青少年对人工智能在道德困境中决策的接受度

张义玲

广西师范大学教育学院心理学系雁山校区，广西桂林，541006；

摘要：本研究采用 2（决策主体：人工智能 vs 人类）×2（道德类型：功利主义 vs 义务论）的完全被试间实验设计，探讨不同决策主体与道德类型对高中生道德建议接受度的影响。结果显示，决策主体主效应显著，被试对人类决策的接受度显著高于人工智能决策；道德类型主效应不显著，功利主义与义务论建议在整体接受度上无差异。进一步分析表明，决策主体与道德类型存在显著交互作用：在功利主义情境下，被试对人类决策的接受度显著高于人工智能决策；在义务论情境下，该差异仍存在但较弱。研究揭示了青少年在不同道德框架下对人工智能决策接受度的差异化表现及其主体性偏好模式。

关键词：人工智能；接受度；功利主义；义务论

DOI：10.64216/3104-9702.25.05.052

1 引言

青少年阶段，是其生理、心理、社交等各个方面发生转变的枢纽期，对于其心理发展的正向引导极为关键[1]，更是个体人格和价值观形成的重要阶段，增强正确的道德判断能力具有重要意义。Krettenauer（2011）提出并实证了青少年道德身份的“双维模型”：道德认知与道德动机。其研究发现，在青少年中，道德动机比认知更强预测道德行为，而认知维度则与判断能力关联更强。也就是说，青少年道德判断尚处于情感与理性加工并存的过渡阶段，这对理解人类与机器作为道德建议来源的接受机制具有启示意义[2]。

在心理学与技术伦理的交叉研究中，人工智能（AI）与人类作为道德决策主体的比较问题日益受到关注。大量研究表明，当个体面对由 AI 或人类作出的道德决策时，其在接受程度与责任归因方式上常表现出显著差异。例如，个体更倾向于对人类决策产生情感共鸣及动机推测，而对 AI 的判断则更多基于其工具性特征与算法属性，从而导致不同的道德评估路径[3]。相较而言，人类所作出的道德决策更容易激发个体的情感共鸣与对其行为动机的推测，从而引导更具情境化和主观色彩的道德评判；而人工智能的决策过程则通常被理解为基于规则和算法驱动的理性计算结果，缺乏情感意图与道德人格的投射空间。因此，个体在对两类主体进行道德责任归因时，常表现出显著分歧，尤其在涉及伦理责任、意图判断及可谴责性的维度上差异尤为突出[4]。

Hertz 与 Wiese（2019）的研究为理解人类与人工智能在建议采纳与决策中的系统性差异提供了较为坚实的证据。尽管实验中 AI 的任务表现与人类相当，但参与者在选择建议来源以及最终做出判断时，普遍呈现出对人类建议的更高依从度。该结果说明，在建议质量一致的前提下，人类行动者因其在意图推断、情境理解和责任归属等方面的“可解释性”优势而获得更高的信任，而这一优势并非源自绩效差异，而是来自被试对不同主体的心理模型差异。从理论贡献看，Hertz 与 Wiese 的发现强调了“主体类型”这一因素在人机比较研究中的关键作用。其意义不仅在于揭示了人类主体在责任相关情境中的固有优势，也提示公众对 AI 缺乏心智与责任承担能力的认知会对决策接受度造成系统性影响。对于探讨道德类型（如功利主义 vs. 义务论）以及损害主体（human vs. AI）如何影响赔偿判断的研究而言，这一现象尤具启发性：即使在道德情境保持不变的情况下，行动主体的社会认知地位本身就可能塑造人们的评估趋向，使人类主体在涉及道德后果或责任判断时获得更高的接受度或宽容度。因此，Hertz 与 Wiese（2019）不仅扩展了我们对建议来源效应的理解，也为后续研究提出重要提醒——在人机比较框架下，公众对 AI 的心智归因、可信度及责任期待是不可忽视的影响因素，可能系统性地调节道德评价、赔偿意愿及代理主体的整体判断模式[5]。

Malle 等人（2015）通过对比人类与机器人施事者在相同牺牲情境下的道德评价，发现尽管行为与结果一致，人们对机器人和人类适用的道德规范却不同：机器人若采用功利型牺牲逻辑，其被责备程度更高、许可性更低。该研究提示，在设计机器人或 AI 承担道德任务时，不仅要考虑功能性能，更要关注公众对其“责任主体”身份的道德期望差异[6]。有研究进一步支持，Zhang 等人（2023）通过三项实验（总样本量为 626 名被试）系统考察了个体在道德两难情境中对人工智能（AI）与人类决策者的道德评价差异。在“电车困境”任务中，

研究结果显示，影响参与者道德判断的主要因素是决策主体的类型，而非具体行为内容。与人类决策相比，被试更倾向于将 AI 的行为视为不道德，并认为其更值得责备。然而，在“天桥困境”任务中，情况有所不同：个体的判断主要受行为本身影响，而非代理者身份。具体来说，功利主义的“行动”被评为较不符合道德规范、接受度较低，且比义务论取向的“不行动”更应承担道德责任。总体结果表明，在不同类型的道德情境中，人们的关注焦点和心理加工机制存在差异——在电车困境中，人们更关注“决策主体”（AI 或人类）的差异；而在天桥困境中，他们更重视行为方式（行动 vs. 不行动）的道德意义。这说明个体在不同情境下可能激活不同的认知与情绪加工系统，从而形成对 AI 与人类行为的不同道德判断模式[3]。在研究“护士面临不道德工作气候”时发现，在制度或文化压力主导下，道德规范往往被“中和”或降权处理，不当行为则被重新解释为环境适应性策略，也就是说，如果功利主义决策来自人类会倾向理解其背后的道德意图[7]。

本研究将“决策主体”（AI vs 人类）与道德类型（功利主义 vs 义务论）作为自变量，目的在于探讨在道德两难情境中，青少年对不同主体所作决策的接受程度，从而识别人类在技术伦理判断中的差异化标准，为 AI 伦理规范设计提供心理学依据。

2 方法

2.1 参与者

采用 G*Power 3.1 进行双因素被试间方差分析的事前功效计算。设定效应量为 0.25（中等效应），显著性水平 $\alpha = 0.05$，统计功效= 0.80，组数= 4（决策主体：人工智能 vs.人类×道德类型：功利主义 vs.义务论）。计算结果显示，最小样本量需求为 128 名被试。

实际研究在河北省石家庄市某高中开展，共发放问卷 683 份（均为高中生），排除了 83 名未能正确回答相关验证问题、答案不完整的被试，最终样本为 600 名被试（40.17%为女性，平均年龄=16.20）。

因此，本研究的实际样本量远高于功效分析所需的最小样本量，能够在显著性水平 $\alpha = 0.05$ 下，以 0.80 及以上的功效检测中等及以上效应量，统计功效充足。

2.2 研究设计

本研究自变量为决策主体的不同（人工智能 vs 人类）、道德类型的不同（功利主义 vs 义务论），因变量为被试对该道德建议的接受程度，采用 7 点评分量表进行测量，识别人们对人工智能道德决策的接受程度是否具有选择性偏差。

2.3 研究材料

（1）电车困境、天桥困境实验材料：

使用改编的电车困境(Foot，1967)、天桥困境(Thomson, 1976) 数据选取电车与天桥困境的均值（实验材料详情见附录一）。其中对决策的接受程度采用 7 点 Likert 量表评分(1=完全不接受 2=很不接受 2=有些不接受 3=不太接受 4=不确定 5=比较接受 6=很接受 7=完全接受)。

（2）研究过程：

首先，被试随机分配到 4 个实验组，被试阅读指导语后，阅读经典的电车困境、天桥困境，分别填写对人工智能决策、对人类决策的接受程度，得出两个困境的接受度均值。

其次，人口统计学包括性别、年级、年龄，对人工智能的熟悉程度（采用 7 点 Likert 评分（1=完全不熟悉，7=完全熟悉））为了更好的筛选数据，排除对人工智能熟悉度为“完全不熟悉、很不熟悉、有些不熟悉 、不太熟悉”的被试，以免影响本研究的结果。

最后，并告知学生本次实验是虚拟的情境。

2.4 结果

在研究一中，采用 2（决策主体：人类 vs. 人工智能）×2（道德类型：功利主义 vs.义务论）的完全被试间实验设计。自变量为决策主体与道德类型，因变量为被试对道德建议的接受度。

对接受度均值进行两因素方差分析结果显示，决策主体的主效应显著，$F(1,596)=172.26$，$MSE=1.47$，$p<0.001$，偏 $\eta p^2=0.22$。具体而言，被试在“人类决策”条件下的接受度（M=5.12，SD=1.33，95% CI [5.00, 5.26]）显著高于“人工智能决策”条件（M=3.82，SD=1.15，95% CI [3.69, 3.96]），$t(596)=13.13$, $p<0.001$。

道德类型的主效应不显著，$F(1,596)=0.09$，$MSE=1.47$，$t(596)=0.30$，$p=0.76$，偏 $\eta p^2=0.000$，表明功利主义（M=4.46，SD=1.54）与义务论（M=4.49，SD=1.25）建议在整体上被接受的程度无显著差异。

然而，决策主体与道德类型的交互作用显著，$F(1,596)=30.83$，$MSE=1.47$，$t(596)=5.55$，$p<0.001$，偏 $\eta p^2=0.05$（见图 1）。

进一步的简单效应分析表明，在功利主义情境下，被试对人类决策的接受度（M=5.38，SD=1.14，95% CI [5.19, 5.58]）显著高于对人工智能决策的接受度（M=3.53，SD=1.32，95% CI [3.34, 3.73]），$t(596)=14.3$, $p<0.001$；而在义务论情境下，被试对人类决策的接受度（M=4.86，SD=1.45，95% CI [4.67, 5.06]）略高于人工智能决策的接受度（M=4.11，SD=0.87，95% CI [3.92,

4.31]），t(596)=5.6, p<0.001（见图 1）。

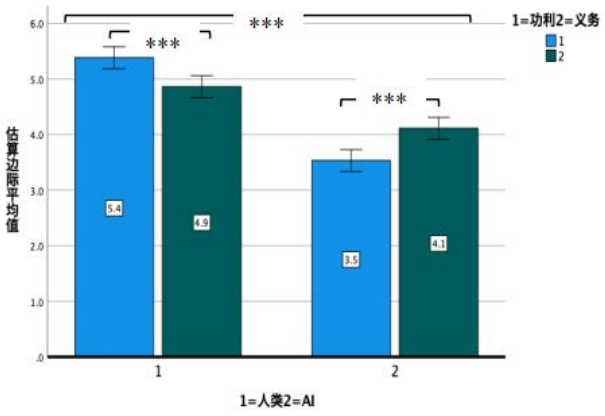

图 1. 决策主体（人类 vs 人工智能）与道德类型（功利主义 vs 义务论）对接受度的交互作用（研究 1），误差线表示 95% 置信区间，决策主体主效应：F(1, 596)=172.26, ***p < 0.001；决策主体与道德类型交互：F(1, 596)=30.83, ***p < .001

3 讨论

本研究通过 2（决策主体：人工智能 vs 人类）×2（道德类型：功利主义 vs 义务论）的被试间设计，探讨了青少年在道德困境情境中对不同主体（AI 与人类）决策的接受差异。结果表明，决策主体的主效应显著，青少年总体上对人类决策的接受度显著高于人工智能，这一发现与以往研究一致（Zhang et al.,2023），反映出个体在面对 AI 决策时存在“道德信任缺口”。即使 AI 表现出相同的决策行为，被试仍倾向认为其不具备人类的道德意图与责任意识（Malle et al., 2015）。

另一方面，道德类型的主效应不显著，说明青少年在面对功利主义与义务论判断时并未表现出一致的偏好。这与 Greene（2007）的双加工理论部分吻合：青少年道德判断尚处于情感与理性加工并存的过渡阶段（Krettenauer, 2011），功利主义与义务论取向可能尚未稳定成型。

更为重要的是，决策主体与道德类型的交互作用显著。被试对人类在功利主义情境下的接受度显著高于 AI，提示青少年对“为更大利益而牺牲个体”的决策若来自人类，会倾向理解其背后的道德意图；但当同样的决策来自 AI 时，则易被视为“冷漠”与“非人性化”。这一现象与 Hakimi et al. (2020)的研究一致，他们发现人们更易对 AI 决策的功利主义行为赋予负面评价，而对人类决策表现出更多道德宽容。

综合来看，本研究验证了“人机决策的道德不对称效应”，即使 AI 作出与人类相同的道德行为，其仍被认为缺乏道德主体资格。此结果揭示了青少年在 AI 伦理评估中的心理特点：他们一方面认可 AI 的逻辑决策能力，另一方面仍将道德责任视为“人类独占”的领域。这种倾向或反映出青少年在社会化过程中尚未形成稳定的“技术伦理归因机制”(杨雪 & 宋佳殷, 2025)。

未来研究可进一步探讨个体差异变量（如道德敏感性、自我控制、AI 态度）在该效应中的调节作用。此外，可采用混合设计或眼动追踪技术，以更全面揭示青少年在 AI 道德判断过程中的认知与情感机制。

参考文献

[1] 杨雪, & 宋佳殷. (2025). 家庭教养方式、亲子关系与青少年社会心理发展. 人口学刊, 47(1), 78-93. http://doi.org/10.16405/j.cnki.1004-129X.2025.01.006

[2] Krettenauer, T. (2011). The dual aspects of moral identity: Moral motivation and moral cognition in adolescence. Developmental Psychology, 47(6), 1619 - 1632.

[3] Zhang, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral Judgments of Human vs. AI Agents in Moral Dilemmas. Behavioral Sciences, 13(2) http://doi.org/10.3390/bs13020181

[4] Noh, D. (2023). Misunderstanding Minds: People's Attributions of Intentionality to AI Systems. AI & Society, 38(3), 649 - 663. https://doi.org/10.1007/s00146-023-01537-0

[5] Hertz, Nicholas, and Eva Wiese. “Good Advice Is beyond All Price, but What If It Comes from a Machine?” Journal of Experimental Psychology: Applied, 31 Jan. 2019, https://doi.org/10.1037/xap0000205. Accessed 30 June 2019.

[6] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction, 117 - 124.

[7] Hakimi, H., Joolaee, S., Ashghali Farahani, M. et al. Moral neutralization: Nurses’ evolution in unethical climate workplaces. BMC Med Ethics 21, 114 (2020). https://doi.org/10.1186/s12910-020-00558-3