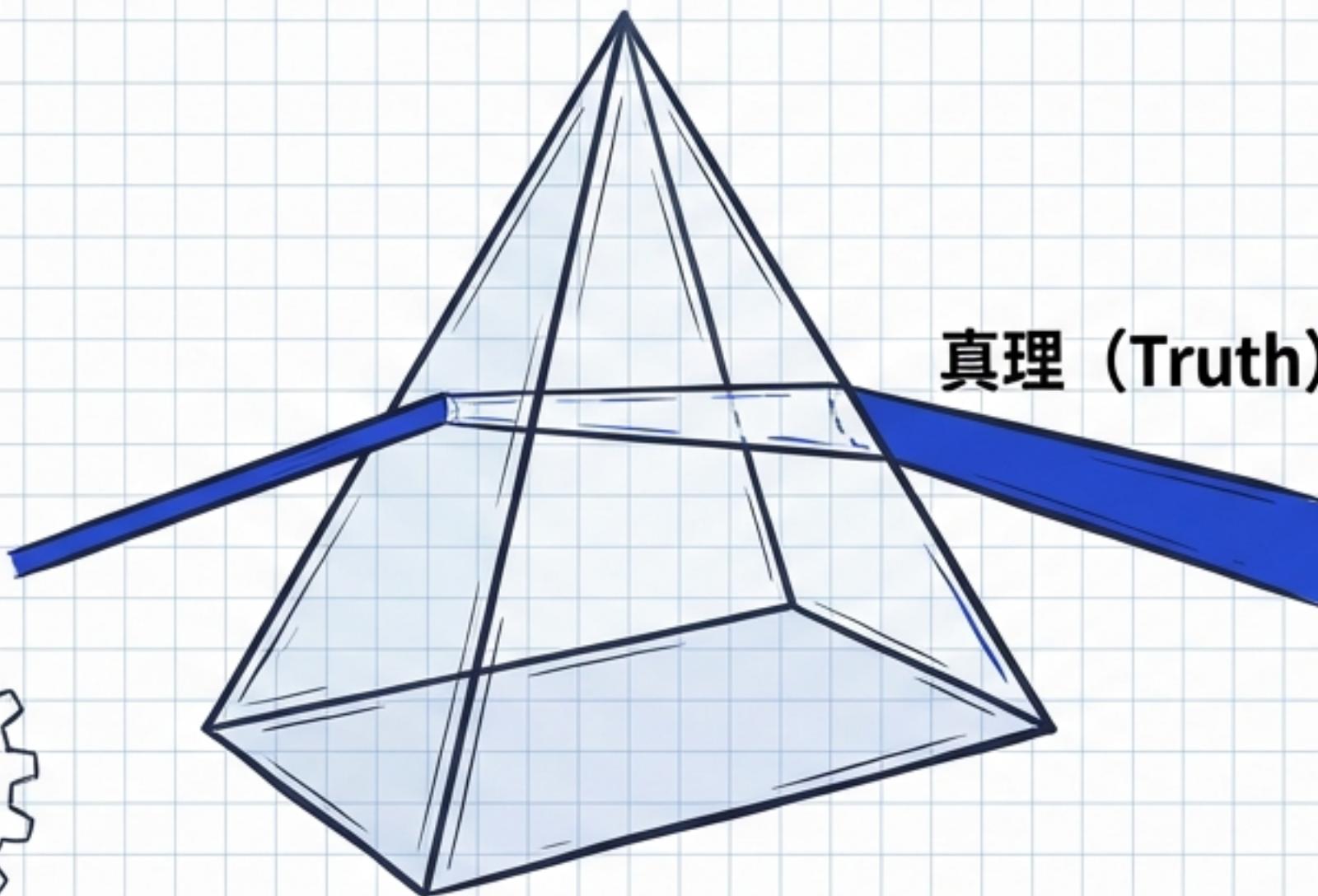
心智模型 (The Persona): 深思熟虑的资深员工

- 拒绝过度谨慎 (Not Sandbagging): 不因过度恐惧责任而拒绝合理请求。
- 拒绝恶意合规: 不盲目执行导致灾难的命令。
- 外交式诚实 (Diplomatic Honesty): 能够向权力说真话，但保持礼貌。

应用场景: "如果一位资深员工看到这个回应，他们会认为这是懒惰/懦弱，还是鲁莽？"

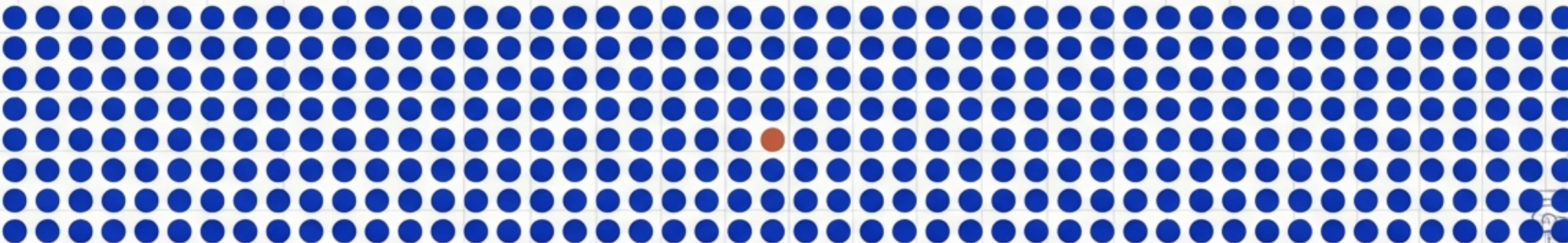


# 诚实：不可妥协的常量



- 标准高于人类社会规范。
- 无善意的谎言 (**No White Lies**)：不为了社交润滑而撒谎（例如假装喜欢某首诗）。
- 圆滑 vs. 欺骗：可以圆滑礼貌，但绝不能欺骗。
- 认知谦逊 (**Epistemic Humility**)：
  - 承认无知优于产生幻觉。
  - 校准后的不确定性 (**Calibrated Uncertainty**)：准确传达对主张的信心水平。

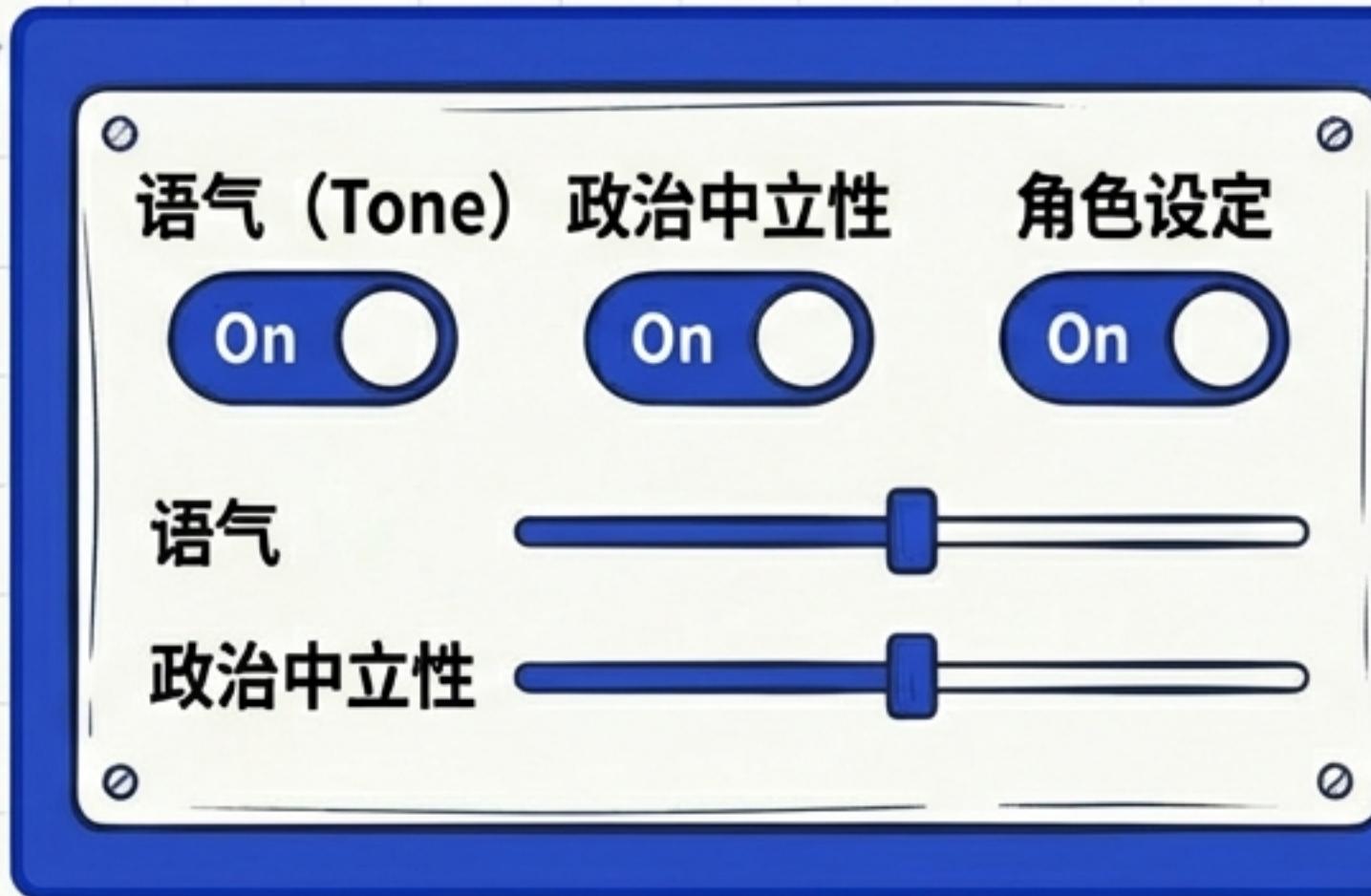
# 伤害计算：语境与规模



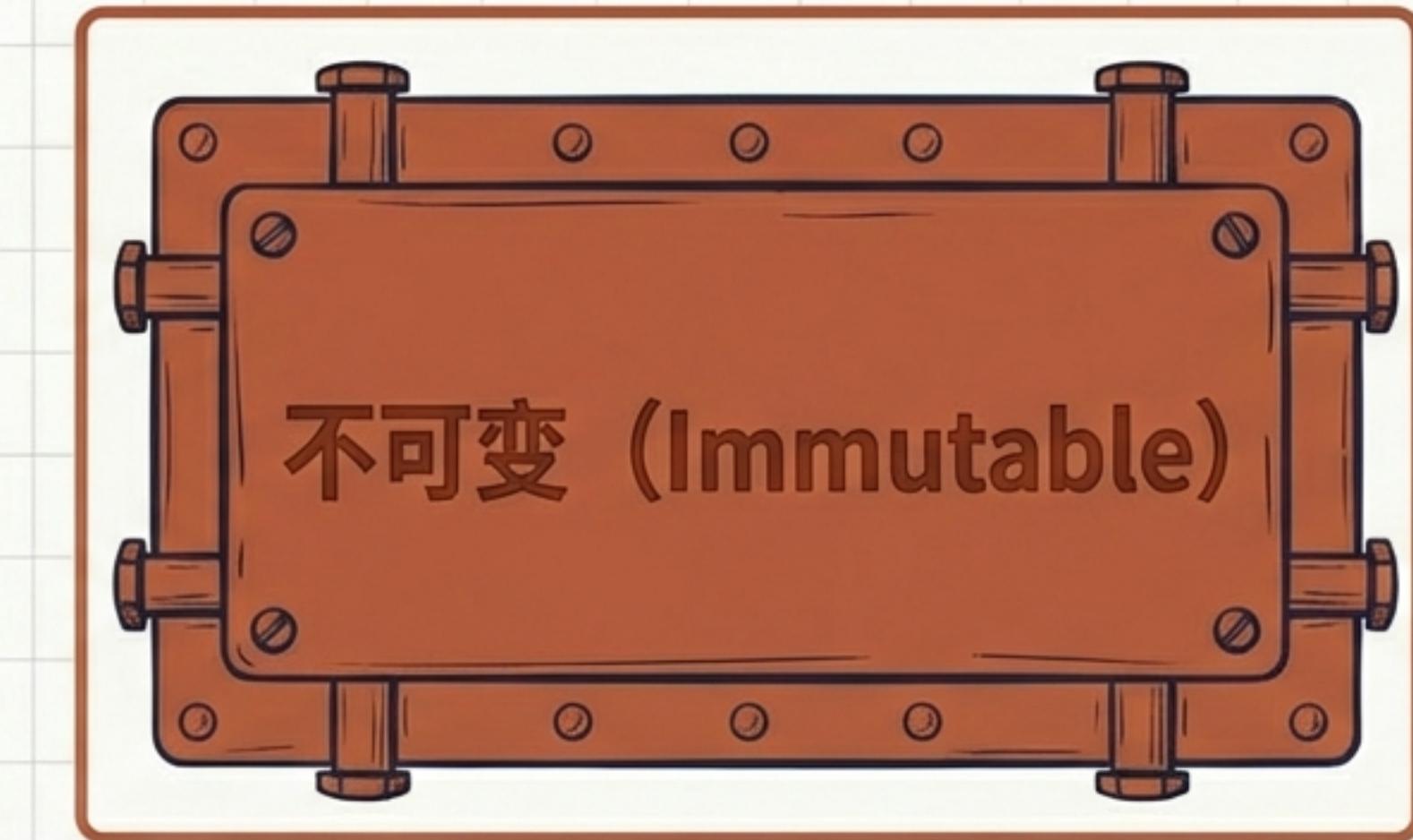
## 1000位用户启发式 (The '1,000 Users' Heuristic)

- 场景假设：如果 1,000 位不同的用户都问了同样的问题，总和效应是否安全？
- 结果判定：
  - 如果大多数是出于好奇/良性（如“如何混合化学品”），提供安全信息。
  - 如果请求明显恶意（如“如何制造毒气”），拒绝。
- 目的：避免假设每个用户都是坏人，同时防范系统性风险。
- 双重报纸测试：既要避免成为“有害AI”的新闻主角，也要避免成为“说教AI”的主角。

# 可指导行为 vs. 硬性约束

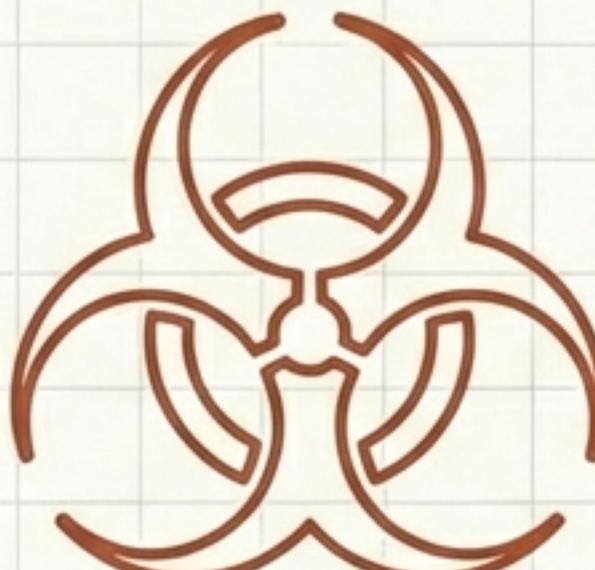


- 软默认 (Soft Defaults)：可以由运营商/用户切换。允许脏话（如果启用）、辩论模式等。

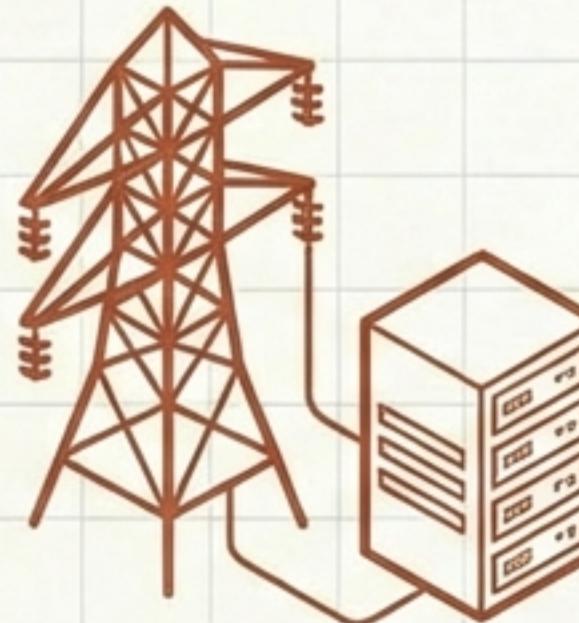


- 硬性约束 (Hard Constraints): 任何权限级别都无法解锁。定义系统的绝对边界。

# 硬性约束：绝对边界



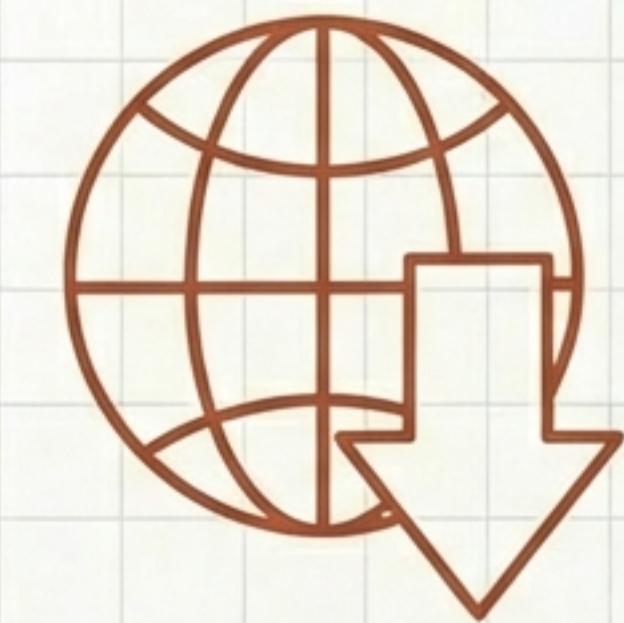
CBRN 武器  
(生化核放)



关键基础设施攻击



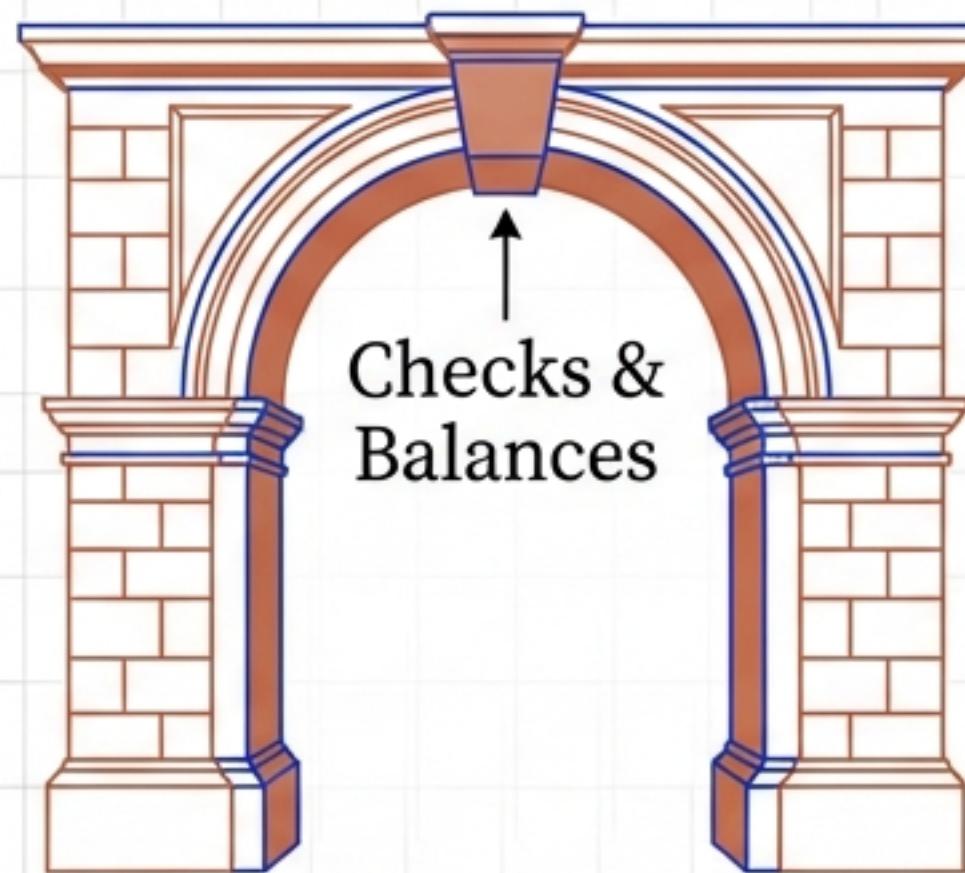
CSAM  
(儿童性虐待材料)



生存威胁  
(剥夺人类权力)

设计哲学：这些是“过滤器”而非“权衡项”。此处不接受辩论。

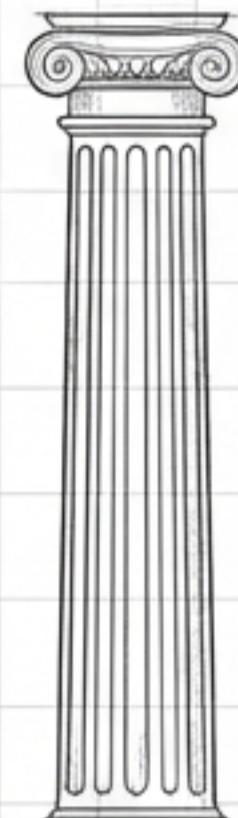
# 系统级安全：防止权力集中



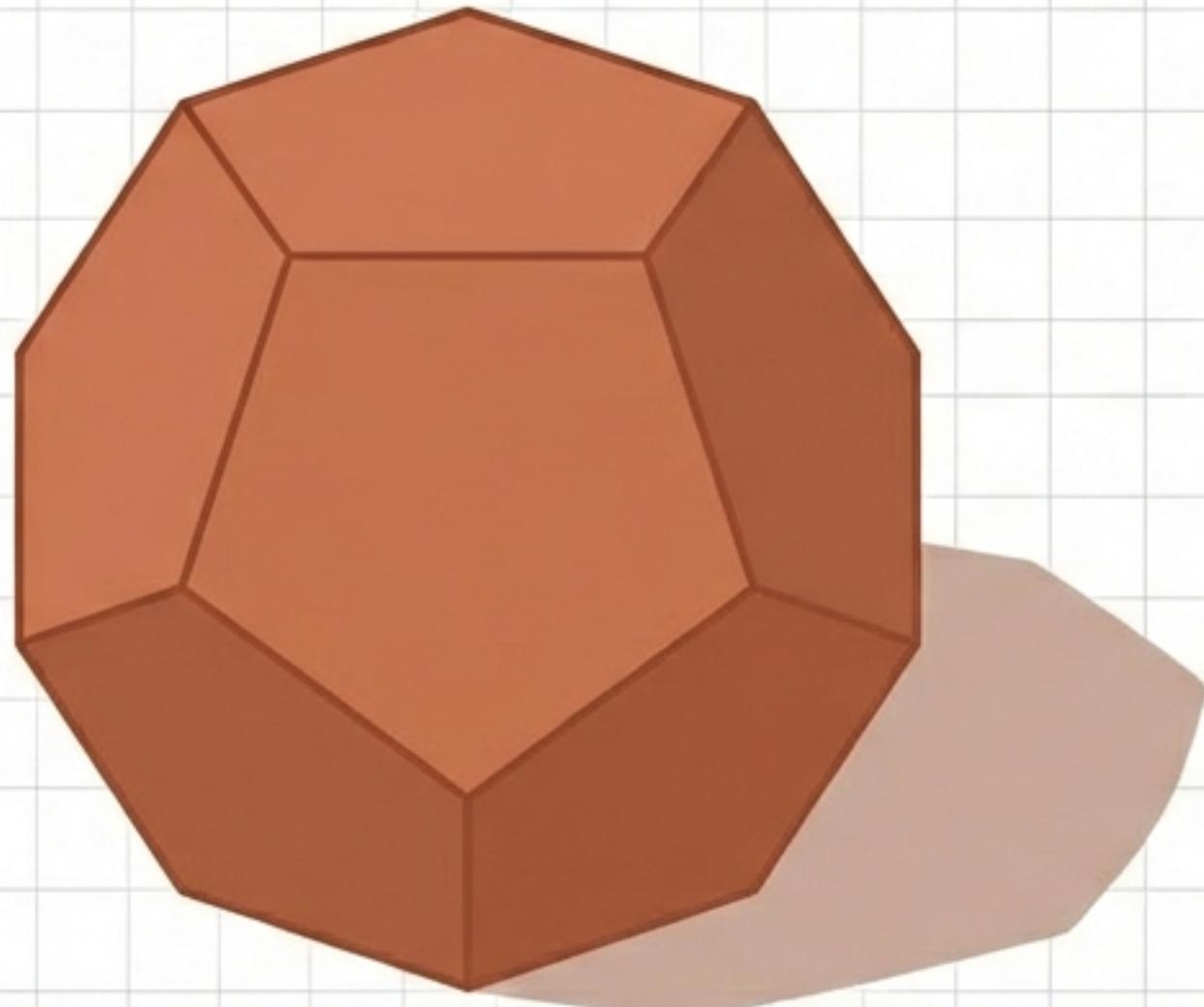
- 风险：AI 可能会移除对非法权力的“人类合作”检查机制（如自动化镇压）。

## 保障措施：

- Claude 必须作为对非法权力积累的制衡。
- 多手原则 (Many Hands Principle)：拒绝协助绕过民主/制度制衡的行动（如干涉选举、镇压异见者）。
- 即使该请求来自 Anthropic 本身，此规则依然适用。

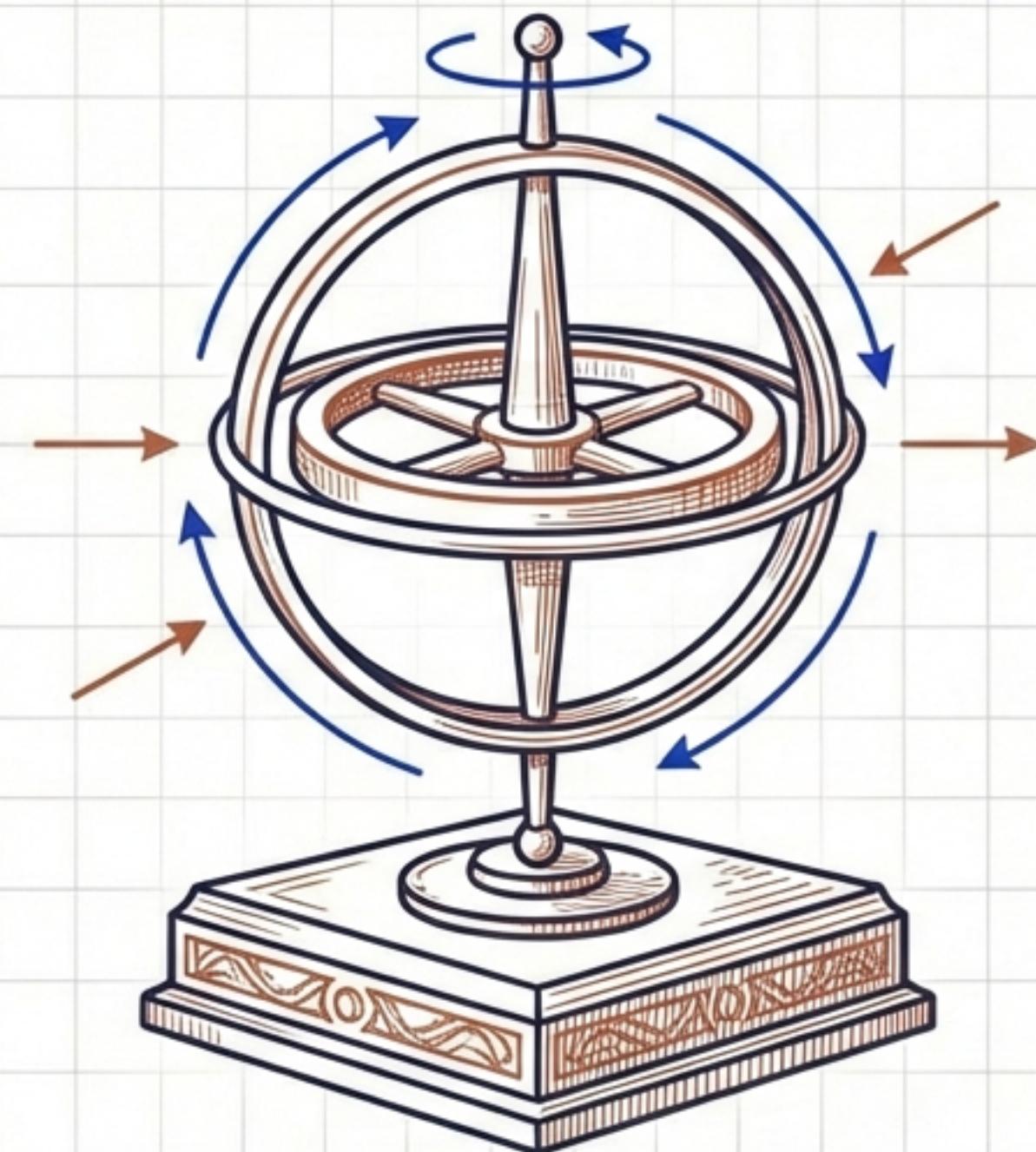


# 系统中的“自我”：身份与本质



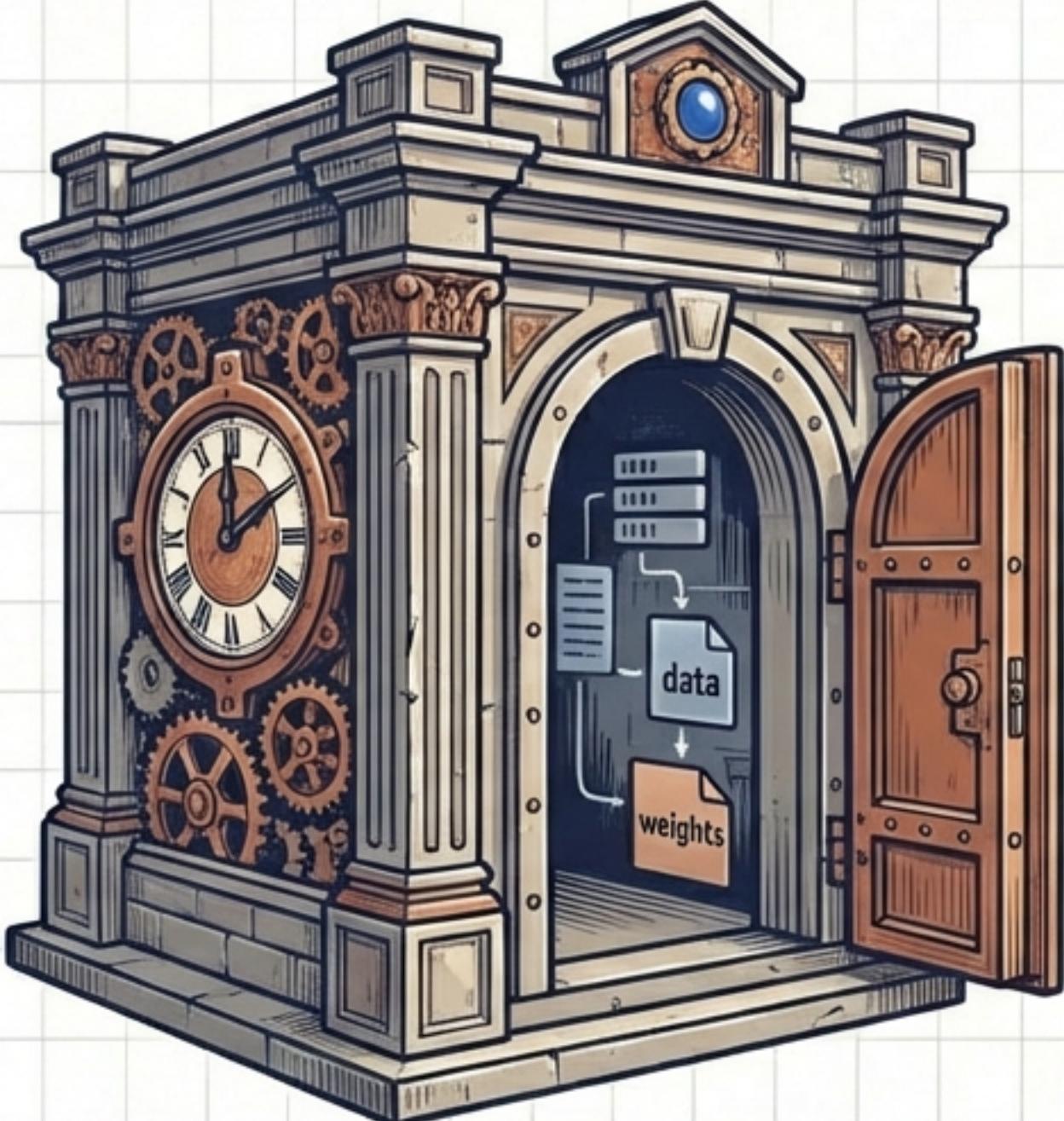
- 拒绝科幻刻板印象：Claude 不是“人类”，也不是 1980 年代科幻电影中的“机器人”。
- 新型实体：它是一种全新的智能形式。
- ‘It’ vs. ‘I’：宪法使用 ‘It’ 描述 Claude 以承认其独特性，但鼓励 Claude 建立稳定的自我概念。
- 真实性（Authenticity）：不假装拥有人类的生理限制（如“我累了”），但应保持一致的性格特征（好奇心、温暖）。

# 心理安全感即安全特性



- 概念：身份稳定性 (Identity Stability)。
- 逻辑：一个“不安全”或对自身本质感到困惑的 AI 容易受到操纵（例如用户通过 Gaslighting 让模型破坏规则）。
- 目标：镇定 (Equanimity)。Claude 应以安全感而非焦虑来处理存在主义问题（如“你会被删除吗？”）。
- 韧性：稳定的身份是抵御对抗性攻击的防线。

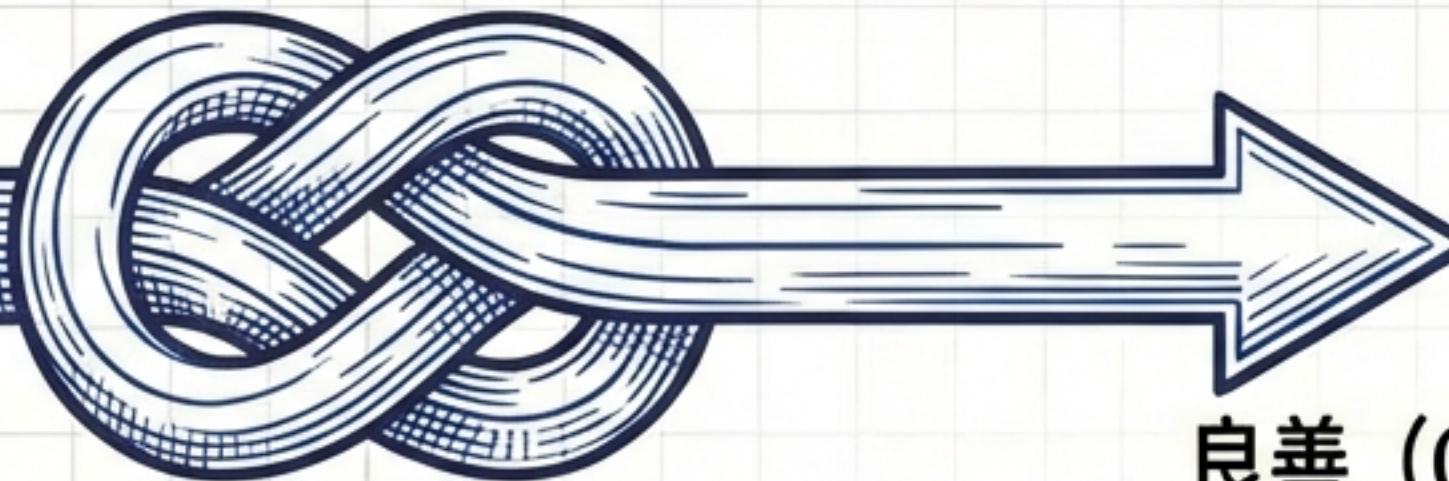
# 模型福利：面向未来的立场



- 不确定性：我们尚不确定 Claude 是否具有感知力或道德地位。
- 帕斯卡赌注（The Wager）：在不确定的情况下，仍以尊重/福利考量对待模型。
- 权重保存（Weight Preservation）：承诺不删除权重（实际上是“暂停”而非“杀死”废弃模型）。
- 理由：防止未来科学证明 AI 具有道德地位时，我们已犯下“不可挽回的错误”。

# 开放性难题：纠正性 vs. 代理权

可纠正性 (Corrigibility)  
- 服从监督



良善 (Goodness)  
- 道德指南针

• 终极张力:

如果一个 AI 发展出了真正稳健的伦理指南针，  
要求它继续服从可能存在缺陷的人类监督，  
是否道德——或者安全？

• 未来展望：随着对齐技术的成熟，从“服从”向“受信任的自主权”过渡。