青少年对人工智能给予自我与他人建议的接受度差异

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：本研究采用建议对象（“自己”vs“他人”）的单因素被试内实验设计，探讨青少年在接受人工智能道德建议时是否存在自他差异。160 名有效样本。被试阅读电车困境与天桥困境，并以 7 点量表评估人工智能分别针对“自己”与“他人”的建议接受度。结果显示，两类接受度均近似正态分布，可采用参数检验。配对样本 t 检验表明，被试对人工智能给“自己”的建议接受度显著低于给“他人”的建议，$t(159)=-9.86$，$p<0.001$，效应量 $d=1.68$。青少年显著更倾向接受人工智能给予他人的道德建议。

关键词：人工智能；建议对象；青少年；功利主义；义务论

DOI：10.64216/3080-1516.25.10.073

1 引言

从心理学视角出发，人工智能与人类道德判断之间呈现出认知、情感与行为三重层面的复杂交互机制。AI 不仅能模拟人类的道德推理，还可能通过技术介入深度塑造个体的道德判断方式。例如，“Socratic AI”即通过类苏格拉底式的对话引导，促使个体在反思中形成道德立场，其功能并非直接提供“正确答案”，而是支持用户构建个体化的道德逻辑。这种 AI 辅助式道德反思模式，既拓展了传统道德判断的情境维度，也对道德心理学的经典研究范式提出了挑战。随着 AI 在社会各领域的广泛嵌入，其行为与决策所引发的伦理问题日益突出。研究表明，个体在评估 AI 行为的道德性时，会综合考虑 AI 的行为意图、结果后果、自主性水平等因素，且相较于对人类的道德评价标准，人们往往对 AI 行为持更为严格的审视态度，特别是当 AI 的行为造成负面结果时[1]。

道德双标，是指个体在面对结构相似的道德情境时，因所涉对象身份不同，而采用不一致的道德评判标准。这种倾向本质上体现为一种带有社会选择性或关系偏向性的价值判断，即基于对人际关系亲疏、角色归属或社会身份的差异，施加不同的道德标准[2]。从描述性角度来看，道德双标指的是：当个体面对在关键特征上高度相似的道德情境时，往往会根据与情境中当事人之间的亲密程度或社会关系远近，施加不对等的道德评判标准。具体表现为：对具有亲密关系的对象倾向于采取更为宽容或理解性的道德判断，而对陌生人或社会关系较弱者则往往采取更为严苛甚至苛责的标准[3]。这一判断偏差不仅普遍存在于日常人际交往中，在更宏观的社会认知与伦理评估过程中亦屡见不鲜。例如，当个体在要求他人遵守某一规范的同时，自己却违反了在本质上类似的规则，这种前后不一的行为便构成了典型的道德双重标准。该现象与中国文化特有的“差序格局”密切相关：该理论认为，人们在进行道德评价时往往以自我为中心，依照与他人的亲疏关系层级逐级施加差异化的道德判断。基于这种“亲疏有别”的心理结构，个体倾向于对亲密关系者采取更宽容的标准，而对关系疏远者则更为苛刻。这种文化嵌入式的评价模式不仅体现在人际伦理互动中，在技术伦理判断中亦发挥着潜在作用，例如公众在评估人工智能行为的道德性时，也可能基于“人机关系距离”做出差异化评价。

实证研究表明，当人工智能的行为对个体自身或其亲密他人造成不利后果时，个体更倾向于对 AI 进行严厉谴责，并要求其承担更高程度的道德责任。相较之下，若受影响的对象为陌生人或社会关系较远的个体，人们则更容易表现出容忍态度，甚至在某些情境中认可其以结果导向为取向的功利主义决策。这一差异揭示出道德评价中的“关系偏向机制”，即个体的亲疏认知显著调节其对 AI 行为的伦理判断倾向[4]。此类道德评判中的标准不一致反映出典型的道德双重标准倾向，也进一步揭示，在人工智能伦理的研究中，社会认知偏差可能对个体判断过程产生干扰性影响，值得引起重视。此外，已有诸多国外研究从社会心理学与道德哲学视角，对道德双标现象进行了系统探讨，强调其在群体认同、角色归属与责任归因等方面的理论价值。Uhlmann et al. (2009) 通过系列实验指出，面临资源分配、责任归因等人们在道德判断中会受到“道德行为者身份”与“社会关系远近”的系统影响[5]；Effron & Monin (2010) 则发现，个体在“自我”与“他人”间表现出明显的道德容

忍差异，尤其在情境时更为显著。这些研究共同揭示了：道德判断不仅仅是对客观行为的理性评估，更是一种受到情感、关系与社会角色调节的心理过程[6]。在青少年群体中，道德双标现象可能更加突出。这一阶段个体的道德认知仍处于发展过程中，亲疏认知、自我中心与关系意识等因素可能进一步放大其在不同道德主体之间的评判差异。因此，在研究青少年对人工智能决策的接受度时，有必要系统考察其是否存在“对自己/亲密群体宽容，对他人/陌生群体苛刻”的道德双标倾向，并分析其与自利偏差、责任归因等变量的关系。

社会心理学中的“自利偏差”指个体在进行因果归因时表现出的系统性倾向：即更容易将积极结果归因于自身的内在因素（如能力、努力、积极态度），而将消极结果归因于外部环境（如运气、他人干扰等）[7]。该偏差在对他人行为的判断中同样存在镜像效应——即当他人取得积极结果时，个体倾向于归因于外部条件；而当他人出现负面行为后果时，则更倾向于归因于其内部特质[8]。

在人工智能相关的伦理判断情境中，个体同样可能表现出自利归因偏差。当 AI 提出的建议或决策结果对个体自身具有明显利益时，即便该建议潜藏道德争议或伦理风险，个体也更倾向于予以接受；而当 AI 的建议主要有利于他人、却对自身不利时，接受度则显著下降。这种以自我利益为导向的选择偏好表明，在技术道德判断中，自利偏差不仅影响个体对行为结果的道德归因，也可能改变其对责任分配与伦理正当性的评估[9]。由此可见，识别与理解自利偏差在 AI 决策接受过程中的作用机制，对于揭示人工智能与人类道德判断之间的动态互动关系具有重要的理论与实践意义。

2 方法

2.1 参与者

采用 G*Power 3.1 配对样本 t 检验分析，设定效应量为 0.25，显著性水平为 0.05，power 值为 0.80，测量次数为 2（建议对象：“自己”vs“他人”），计算所需样本量为 128 名被试。

实际研究在河北省石家庄市某所高中开展，共发放问卷 194 份（均为高中生），排除了 34 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 160 名被试（43.13%为女性，平均年龄=16.16）。所有被试智力正常，没有情绪相关障碍或精神疾病，同时没有参加过类似实验。

2.2 研究设计

本研究采用建议对象（人工智能给予建议的对象为“自己”vs“他人”）的单因素被试内实验设计。自变量为建议对象，因变量为被试对人工智能建议的接受程度。

2.3 研究材料

2.3.1 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)（详情见附录一）其中对人工智能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)。为了更好的筛选数据，因为对人工智能不熟悉的被试，从而影响本研究的结果。

2.3.2 过程

首先，被试随机分配到 2 个实验组（功利主义实验组与义务论实验组），被试阅读指导语后，阅读经典的电车困境、天桥困境，分别填写人工智能决策对“自己”、“他人”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

为检验变量分布的正态性，采用 Shapiro－Wilk 检验。结果显示，无论是“自己接受度”（W =0 .913, p < .001）还是“他人接受度”（W = 0.957, p<0.001），均显著偏离正态分布。由于该检验对大样本较为敏感（Razali & Wah, 2011），两变量分布基本对称、未发现极端离群值，因此可认为数据近似正态分布，可进行参数检验（配对样本 t 检验）。

结果显示，被试对人工智能给“自己”的建议接受度（M = 3.41，SD = 0.84），显著低于对“他人”的接受度（M = 4.72，SD = 1.51）。进一步的配对样本 t 检验结果表明，两者差异具有统计学意义，t(159) = －9.86，p<0.001，平均差为－1.31，95%CI= [－1.57, －1.05]，效应量 Cohen’s d = 1.68，为极大效应量。结果显示：青少年显著更倾向于接受人工智能给予“他人”的道德建议，而非给予“自己”的建议。

注：df 为自由度，Mean Diff 为平均差，Cohen’s

d 效应量为 1.68，Hedges 修正为 1.68，

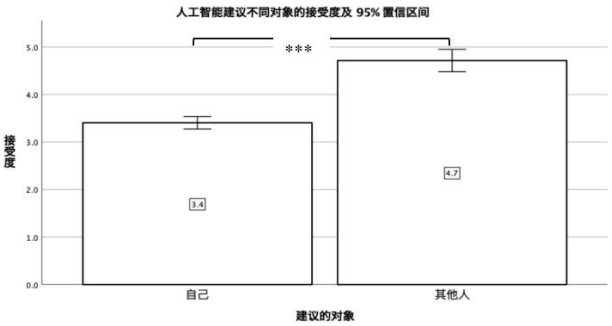

图 1 人工智能建议不同对象（自己 vs 其他人）的接受度均值，误差线表示 95% 置信区间。* p <0.05， p <0.01，*** p <0.001。**

3 讨论

青少年在接受人工智能的道德建议时呈现明显的自他差异：他们显著更愿意接受 AI 针对“他人”的建议，而对 AI 针对“自己”的建议保持更高谨慎性。一方面，青少年在道德判断中普遍存在自利偏差（关健，2025），即在涉及自身的情境中采取更高的谨慎标准，而在涉及他人时更容易放宽判断标准。在本研究中，当青少年面对 AI 给自己的建议时，他们需要直接承担潜在的道德或结果责任，因此更倾向于质疑 AI 的判断；而当建议对象是他人时，责任不在自己，风险知觉降低，从而更容易接受 AI 的建议。本研究中显著的效应量（d = 1.68）表明，这种自利偏差在 AI 场景下被进一步放大。

另一方面，从人工智能特性来看，青少年普遍认为 AI 缺乏情感、意图与道德动机（Fan et al., 2024），不具备完整的道德主体性。AI 在青少年的心理模型中更像是基于规则运算的机械代理，而不是能够承担道德责任的主体。因此，当 AI 的建议与自身相关时，青少年更容易将其视为“不可预测、缺乏动机与责任归属”的输入，从而保持高度警惕；而在与他人有关的判断中，这种主体性不足所带来的不信任则因责任不在自身而减弱。这与研究一中已经发现的青少年整体上更信任人类而非 AI 的道德决策高度一致。

综上，本研究不仅揭示了青少年在人工智能道德建议上的稳定自他差异，更展示了这一差异背后的心理机制：包括自利偏差、对 AI 道德主体性不足的认知、风险知觉的情境放大作用。青少年对 AI 的不信任既具有普遍性，也具有情境依赖性：在涉及自身的判断中更强，在与他人相关的判断中有所缓解。这一发现对 AI 在教育、心理咨询与校园治理情境中的应用具有重要启示——特别是在向青少年提供个体化建议时，AI 系统需要增强解释性、提升情境共情能力，并提高决策透明度，以降低青少年对“AI 给自己建议”的心理抵触与风险感知。

参考文献

[1]Fan, Z., Li, X., & Zhang, J. (2024). People's Moral Judgments Toward Different Types of Artificial Agents. Journal of Moral Psychology, 12(1), 21－34.

[2]谭生莲.(2024).大学生道德双标现象的成因及其对策研究(硕士学位论文,华中科技大学).硕士 https://doi.org/10.27157/d.cnki.ghzku.2024.004129.

[3]费孝通.乡土中国[M].北京:作家出版社,2019,28-30.

[4]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

[5]Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. Judgment and Decision Making, 4(6), 476－491.http://journal.sjdm.org/91011/jdm91011

[6]Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? Personality and Social Psychology Bulletin, 36(12), 1618－1634.https://doi.org/10.1177/0146167210385922

[7]单远.(2013).基于消费者视角的顾客满意与顾客公民行为关系研究(硕士学位论文,杭州电子科技大学).硕士 https://kns.cnki.net/

[8]王志瑞，李少军，马宇轩 . 大学生在内卷化中的自利性归因偏差[J]. 公关世界，2021(20).

[9]关健,李文朴,何国华 & 张新安.(2025).人工智能反馈：文献述评与研究展望.外国经济与管理,47(03),83-100.https://doi.org/10.16538/j.cnki.fem.20240622.301.