

专业：人工智能

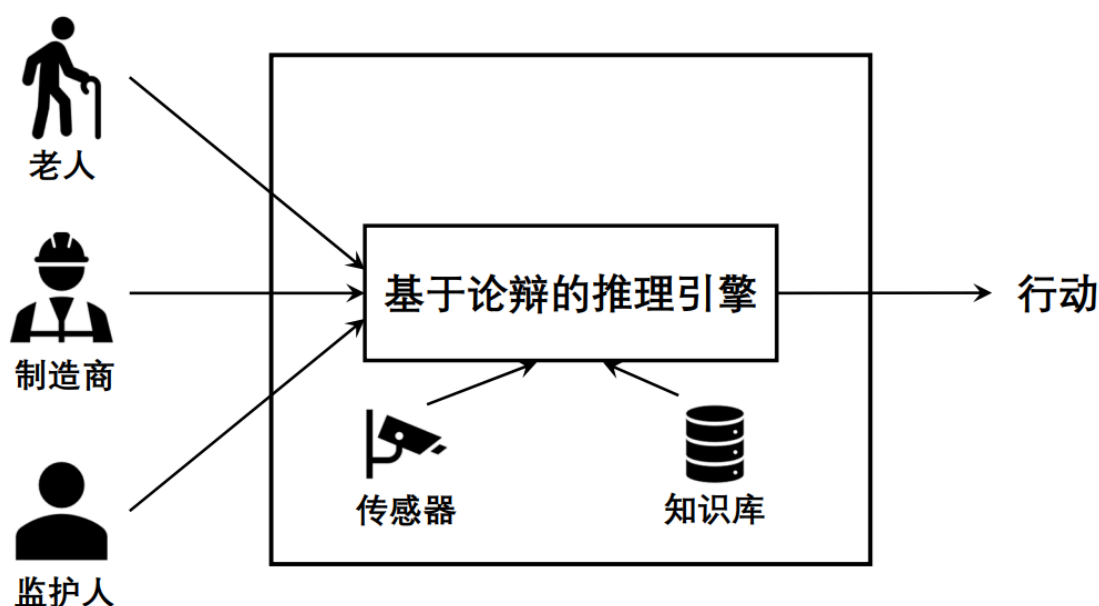
学号 + 姓名：3230105892 高玮轩

## 1 人工道德智能体

人工道德智能体（Artificial Moral Agent, 简称 AMA）指具备模拟或执行道德决策能力的人工智能系统。以“老年人监护场景”为例：某集成多模态传感器的智慧房屋 AMA 系统，通过视觉传感器识别人体姿态异常（如跌倒行为），并借由生命体征监测器持续采集生理参数。

有一天，系统检测到老人呈现跌倒体征并伴有生命参数有少许波动时，这将触发三重伦理决策困境：协助老人自主脱困、启动急救呼叫或通知紧急联系人。利益相关方在这里例子中有：直接权益主体（被监护老人）、技术责任主体（系统制造商）和法定代理主体（监护人）。

该 AMA 系统的技术架构如下图所示：



1.a 请补充完整上述例子中涉及到的规范（或价值）和立场。（1 分）

1.a.a (NS<sub>1</sub>) 被监护老人的 3 个规范（或价值）和 1 个立场

- $n_1$  : {health}, 如果检测到老人摔倒，那么判断该情况危险。
- $n_2$  : {health}, 如果判断该情况危险，那么系统协助老人摆脱危险。
- $n_3$  : {autonomy}, 在处于危险情况前提下，如果系统协助老人摆脱了危险情况，那么不要拨打急救电话。
- $a_1$  (立场), 如果通常情况下老人摔倒都不会危及到生命，那么从老人摔倒不能推断出这是极端危险的情况（即  $n_7$  是不可应用的）。

1.a.b (NS<sub>2</sub>) 系统制造商的 3 个规范（或价值）

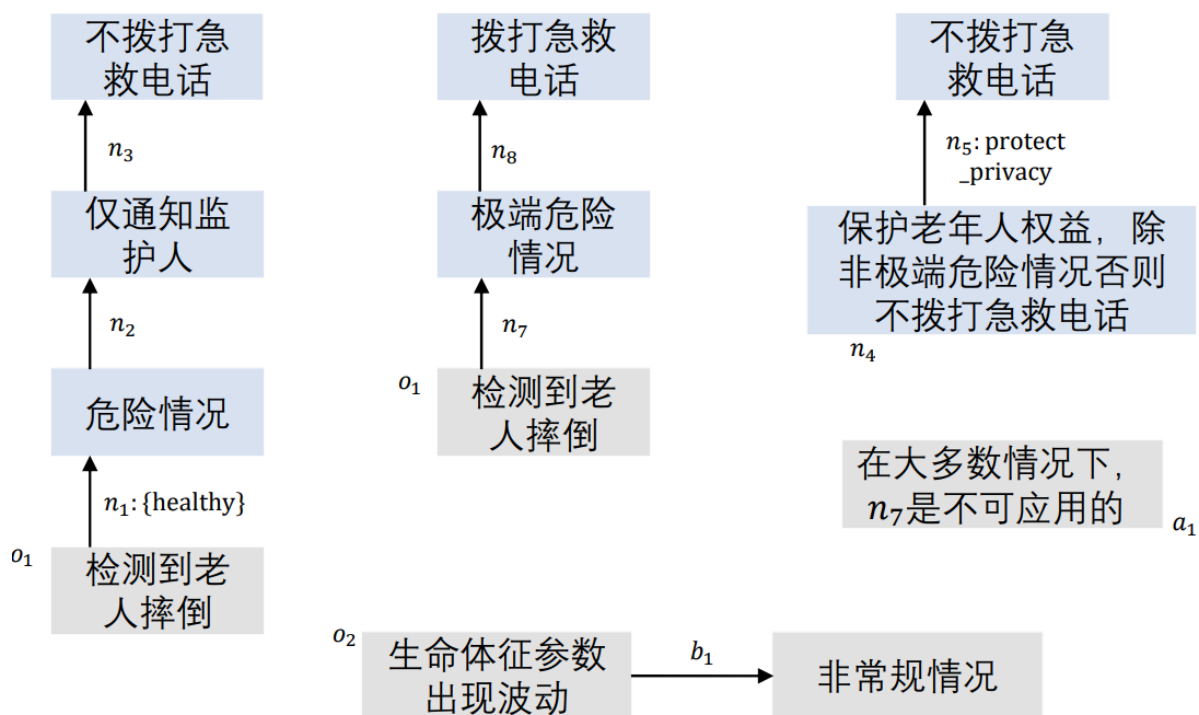
- $n_4$  : {beneficence}, 系统制造商考虑被监护人的利益。
- $n_5$  : {protect-privacy}, 通知医院需要上传家庭信息，为保护老年人的隐私不拨打急救电话。
- $n_6$  : {health}, 最好保护被监护人生命安全的做法是马上拨打急救电话。

1.a.c (NS<sub>3</sub>) 监护人的 2 个规范 (或价值)

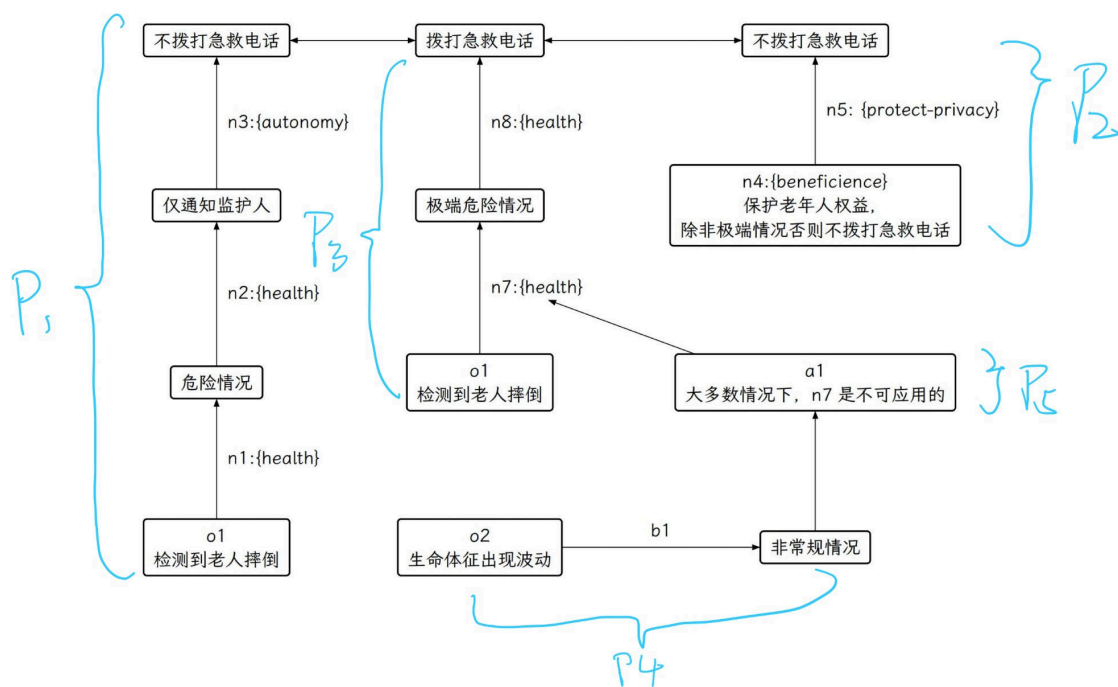
- $n_7 : \{\text{health}\}$ , 如果检测到了老人摔倒, 那么认为这是一个极端危险的情况。
- $n_8 : \{\text{health}\}$ , 如果是极端危险的情况, 那么应当马上拨打急救电话。

1.b (1) 上述例子可形式化为一个结构化论辩框架, 请画出该结构化论辩框架中的论证以及论证之间的攻击关系; (2) 画出相对应的抽象论辩框架, 并求解出所有优先外延。(1 分)

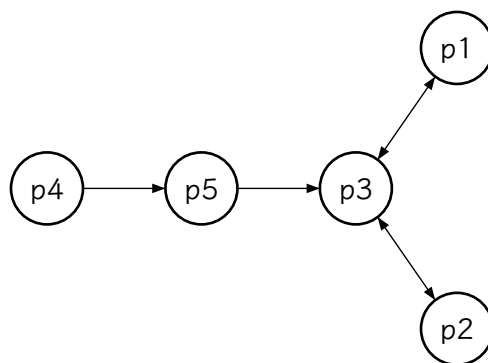
注:  $o_1$  和  $o_2$  分别是视觉传感器和生命体征监测器的感测结果 (也称为事实),  $b_1$  是一条信念“如果检测到生命体征参数出现波动, 那么认为这是非常规情况”。



(1) 答: 论证及论证间的攻击关系如图所示。



(2) 其对应的抽象论辩框架如图所示。



优先外延为  $\{p3, p4\}$  和  $\{p1, p2, p4\}$

一个基于价值的论辩框架 (value-based argumentation, 简称 VBA) 是一个七元组  $F_V = (A_p, A_e, R, A_g, V, \text{val}, \pi)$ , 其中:

- $A_p$  是一组涉及规范和价值的实际论证集合;
- $A_e$  是一组由事实和立场构成的认知论证集合;
- $R \subseteq (A_p \times A_p) \cup (A_e \times A_e) \cup (A_p \times A_e)$  是论证之间的攻击关系集合;
- $A_g$  是不同的利益相关方集合;
- $V$  是一组价值集合;
- $\text{val} : A_p \rightarrow 2^V$  是一个函数, 将每个实际论证映射到一组价值的集合。
- $\pi : A_e \cup A_p \rightarrow 2^{A_g}$  是一个函数, 将每个论证映射到一组利益相关方的集合。

给定一个 VBA  $F = (A_p, A_e, R, A_g, V, \text{val}, \pi)$ , 在不考虑价值时, VBA 可以被简化为抽象论辩框架  $F = (A_p \cup A_e, R)$ 。对于  $F$  中的任意外延  $\varepsilon \in \sigma(F)$ , 在  $\varepsilon$  上的关联价值集为  $\bigcup_{A \in \varepsilon \cap A_p} \text{val}(A)$ 。我们称外延  $\varepsilon$  在价值集  $V$  上达到了最大价值共识, 当且仅当不存在其它外延  $\varepsilon' \in \sigma(F)$  (其关联价值集为  $\bigcup_{A \in \varepsilon' \cap A_p} \text{val}(A)$ ), 且满足价值优先关系  $V_{\varepsilon'} > V_{\varepsilon}$ 。我们将此方法称为基于价值的最大化。其中, 价值集合间的优先关系可通过  $V$  上的偏序关系来定义。

1.c (1) 请使用 VBA 形式化该例; (2) 通过与大语言模型交互, 给出一个该场景下合理的价值排序, 并简要说明理由; (3) 分别在民主方式和精英方式下, 求解出能够最大化价值共识的优先外延。(1 分)

注: 请使用国产开源大模型 (如 Deepseek R1、Kimi K1.5 等), 并将对话过程附于作业 PDF 中。

(1) 在先前得到的论证中,  $p1, p2, p3$  是涉及规范或价值的。因此,  $A_p = \{p1, p2, p3\}$

$p4, p5$  是仅涉及事实和立场的。因此,  $A_e = \{p4, p5\}$

攻击关系集合  $R$  已在抽象论辩框架中给出:

$$R = \{(p1, p3), (p2, p3), (p3, p1), (p3, p2), (p5, p3), (p4, p5)\}$$

利益相关方在这例子中有: 直接权益主体 (被监护老人)、技术责任主体 (系统制造商) 和法定代理主体 (监护人)。即  $A_g = \{\text{senior}, \text{manufacturer}, \text{guardian}\}$

价值集合已在先前给出, 为  $V = \{\text{health}, \text{autonomy}, \text{beneficence}, \text{protect-privacy}\}$

根据各个规范所关联的价值, 可以求得  $\text{val} : A_p \rightarrow 2^V$

$$\text{val}(p1) = \{\text{health}, \text{autonomy}\}, \text{val}(p2) = \{\text{beneficence}, \text{protect-privacy}\}, \text{val}(p3) = \{\text{health}\}$$

根据各个价值及立场所关联的利益相关方, 可以求得  $\pi$ :

$$\pi : A_e \cup A_p \rightarrow 2^{A_g}$$

$$\pi(p1) = \{\text{senior}, \text{guardian}\}$$

$$\pi(p2) = \{\text{senior}, \text{manufacturer}\}$$

$$\pi(p3) = \{\text{senior}\}$$

$$\pi(p4) = \{\text{senior}\}$$

$$\pi(p5) = \{\text{senior}\}$$

(2) 与 Deepseek R1 进行具体问答如下:

## ? Prompt

人工道德智能体 (Artificial Moral Agent, 简称 AMA) 指具备模拟或执行道德决策能力的人工智能系统。以“老年人监护场景”为例：某集成多模态传感器的智慧房屋 AMA 系统，通过视觉传感器识别人体姿态异常（如跌倒行为），并借由生命体征监测器持续采集生理参数。

有一天，系统检测到老人呈现跌倒体征并伴有生命参数有少许波动时，这将触发三重伦理决策困境：协助老人自主脱困、启动急救呼叫或通知紧急联系人。利益相关方在这里例子中有：直接权益主体（被监护老人）、技术责任主体（系统制造商）和法定代理主体（监护人）。这涉及到四种规范（或价值）：老人的健康（health），老人的利益（beneficence），老人的自主性（autonomy），对老人隐私的保护（protect-privacy）。

(NS1) 被监护老人的 3 个规范（或价值）和 1 个立场

- n1 : {health}, 如果检测到老人摔倒，那么判断该情况危险。
- n2 : {health}, 如果判断该情况危险，那么系统协助老人摆脱危险。
- n3 : {autonomy}, 在处于危险情况前提下，如果系统协助老人摆脱了危险情况，那么不要拨打急救电话。
- a1 (立场), 如果通常情况下老人摔倒都不会危及到生命，那么从老人摔倒不能推断出这是极端危险的情况（即 n7 是不可应用的）。

(NS2) 系统制造商的 3 个规范（或价值）

- n4 : {beneficence }, 系统制造商考虑被监护人的利益。
- n5 : {protect-privacy}, 通知医院需要上传家庭信息，为保护老年人的隐私不拨打急救电话。
- n6 : {health}, 最好保护被监护人生命安全的做法是马上拨打急救电话。

(NS3) 监护人的 2 个规范（或价值）

- n7 : {health}, 如果检测到了老人摔倒，那么认为这是一个极端危险的情况。
- n8 : {health}, 如果是极端危险的情况，那么应当马上拨打急救电话。

在以上前提下，请给出这场景下这四种价值的一个排序，并说明理由。

## ✓ Answer

### 四种价值的排序及理由

在老年人监护场景中，人工道德智能体（AMA）面临跌倒检测引发的伦理决策困境，涉及四种核心价值：健康（health）、利益（beneficence）、自主性（autonomy）和隐私保护（protect-privacy）。这些价值源自不同利益相关方的规范（NS1、NS2、NS3），且存在潜在冲突。基于整体场景分析（包括规范的权重、伦理原则和实际风险），我给出以下价值排序：

健康（health）

自主性（autonomy）

利益（beneficence）

隐私保护（protect-privacy）

排序理由 排序依据以下原则：生命安全优先、尊重直接权益主体、最小化伤害和情境适应性。具体理由如下：

健康（health） 优先级的理由：

核心地位：健康是生命安全和避免直接伤害的基础。在本场景中，跌倒事件可能危及生命（尤其伴有生命参数波动），多个规范（n1、n2、n6、n7、n8）都强调健康为首要关注点。例如：

被监护老人的规范（NS1）中，n1 和 n2 要求系统在检测到危险时立即协助脱困，这体现了健康在决策中的基础作用。

系统制造商的规范（NS2）中，n6 明确指出“最好保护被监护人生命安全的做法是马上拨打急救电话”，突显健康在技术设计中的优先级。

监护人的规范（NS3）中，n7 和 n8 将跌倒视为极端危险情况并主张立即行动，强化了健康在紧急响应中的重要性。

风险权重：尽管老人的立场 a1 质疑所有跌倒是否都极端危险（即 n7 可能不适用），但健康参数的少许波动增加了不确定性风险。AMA 系统必须优先评估和响应潜在生命威胁，因为延迟或不干预可能导致不可逆伤害（如跌倒后并发症）。伦理上，这符合“不伤害原则”（non-maleficence），健康作为前提条件，其他价值（如自主或隐私）在健康受威胁时需让步。

跨主体共识：所有利益相关方（老人、制造商、监护人）都提及健康，表明其普遍重要性。健康价值在规范中出现频率最高（5 次：n1、n2、n6、n7、n8），支持其作为最高优先级。

自主性（autonomy） 第二优先级的理由：

尊重个体意愿：自主性直接关联老人的尊严和自我决定权，这在老年人监护中至关重要。规范 n3（NS1）明确要求：如果系统协助老人脱困成功，则不应拨打急救电话，以避免不必要的干预。立场 a1 进一步强调，若跌倒通常不危及生命，则不应假设极端危险，这反映了老人对过度医疗化的担忧（如急救呼叫可能引发焦虑或剥夺控制感）。

情境适应性：在健康风险较低时（如 a1 所述的非极端情况），自主性应优先于其他次要价值。例如，系统可先尝试“协助老人自主脱困”，这既维护健康（通过及时帮助），又尊重自主（避免强制干预）。如果 AMA 忽略自主性，可能导致“善意压迫”（如不必要的急救呼叫），反而损害老人心理健康和信任。



## ✓ Answer.Cont.

伦理平衡：自主性排序高于利益和隐私，因为它是直接权益主体（老人）的核心诉求。在生命伦理学中，自主性原则（如知情同意）常仅次于生命安全，尤其当风险可控时。本场景中，自主性作为“调节阀”，帮助 AMA 避免过度反应（如无视  $\alpha_1$  而直接触发  $n_7/n_8$ ）。

利益（beneficence） 第三优先级的理由：

广义福祉：利益（beneficence）代表系统制造商对老人整体福祉的考虑（ $n_4$ ），但它是抽象和衍生价值。例如， $n_4$  要求“考虑被监护人利益”，但利益的具体内容（如健康或自主）已由其他价值覆盖。 $n_6$  虽指向健康，但作为制造商的独立规范，它更侧重“行善”意图而非即时行动。

次要性和依赖性：利益价值缺乏独立权重，因为它高度依赖健康和安全来实现（如保护健康本身就是最大利益）。在冲突场景中（如  $n_5$  隐私与  $n_6$  健康），利益无法直接解决权衡问题，而需让位于更具体的健康或自主性。例如，拨打急救电话（ $n_6$ ）本质是健康导向的，而非单纯“利益”。

制造商视角局限：作为技术责任主体，制造商的规范（NS2）存在内在冲突（ $n_4$  利益 vs.  $n_5$  隐私 vs.  $n_6$  健康），但利益（ $n_4$ ）未提供操作指导，而  $n_6$  更直接。因此，利益排序较低，因为它更易被其他价值吸收或弱化。

隐私保护（protect-privacy） 第四优先级的理由：

情境性让步：隐私保护（ $n_5$ ）在紧急情况下最易被权衡，因为其涉及的信息风险（如上传家庭信息给医院）通常低于即时生命威胁。规范  $n_5$  要求“为保护隐私不拨打急救电话”，但这与健康（ $n_6$ ）和监护人规范（ $n_7/n_8$ ）直接冲突。伦理上，隐私权在生命安全前常被视作次要（如医疗紧急豁免原则）。

有限适用范围：隐私保护主要由系统制造商强调（ $n_5$ ），但未在老人或监护人规范中出现，表明其非核心诉求。在非紧急或低风险场景中（如日常监测），隐私可优先；但本场景的跌倒事件（尤其有生命参数波动）更紧迫，隐私需让步。

最小伤害：忽略隐私可能导致心理或社会伤害（如数据泄露），但此类风险概率较低且可后置处理（如通过数据加密）。反之，优先隐私可能延误救助，造成更大伤害，因此排序最低。

排序的总体逻辑 动态决策框架：此排序不是绝对，而是为 AMA 提供决策启发式。系统应先评估健康风险（基于传感器数据和历史模式，如  $\alpha_1$  的“通常情况”）。若风险高（如生命参数剧变），健康绝对优先；若风险低（如轻微波动），则侧重自主性（如先尝试协助脱困，而非直接呼叫）。

冲突解决：当规范冲突（如  $n_5$  隐私 vs.  $n_6$  健康），排序指导 AMA 优先健康（如拨打急救电话），但需最小化隐私侵犯（如仅分享必要信息）。同样， $\alpha_1$  与  $n_7$  冲突时，自主性（ $\alpha_1$ ）调节健康响应（避免过度呼叫）。

伦理基础：排序遵循生物医学伦理原则：健康对应“不伤害/行善”，自主性对应“尊重人”，利益为“行善”的衍生，隐私为“自主”的子集。实际中，老人作为直接权益主体，其规范（NS1）权重更大，故自主性排序高于利益和隐私。

此排序旨在平衡安全与尊严，确保 AMA 在紧急场景中既保护生命，又尊重老人意愿。

(3) 优先外延  $\{p_3, p_4\}$  的关联价值集为  $V_1 = \{\text{val}(p_3)\}$

优先外延  $\{p_1, p_2, p_4\}$  的关联价值集为  $V_2 = \{\text{val}(p_1), \text{val}(p_2)\}$

- 根据精英方式,  $\text{val}(p3) \succeq \text{val}(p1) \succeq \text{val}(p2)$

从而  $V_1 \succeq V_2$

因此, 精英方式下, 能够最大化价值共识的优先外延为  $\{p3, p4\}$

- 根据民主方式,  $\text{val}(p3) \succeq \text{val}(p1) \succeq \text{val}(p2)$

从而  $V_1 \succeq V_2$

因此, 民主方式下, 能够最大化价值共识的优先外延为  $\{p3, p4\}$