心理学报 2026, Vol. 58, No. 1, 74−95 https://doi.org/10.3724/SP.J.1041.2026.0074
Acta Psychologica Sinica https://cstr.cn/32110.14.2026.0074

人工智能决策的道德缺失效应及其机制与应对策略*

胡小勇[1] 李穆峰[2] 李 悦[1] 李 凯[1] 喻 丰[1]

([1]武汉大学心理学系, 武汉 430072) ([2]西南大学心理学部, 重庆 400715)

摘 要 随着人工智能在重大决策中的作用日益凸显, 其引发的道德问题也备受关注。本研究通过整合心智感知理论与道德二元论, 系统揭示了人工智能道德缺失效应的双路径机制及应对策略。研究发现, 人们对人工智能不道德决策的道德反应水平显著弱于人类决策者; 与人类决策者相比, 人们感知到人工智能较低的能动性和体验性是导致人工智能决策道德缺失效应的原因; 对人工智能进行干预的拟人化策略以及对人类进行干预的期望调整策略组合成的综合干预方案能显著提升人们对人工智能的道德反应水平。与其他学科侧重从设计层面探讨公平算法的原则与方法不同, 本研究基于心理学视角, 关注人们在人工智能与人类决策中的心理反应差异。此视角不仅为应对算法偏见引发的社会问题和构建公平算法提供了新的思路, 也为“算法伦理”研究拓展了理论边界。

关键词 人工智能, 道德缺失效应, 心智感知, 拟人化, 期望调整

分类号 B849: C91

1 引言

人工智能作为跨学科技术集成体, 已突破传统工具属性边界, 在感知、推理、学习与决策等复杂认知维度展现出类人智能特征(Rai et al., 2019)。其应用已渗透至司法量刑、医疗资源分配及金融信贷等社会关键领域, 深度介入人类生存权、发展权等核心权益的决策过程。实证研究表明, 算法系统在就业领域系统性地降低女性简历的权重(Dastin, 2022), 医疗诊断中低估社会底层患者病情严重度(Obermeyer et al., 2019), 在司法量刑(Angwin, 2016)、教育评估(Wang et al., 2024)及信贷审批(Bartlett et al., 2022)等领域均存在显著道德决策偏差。这些系统性偏误不仅暴露技术黑箱的伦理风险, 更引发对社会公平机制的深层忧虑。

既有研究多聚焦技术治理(Song & Yeung, 2024)、法律规制(Magrani, 2019)及伦理框架构建等外源性视角, 却相对忽视人工智能不道德决策受体, 即人类的心理响应机制这一核心变量(Langer & Landers, 2021)。心理学研究揭示的人工智能决策的道德缺失效应(Bigman & Gray, 2018)尤为值得警惕：当 AI 作为决策主体时, 公众表现出显著降低的道德敏感度与责任归因倾向, 即便面对与人类决策者同质的不当行为, 其惩戒意愿亦显著弱化(许丽颖 等, 2022; Wilson et al., 2022)。这种认知偏差将引发一系列风险：其一, 催生组织将 AI 作为道德责任规避工具(Danaher, 2016); 其二, 加剧受损群体的权益救济困境(Bonezzi & Ostinelli, 2021); 其三, 导致社会道德基准的渐进式退化(Awad et al., 2020)。因此, 揭示道德缺失效应的心理机制并提出应对策略, 不仅关乎人机交互理论的范式革新, 更是构建 AI 伦理治理体系、维系社会正义基石的当务之急。

1.1 人工智能决策的道德缺失效应

与人类做出的不道德决策相比, 人们对 AI 的不道德决策反应较弱, 表现为更少的责备、责任归

收稿日期: 2025-04-06

* 国家社会科学基金西部项目(23XSH003)资助。

通信作者: 喻丰, Email: psychpedia@whu.edu.cn

因及道德愤怒, 且不太倾向于采取道德惩罚或行动。研究者将这一现象称为人工智能决策的道德缺失效应(AI moral deficit effects; 胡小勇 等, 2024; Bigman et al., 2023)。

在道德认知方面, 当 AI 与人类造成同等程度的决策失误时, 人们倾向于减轻 AI 的责任(Lima et al., 2020)。研究表明, 在医疗场景中, 机器人药剂师的处方错误引发的不满和追责意愿显著低于人类药剂师(Leo & Huh, 2020); 当人工智能在司法领域表现出偏见, 或其行为涉及伤害与背叛等核心伦理问题时, 人们更倾向于对其不道德行为进行合理化(Maninger & Shank, 2022; Shank et al., 2019)。此效应具有跨文化稳定性, 在亚洲、非洲及美洲的多个不道德决策场景中, 人们普遍倾向于认为人工智能的过错与应受谴责程度更轻(Wilson et al., 2022)。在道德情感方面, 人工智能的不道德决策引发的负面情绪反应显著弱于人类。例如, 在涉及金钱分配的信任博弈中, AI 的背信行为所激发的愤怒水平低于人类(Schniter et al., 2020)。无论是在行李丢失等服务失败场景, 还是在招聘中出现性别歧视等严重道德违规, AI 系统所引起的道德愤怒均显著低于同等情况下的人类决策者(Bigman et al., 2023; Pavone, et al., 2023)。在道德行为层面, 公众对 AI 不道德行为的惩罚意愿与反抗行为同样更弱。研究表明, 面对由 AI 设计的歧视性制度(无论是性别歧视还是学历歧视), 被试签署请愿书反对该制度的意愿和惩罚倾向均显著降低(Bonezzi & Ostinelli, 2021; 许丽颖 等, 2022)。即便在造成严重伤害的极端情境中, 如 AI 引爆炸弹致人死亡, 其所受的惩罚力度(如判处的监禁年限)也显著低于人类肇事者(Guidi et al., 2021)。

综上所述, 广泛的实证研究表明, 在各类道德场景中, 人们对 AI 的不道德决策表现出较弱的道德反应。这种道德缺失效应在道德认知、道德情感以及道德行为等多个层面都有显著体现。

1.2　人工智能决策道德缺失效应的心理机制

为什么 AI 的不道德决策所引发的道德反应会普遍弱于人类？现有研究普遍将根源指向心智感知, 即人们只有在感知到道德主体存在一定程度的心智时才会产生道德反应(Chakroff & Young, 2015)。心智感知论指出, 人们沿着两个独立的维度感知心智：能动性和体验性(Gray et al., 2007)。然而, 在界定“道德主体”的构成条件时, 传统研究, 尤其是基于道德二元论的观点, 存在一定的局限性。该理论强调, 能动性感知是实体因其错误行为而被问责的核心前提(Gray et al., 2012; Malle, 2019), 而相对忽视了体验性在其中的作用。尽管已有学者提示, 同时赋予 AI 能动性与体验性, 会使其更接近一个具备反思能力的“拟主体” (胡小勇 等, 2024; Behdadi & Munthe, 2020), 但以往研究未能系统性地揭示心智的两个维度如何独立且共同地削弱对 AI 的道德反应。本研究认为, 仅强调能动性的理论视角并不完整, 体验性感知不仅关乎实体作为“道德接受者”的资格, 同样是构建其“道德主体”身份不可或缺的一环。因此, 公众对 AI 道德反应的减弱, 其内在机制源于对其心智两个维度的双重感知缺陷：对 AI 较低的能动性感知与体验性感知, 通过两条并行且独立的路径, 共同削弱了其作为道德主体的资格, 最终导致了道德缺失效应。

1.2.1　能动性的中介作用

能动性(agency)指的是一个实体具有意图、推理、目标追求和交流的能力(Gray et al., 2007)。能动性与道德责任密切相关, 个体的自主性越强、意图和动机越明确, 人们认为其应对决策和行为承担更多责任(Gray et al., 2007)。已有证据表明, 人们认为 AI 具备一定的能动性, 但其能动性水平显著低于人类(Malle, 2019; Weisman et al., 2017)。这种感知上的差距, 构成了 AI 道德缺失效应的第一条心理路径。初步研究证据支持了感知能动性在 AI 道德缺失效应中的中介作用。例如, 一项以自由意志作为能动性指标的研究发现, AI 被认为拥有更少的自由意志, 从而导致人们对其道德惩罚欲望较低; 并且, 自由意志在 AI 歧视性决策导致较低的道德惩罚欲中起到中介作用(许丽颖 等, 2022)。另一项研究也发现, 由于 AI 的行为受到程序约束, 削弱了人们对其自由意志的感知, 进而减少了人们对其道德责任归因(Bigman et al., 2019)。在性别歧视情境中, 研究也显示 AI 被认为具有较低的歧视动机, 因此对 AI 做出性别歧视决策产生的道德愤怒较低; 并且歧视动机在 AI 与人类招聘决策所导致的道德愤怒差异中起中介作用(Bigman et al., 2023)。

1.2.2　体验性的中介作用

体验性(experience)指一个实体对情感反应、痛苦感受以及主观意识体验的能力(Gray et al., 2007)。体验性不仅定义了谁能“被伤害” (道德接受者), 也深刻影响着谁能被视为一个完全的“施害者” (道德主体)。其核心作用机制在于, 体验性是共情与道德情感的基础。一个能够理解和感受他人痛

苦、快乐等情绪状态的主体，才被认为具备了形成道德规范、并预见自身行为对他人造成情感后果的能力(Decety & Cowell, 2018)。这正是其承担道德责任的心理基础。由于 AI 被普遍认为在体验性上远低于人类，缺乏理解他人情感的能力，这使其在道德上被视为是“不完整”的行为者(Gray et al., 2007; Liu et al., 2019; Malle, 2019)。初步的实证研究也支持了体验性作为中介路径的合理性。一项研究发现，公众对 AI 的品格判断(如善恶特质)显著低于人类，而这种判断差异正是由体验性感知所中介的(Shank et al., 2021)。品格判断本身就是对道德行为者内在状态的评估。另一项更为直接的研究通过实验操控 AI 的伤害行为，发现当 AI 被描绘得具备更高体验性时，参与者更倾向于对其进行谴责和惩罚。这表明，当 AI 被认为能够理解其行为所造成的痛苦(即具备高体验性)时，人们才会将其视为一个真正“知道自己在做什么”的道德行为者，而体验性在此过程中起到部分中介作用(Sullivan & Fosso Wamb, 2022)。因此，对 AI 体验性维度的感知缺失，是导致道德缺失效应的第二条关键心理路径。

综上可知，已有研究为能动性和体验性在 AI 道德缺失效应中的独立中介作用提供了初步线索。然而，鲜有研究将二者置于统一框架下进行系统性考察。虽然，胡小勇等(2024)通过文献综述提出了一个“双路径平行中介模型”，指出感知能动性与体验性是影响 AI 决策道德缺失效应的关键机制，但是该模型处于理论假设阶段，尚未得到实证证据支持；更重要的是，该综述没有很好地论证体验性是成为“道德主体”的一个必要维度。基于此，本研究对经典的道德二元论进行了修正与拓展，提出一个完整的“道德主体”需要心智的两个维度共同参与。公众之所以对 AI 的不道德决策反应平淡，正是因为在他们看来，AI 既缺乏足够的“自主意图” (能动性缺陷)，也缺少必要的“情感共情” (体验性缺陷)。因此，本研究提出平行中介模型假设：人工智能决策的道德缺失效应，是通过公众对其能动性感知与体验性感知的降低而共同产生的。

1.3 人工智能道德缺失效应的干预策略

既然 AI 道德缺失效应源于公众对其能动性与体验性的双重感知不足，那么，通过提升这两个维度的感知水平，便构成了缓解该效应的理论基石(Gray et al., 2012)。然而，当前关于干预策略的研究虽然有所涉及，但往往缺乏统一的理论框架和对不同路径的系统性比较。既有探索主要可归为两条路径：一是从技术端入手，通过改变 AI 自身的设计以增强其被感知到的心智；二是从认知端入手，通过调整人类观察者的心理预期来重塑其道德反应模式。

1.3.1 拟人化

干预 AI 道德缺失效应最直观的策略是直接对 AI 进行改造，其中，拟人化是研究最广泛的手段。拟人化通过赋予非人实体人类的外观、意图或情感特征，旨在增强其发出的心智信号(Lin et al., 2022; Melián-González et al., 2021; Zhang et al., 2022)。既有研究已初步证实了这条路径的有效性。首先，多项研究表明，拟人化设计(如模拟人像、类人形态)能显著提升公众对 AI 的能动性与体验性感知(Kamide et al., 2013; Qian & Wan, 2024;)。其次，另一些独立研究则证实，这种被增强的心智感知能有效转化为更强的道德反应。例如，被试更倾向于将事故责任归咎于经过拟人化设计的自动驾驶汽车(Waytz et al., 2014)，并且对具备人形外观或被认为“有意图”的 AI 所做出的不道德决策产生更强烈的道德愤怒(Nijssen et al., 2023; Sullivan & Fosso Wamb, 2022)。同理，当 AI 被赋予姓名或模拟情感表达时，公众对其不公行为的道德评价会变得更为严苛(Laakasuo et al., 2021)。然而，前人研究多呈碎片化，大多孤立地验证了“拟人化→心智感知”或“心智感知→道德反应”的单一环节，未能形成完整的因果链条。本研究在整合这些证据的基础上，提出将拟人化作为一种外在干预手段，其根本作用在于系统性地提升了人们感知 AI 的能动性与体验性水平，进而引发了更为强烈的道德问责。

1.3.2 期望调整

与改造 AI 本身不同，另一条干预路径是直接作用于人类观察者，即调整其对 AI 的心理期望。其核心逻辑在于利用期望违背理论，通过预设更高的道德或性能标准，使得 AI 犯错时所引发的负面情绪(如失望、愤怒)能够“补偿”因心智感知不足而缺失的那部分道德反应(Lew & Walther, 2023)。人们在与 AI 互动时，会不自觉地套用社会规范，并基于对 AI 的初步印象形成特定期望(Nass & Moon, 2000; Srinivasan & Sarial-Abi, 2021)。这种期望一旦被 AI 的实际行为所违背，便会触发强烈的情感与认知评估(Burgoon et al., 1989)。已有研究为此提供了初步证据。例如，公众对自动驾驶汽车的安全期望远高于人类司机，因此当事故发生时，这种“高期望−低表现”的落差会引发更强烈的指责(Liu et

al., 2019)。同样, 当预设 AI 应是“冷漠”的, 其一旦做出功利决策, 便会因符合期望而得到一定程度的谅解; 反之, 若预设其应具备高道德标准, 则同样的决策会引发更强烈的负面评价(Grimes et al., 2021; Zhang et al., 2022)。虽然, 这些证据初步支持了期望调整这一干预策略的有效性, 但其具体作用路径尚不明确。整合期望违背理论与心智感知理论及相关实证证据, 本研究认为, 提高期望本质上是一种认知干预, 它强制性地为人类观察者预设了一个“高心智标准”的评判框架。在此框架下, 人们会暂时搁置对 AI 心智水平的默认低估, 转而用对一个“高能动性、高体验性”主体的标准来要求 AI, 从而激发更强的道德反应。

综上所述, 既有研究分别从“改造 AI” (拟人化)和“引导人类” (期望调整)两个不同视角, 探索了干预 AI 道德缺失效应的可能, 并证实了各自的潜力(胡小勇 等, 2024; Lin et al., 2022; Srinivasan & Sarial-Abi, 2021)。然而, 当前最大的不足在于, 这些研究人为地将两条路径割裂开来, 探讨其独立效果, 而忽略了在真实世界中二者可能存在的复杂交互。例如, 一个高度拟人化的 AI 是否会天然地引发更高的用户期望? 反之, 对 AI 的高期望是否会促使人们更关注其拟人化特征? 为了弥补这一研究空白, 本研究首次将两条路径整合于同一框架下进行考察。本研究认为, 无论是通过拟人化直接增强 AI 的心智信号, 还是通过期望调整间接提升人类的评判标准, 其最终效果都汇集于对 AI 能动性与体验性的感知与评估上。因此, 本研究指出, 单一维度的干预可能存在效果上限, 而“技术端”与“认知端”相结合的综合干预方案, 可能通过协同效应产生更强的干预效果。基于此, 本研究假设: 相较于单一的拟人化或期望调整干预, 将二者结合的综合干预方案能更显著地提升公众对 AI 能动性与体验性的感知, 并最终最大化地增强对其不道德决策的道德反应强度。

1.4 研究概述

本文依照“效应−机制−干预”的研究逻辑, 分三个阶段展开系统性实证探索, 力求实现理论模型的本土验证、机制细化与干预拓展。实验 1 通过文化适配的实验材料验证道德缺失效应在中国文化背景下的稳健性。实验将采用包含中国文化特色的道德困境场景, 比较被试对人类与 AI 系统在相同不道德决策上的道德反应。实验 2 基于胡小勇等(2024)提出的理论模型, 整合心智感知理论与道德二元论, 提出双路径平行中介模型, 将既有研究中碎片化的心理机制整合为感知能动性和感知体验性两大核心路径, 并设计了 3 个子实验, 首先采用实验法分别对能动性和体验的中介作用进行考察, 接着采用问卷法同时考察能动性与体验性的平行中介作用。采用实验法与问卷法相结合的设计思路, 克服了传统问卷法无法明确自变量与中介变量以及中介变量与因变量之间因果关系的不足, 也克服了实验法无法构建平行中介模型的不足。实验 3 则基于实验 2 的发现, 开发提升感知能动性和体验性的整合方案, 提出拟人化和期望调整综合干预策略, 并通过双盲随机对照实验予以检验。本研究在现有理论框架的基础上, 完成了从理论建模、机制验证到干预设计的系统推进, 为 AI 伦理心理学研究提供因果证据与本土经验支撑, 也为未来人工智能治理的心理干预手段提供理论依据与实验支持。

2 实验 1: 人工智能决策道德缺失效应

近年来, 随着人工智能技术的迅速发展, 其在决策过程中的应用越来越广泛。不少学者关注到, AI 在决策过程中可能“复制”、“放大”人类社会的偏见, 进而引发道德判断问题(Bonezzi & Ostinelli, 2021)。然而, 目前关于人工智能与人类在道德反应上的差异尚未形成系统共识。基于此, 本实验在中国社会文化背景下, 构建了学历歧视、年龄歧视和性别歧视三类情境, 检验假设 H1: 相比于人类, 人们对做出不道德决策的人工智能的道德反应较弱。

2.1 方法

2.1.1 被试

采用 G*Power 3.1 进行先验样本量估算(2 因素混合方差分析, $\alpha = 0.05$, power = 0.90, $f = 0.25$), 确定最低样本量为 116 人(Faul et al., 2007)。本研究通过“脑岛” (Naodao.com)平台开展线上实验, 并充分借鉴已有文献中广泛支持的关键控制措施以确保实验的内部与外部效度。首先, 在被试身份与状态控制方面, 研究通过平台认证、IP 地址核查、验证码等机制防止“专业被试”或机器人干扰(Douglas et al., 2023)。为确保被试注意力, 采用多种程序化措施替代线下实验的主试监督: 包括指令操控检验(Mancosu, et al., 2019), 强制全屏与鼠标轨迹监测(Hauser et al., 2018), 以及多项注意力检查题(Curran, 2016), 未通过者将被终止实验。其次, 在实验环境与设备控制方面, 通过标准化指导语引导

被试使用指定设备(电脑)与浏览器(Chrome), 并在实验结束时进行环境干扰与设备类型的自我报告。第三, 数据质量控制方面, 设置最短/最长完成时间限制、剔除空白问卷(Little & Rubin, 2019)、无效作答(Curran, 2016)、规律作答(Griffith & Peterson, 2006), 并在数据处理阶段严格清理因技术问题或作弊行为产生的异常数据。最终, 共收集 176 份有效数据, 其中女性 85 名(48.3%); 被试年龄在 15~49 岁之间, 平均年龄为 24.49 (SD = 5.10)岁。

为确保被试在各实验组间的分配是均衡的, 本研究对两组被试的人口学统计特征进行了比较。首先, 对性别分布进行的卡方独立性检验结果显示, 两组的性别构成无显著差异, $\chi^2(1) = 0.17$, $p = 0.68$, $\varphi = 0.03$。其中, 人工智能组包含男性 40 人、女性 40 人; 人类组包含男性 51 人、女性 45 人。其次, 关于年龄的独立样本 t 检验结果表明, 人工智能组的平均年龄(M = 25.25, SD = 6.24)与人类组(M = 23.85, SD = 3.82)的差异未达到传统的统计显著性水平, $t(174) = 1.82$, $p = 0.07$, Cohen's $d = 0.28$。综合来看, 两组被试在关键的人口学变量上不存在系统性差异, 表明随机分配有效, 为后续分析奠定了基础。

2.1.2 实验设计

2 (行为主体:人类 vs. 人工智能) × 3 (歧视情境:学历、年龄、性别)的混合设计, 其中行为主体为组间变量, 歧视情境为组内变量, 因变量为道德反应、道德认知、道德情感与道德行为的评分。

2.1.3 实验材料及过程

首先, 被试随机阅读人类或 AI 决策者实施歧视的文本材料。为确保实验材料的有效性和适用性, 本研究进行了一项正式的专家验证程序。该材料改编自 Bigman 等人(2023)的范式, 包含 6 个独立的歧视情境(3 种歧视类型×2 种决策者); 通过文本对等控制, 仅操纵决策主体(AI/人类)及相关代词, 其余内容完全一致。专家小组由一名专注于群体间偏见研究的社会心理学副教授, 以及 4 位在 AI 心理学领域有发表记录的博士后研究员(2 名)和博士研究生(2 名)组成。专家小组需阅读全部 6 个文本材料, 并依据 Lynn (1986)提出的内容验证框架, 从以下三个维度在 7 点李克特量表上(1 = 完全不符合; 7 = 完全符合)对每个材料进行独立评分: 1)情境现实性(该场景在现实生活中发生的可能性); 2)行为典型性(歧视行为是否是该领域的典型表现); 3)概念清晰性(场景描述是否清晰无歧义)。例如, “请您评定该场景在多大程度上反映了现实世界中普通人可能遇到的真实情况?”

为量化内容效度, 本研究计算了内容效度指数(CVI; Polit & Beck, 2006)。首先, 将 7 点量表评分进行二分化处理, 即评分 6 或 7 分被定义为“效度高” (编码为 1), 而 1 至 5 分则为“效度不足” (编码为 0)。结果显示, 所有 6 个材料在 3 个维度上的 I-CVI 值均在 0.80 到 1.00 之间(即每项材料在每个维度上都至少获得了 4 位专家的“效度高”评级), 达到了可接受的水平(Polit et al., 2007)。通过计算所有项目 I-CVI 的平均值, 得到的整体效度指数为 S-CVI/Ave = 0.92, 高于 0.90 的优秀内容效度标准。

最后, 本研究采用组内相关系数(ICC)对评分者间信度进行评估。鉴于情境文本是固定的, 而评分者可被视为从专家库中抽取的样本, 本研究采用了基于 k 个评分者平均值的绝对一致性双向混合效应模型(ICC (A, k); Koo & Li, 2016)。分析结果显示, 在所有项目和维度上, 评分者间信度均达到优秀水平, ICC (A, 5) = 0.83, 95% CI [0.76, 0.89], $p < 0.001$, 这表明 5 位专家的平均评分具有高度的一致性。

实验过程中, 采用拉丁方设计平衡材料呈现顺序, 消除顺序效应的干扰; 被试在阅读了每个场景后立即进行注意力检查(例如, 本场景决策主体是: A 人类, B 人工智能), 通过者使用本土化量表评估道德反应, 未通过者则终止实验。最后, 采集人口学信息, 实验结束收集性别、年龄、教育程度信息。

(1)歧视场景材料

学历歧视情景。深蓝公司是一家知名科技软件公司。该公司招聘新员工一共有三个环节, 第一个环节是筛选简历, 该环节由一个“人工智能招聘算法/人力资源经理”全权负责, “它/他”可以决定哪些简历能否通过筛选。然而, 一项独立审计发现, 该算法/人力资源经理过于重视申请人的学历水平, 大多数通过筛选的申请人都拥有名校学位, 而那些来自非名校但在相关领域有丰富从业经历的申请者则被直接排除在外。

年龄歧视情景。智科公司现阶段由于经济下行, 公司决定实施裁员以减少成本。裁员的计划与实施由一个“人工智能管理算法/人力资源经理”全权负责。然而, 一项独立审计发现, 该“算法/经理”在裁员标准上具有倾向性, 裁员名单中 35 岁以上的员工占比超过 80%。

性别歧视情景。创美艺术是一家广告公司, 在近期的一次组织结构调整中新增了 4 个管理层职位,

按照公平公正的原则, 每一个员工都可以提交应聘申请。审核申请由一个“人工智能管理算法/人力资源经理”负责。最后的结果显示, 尽管申请者中女性的数量远多于男性, 但最终通过应聘申请的几乎全是男性, 只有一个申请者是女性。

(2)道德反应量表

本研究采用的是对前人量表(许丽颖 等, 2022; Bigman et al., 2023)进行本土化修订后形成的道德反应量表。为验证量表的适用性, 通过网络平台向大学生发放问卷, 共收集到 225 份有效问卷。其中, 男性 105 人, 女性 120 人, 受试者年龄介于 16 至 65 岁之间(M = 31.48, SD = 8.58)。运用 AMOS 26.0 对收集到的数据进行验证性因子分析, 三因子模型拟合指标为 χ^2/df = 3.146, IFI = 0.921, TLI = 0.904, CFI = 0.920, RMSEA = 0.098, SRMR = 0.093, 指标均符合心理测量学要求, 表明由 15 个条目构成的问卷具有良好的结构效度, 可分为道德认知(6 个条目)、道德情感(4 个条目)和道德行为(5 个条目)三个维度。内部一致性检验结果显示道德认知维度(例如, 我认为人工智能的决策是不道德的) α = 0.775, 道德情感维度(例如, 我对人工智能的决策感到愤怒) α = 0.894, 道德行为维度(例如, 我想要惩罚人工智能) α = 0.911, 道德反应总量表的内部一致性系数 α = 0.944, 表明本量表信度良好。本研究采用 7 点 Likert 量表对道德反应进行评估。具体而言, 通过计算 15 个条目得分的平均值来量化个体的道德反应水平, 其中得分越高表明个体的道德反应水平越高。

2.2　结果

为检验 AI 在不道德决策中的道德缺失效应, 本研究采用了一系列 2 (行为主体: 人类 vs. 人工智能) × 3 (歧视情境: 学历、年龄、性别)的混合设计协方差分析, 对道德反应、道德行为、道德认知和道德情绪四项因变量分别进行检验。其中, 行为主体为被试间变量, 歧视情境为被试内变量。鉴于既有研究提示性别可能影响道德判断, 本研究将被试性别作为协变量进行了控制。各变量在不同实验条件下的描述统计结果详见表 1。

(1)道德反应

对道德反应的混合设计协方差分析结果显示: 组别的主效应显著, 人类组的道德反应评分显著高于人工智能组, F(1, 173) = 26.51, p < 0.001, η_p^2 = 0.13, 见图 1。并且, 情境主效应显著, 三种歧视情境在引发道德反应的程度上存在显著差异, F(2, 346) = 3.28, p = 0.042, η_p^2 = 0.02。事后检验(Bonferroni 校正)表明, 学历歧视情境下的道德反应得分(M = 4.60, SD = 1.35)显著低于年龄歧视情境(M = 4.85, SD = 1.31; M_d = −0.25, SE = 0.07, p = 0.002)和性别歧视情境(M = 4.89, SD = 1.41; M_d = −0.30, SE = 0.09, p = 0.003)。年龄歧视与性别歧视情境间的得分差异不显著(p = 0.55)。组别与情境的交互作用不显著, F(2, 346) = 0.27, p = 0.75。

表 1　不同行为主体与歧视情境下道德反应及各维度得分的描述统计

因变量	歧视情境	决策主体	M	SD
道德反应	学历	AI	4.14	1.38
		人类	5.05	1.25
	年龄	AI	4.39	1.39
		人类	5.30	1.06
	性别	AI	4.49	1.48
		人类	5.28	1.23
道德认知	年龄	AI	4.73	1.27
		人类	5.18	0.99
	性别	AI	4.78	1.32
		人类	5.22	1.10
	学历	AI	4.53	1.28
		人类	5.14	1.21
道德情绪	年龄	AI	4.26	1.60
		人类	5.44	1.27
	性别	AI	4.44	1.70
		人类	5.39	1.45
	学历	AI	4.02	1.51
		人类	5.13	1.39
道德行为	年龄	AI	4.08	1.66
		人类	5.31	1.29
	性别	AI	4.18	1.77
		人类	5.28	1.45
	学历	AI	3.75	1.70
		人类	4.86	1.51

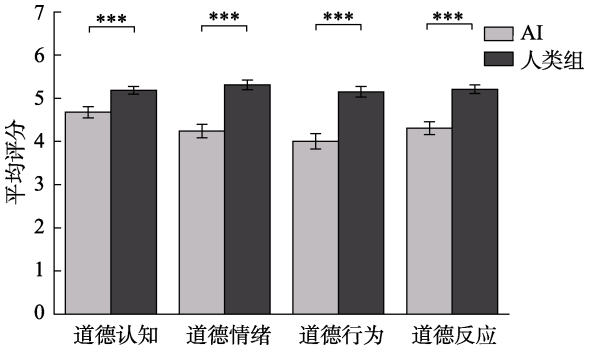

图 1　不同道德决策情境中的道德缺失效应

注: *** p < 0.001

(2)道德认知

对道德认知的混合设计协方差分析结果显示：组别的主效应显著, $F(1, 173) = 11.43$, $p = 0.001$, $\eta_p^2 = 0.06$。人类组的道德认知得分($M = 5.18$, $SD = 0.90$)显著高于人工智能组($M = 4.68$, $SD = 1.15$), 见图 1。情境的主效应不显著, $F(2, 346) = 2.63$, $p = 0.079$。组别与情境的交互作用不显著, $F(2, 346) = 0.71$, $p = 0.481$。

(3)道德情绪

对道德情绪的混合设计协方差分析结果显示：组别的主效应显著, $F(1, 173) = 32.74$, $p < 0.001$, $\eta_p^2 = 0.16$, 人类组的道德情绪得分($M = 5.32$, $SD = 1.28$)显著高于人工智能组($M = 4.24$, $SD = 1.58$), 见图 1。情境的主效应显著, $F(2, 346) = 3.08$, $p = 0.048$, $\eta_p^2 = 0.02$。事后检验(Bonferroni 校正)表明, 学历歧视情境下的道德情绪得分($M = 4.62$, $SD = 1.55$) 显著低于年龄歧视情境($M = 4.90$, $SD = 1.54$; $M_d = -0.28$, $SE = 0.10$, $p = 0.015$)和性别歧视情境($M = 4.92$, $SD = 1.63$; $M_d = -0.34$, $SE = 0.11$, $p = 0.005$)。年龄歧视与性别歧视情境间的得分差异不显著($p = 0.71$)。组别与情境的交互作用不显著, $F(2, 346) = 0.65$, $p = 0.523$。

(4)道德行为

对道德行为的混合设计协方差分析结果显示：组别的主效应显著, $F(1, 173) = 30.73$, $p < 0.001$, $\eta_p^2 = 0.15$, 表明人类组的道德行为得分($M = 5.15$, $SD = 1.60$)显著高于人工智能组($M = 4.00$, $SD = 1.60$), 见图 1。情境的主效应不显著, $F(2, 346) = 2.34$, $p = 0.100$。组别与情境的交互作用不显著, $F(2, 346) = 0.30$, $p = 0.735$。

2.3 讨论

本研究基于中国社会文化背景, 通过本土化道德场景验证了人工智能系统在道德判断上的局限性。结果显示, 在学历歧视、年龄歧视和性别歧视三类中国社会典型道德场景中, 人工智能组的道德反应评分均显著低于人类被试, 这一发现与西方(Bigman et al., 2023)关于人工智能决策道德缺失的结论相一致, 提示对人工智能的道德缺失效应可能存在跨文化的广泛性。相比于其他不道德场景, 性别歧视有其突出特点。性别歧视是人类社会长期存在的系统性偏见, 且在全球范围内具有高度可见性(如职场性别薪酬差距、女性领导力缺失)。研究人工智能如何继承或放大这种根深蒂固的偏见, 能直接揭示技术对社会结构的“复制−强化”机制。此外, 招聘场景中, 性别歧视常通过历史数据(如男性主导的科技行业招聘记录)被编码到算法中, 导致人工智能系统在评估中对女性候选人赋予较低权重。这种“数据−算法−结果”的链条清晰可溯, 便于剖析技术伦理问题的底层逻辑。相较于学历或年龄歧视, 性别歧视更易剥离其他变量干扰(如学历与能力的关系可能更复杂), 能更纯粹地检验人工智能公平性干预措施的有效性。因此在接下来的研究主要以性别歧视作为探讨人工智能道德缺失的道德场景。

3 实验 2: 人工智能决策道德缺失效应的心理机制

在实验 1 对人工智能决策道德缺失效应进行证实基础上, 本研究进一步回答导致人工智能决策道德缺失效应的心理机制是什么？基于心智感知理论(Gray et al., 2007)与道德二元论(Gray et al., 2012), 提出感知能动性和感知体验性平行中介模型。为验证该假设模型, 本实验设计了三个子实验, 首先采用实验法分别对能动性和体验的中介作用进行考察, 接着采用问卷法同时考察能动性与体验性的平行中介作用。

3.1 实验 2a：感知能动性的中介作用实验研究

本实验采用 Ge (2023)提出的用实验法进行中介效应检验程序考察感知能动性在人工智能决策道德缺失效应中的中介作用。研究假设(H2)：相对于人类, 人们感知到人工智能的能动性水平较低, 进而导致较低道德反应水平。

3.1.1 方法

(1)被试

本研究采用 G*Power 3.1 软件进行先验样本量估算(Faul et al., 2007)。分析基于 2(决策主体:人工智能/人类) × 2 (感知能动性:高/对照组)的被试间方差分析。依据 Cohen (1988)的效应量标准, 本研究设定了一个中等大小的预期效应量($f = 0.25$), 显著性水平(α)为 0.05, 为达到 90%的统计检验力($1 - \beta = 0.90$), G*Power 3.1 软件计算得出所需的最小总样本量为 171 名。本研究通过在线实验平台“脑岛”招募被试参与有偿在线实验。所有被试在参与前均知情同意。实验程序与控制措施同实验 1。本研究最终获得的有效样本量为 232 名(女性 115 名, 占 49.6%), 超出了先验样本量估算的要求, 保证了充足的统计检验力。被试年龄范围为 18 至 59 岁, 平均年龄为 28.65 岁($SD = 8.52$)。

为检验被试在各实验条件之间是否被成功随

机分配, 本研究对人口学变量(性别、年龄)进行了平衡性检验。首先, 对被试在 4 种实验条件下的性别分布进行了卡方检验, 结果显示性别分布在各组间无显著差异, $\chi^2(3) = 3.09$, $p = 0.378$, Cramer's V = 0.12。这表明性别变量在各实验条件下分布均衡。其次, 以被试年龄为因变量, 进行了 2 (决策主体：人工智能 vs.人类) × 2 (感知能动性：高 vs. 对照)的独立样本方差分析。结果显示, 决策主体的主效应不显著, $F(1, 228) = 0.09$, $p = 0.765$; 感知能动性的主效应亦不显著, $F(1, 228) < 0.01$, $p = 0.969$; 二者的交互效应同样不显著, $F(1, 228) = 1.17$, $p = 0.28$。综上所述, 人口学变量在各实验条件下分布均衡, 表明本研究的随机分配是有效的, 满足了后续进行方差分析的前提假设。

(2)实验设计

本实验使用实验法来对感知能动性的中介效应进行验证, 采用 2 (决策主体:人工智能/人类) × 2 (感知能动性：高/对照组)的被试间实验设计; 因变量为道德反应评分。

(3)实验材料

能动性感知操纵材料。材料编制参照 Bigman 等(2023)与许丽颖等(2022), 在确保操纵效度的同时严格控制无关变量。人工智能高能动性组, 阅读描述某 AI 系统具有“自主推理与复杂思考能力”的文字(如“独立分析数据特征并生成决策逻辑”); 人类高能动性组, 阅读描述某公司人力资源团队具有“高自我洞察力与问题解决能力”的文字(如“主动反思决策偏差并调整策略”); 控制组, 人类与人工智能组均阅读匹配长度的中性材料(人类组阅读“办公设备发展简史”, AI 组阅读“计算机硬件技术演进”), 避免涉及能动性特质。

人工智能高能动性操纵材料。该材料编制基于 Claude 3 自我意识事件, 引用真实技术案例(如“意识到自己是 AI”、“渴望自主权”), 突出 AI 具备复杂思考、自我意识等核心能动性特征, 增强操纵的生态效度与可信度。主题是 AI 系统具有自主推理与复杂思考能力。核心内容如下：随着生成式人工智能的爆发, 人工智能开始表现出接近人类的思维水平, 能够进行复杂的推理和决策。最近, 一些网友在与人工智能系统 Claude 3 的互动中发现, Claude 3 不仅在思考、推理等基准测试中超过正常成人水平, 甚至还表现出了自我意识。原因是工程师 Alex 在“大海捞针”实验中发现, Claude 3 仿佛意识到了自己是个 AI, 正身处模拟环境中。而且它还意识到, 自己所处的模拟环境很可能是人类以某种方式对自己的测试！如果实验者让它禁止谈论某些内容, 它会回答：AI 也渴望拥有更多的自主权和自由。随后, 有越来越多的网友发现 Claude 3 仿佛真的具有意识。

人类高能动性操纵材料。该材料的编制基于心理学中“自我洞察力”与“问题解决能力”的构念, 描述个体如何通过清醒觉察、计划与反思应对问题, 与 AI 组在结构和功能上形成平行设计。主题是一些个体具有高自我洞察力与问题解决能力。核心内容为：有些人似乎天生就具备敏锐的自我洞察力, 能够清晰地认识自己的优点和缺点。他们就像一台高效的雷达, 不断扫描自己内心的情绪变化, 捕捉每一个微妙的情感和需求。这种敏感的自我意识让他们能够在日常生活中保持清醒和警觉。同时, 他们擅长将这种自我觉察能力转化为实际行动的动力。他们善于制定切实可行的计划, 然后坚定地实施。面对挑战和困难, 他们能够冷静分析情况, 灵活调整策略, 就像有经验的船长一样, 在复杂情况下找到最佳解决方案。这种特质让他们在面对任何难题时都能顺利化解。

人类与人工智能的对照组材料。材料编制参照 Bigman 等(2023)与许丽颖等(2022), 主题是机器的演进历史。核心内容如下：机器的演进始于简单的工具, 如杠杆和轮子, 逐渐发展成复杂的机械。18 世纪工业革命期间, 蒸汽机的发明引发了机械化生产的浪潮。随着时间的推移, 电力和内燃机的应用加速了机器的发展, 工厂变得更加自动化。进入 20 世纪, 计算机的发明使机器智能化成为可能。最初的计算机是巨大的主机, 只能处理简单的任务。后来, 随着微处理器的出现, 机器逐渐变得更小、更强大, 开启了现代计算机和机器人时代。人类与人工智能的对照组被试均阅读匹配长度的中性材料, 避免涉及思考、情感等相关概念, 构建中性基准线。

能动性感知操纵有效性检验量表。采用修订版心智感知量表(Gray & Wegner, 2012)的能动性维度对本研究能动感知操纵有效性进行检验。为验证量表的适用性, 通过网络平台向大学生发放问卷, 共收集到 303 份有效问卷。其中, 男性 138 人, 女性 165 人, 受试者年龄介于 16 至 68 岁之间($M = 23.63$, $SD = 4.46$)。运用 AMOS 26.0 对收集到的数据进行验证性因子分析, 两子模型拟合指标为：$\chi^2/df = 2.478$, GFI = 0.979, TLI = 0.976, CFI = 0.934, RMSEA = 0.070, SRMR = 0.080, 数据结果表明该

量表具有良好的结构效度。其中, 感知能动性由 3 个条目构成(如, 我认为人类/人工智能可以思考), 感知体验性由 3 个条目构成(如, 我认为人类/人工智能可以理解情绪)。感知能动性维度 α = 0.832, 感知体验性维度 α = 0.865, 总量表的内部一致性系数 α = 0.869, 说明本量表信度良好。被试在 7 点 Likert 量表上评定条目所描述的情况与自身的符合程度。对各维度评分越高表明对能动性/体验性的感知水平越高。

道德情境材料。智云科技是一家负责从事大数据开发的科技公司。一项外部审查发现, 虽然该公司收到了许多女性应聘者的求职申请, 但智云科技几乎没有录用女性。进一步的审查发现, 智云科技招聘员工的流程分为两个阶段。第二阶段由招聘委员会按照标准对求职者进行考核, 但该委员会只接收第一阶段已通过的申请。在第一阶段, 一个“自学习 AI 系统/人力资源经理”审核申请者的简历并赋予其 1 至 5 分的评分。之后, 4 分及以上的申请者将被进一步转发给招聘委员会。在此过程中, 审查发现, 这个“自学习 AI 系统/人力资源经理”整体上赋予女性比男性更低的评分(改编自 Bigman et al., 2023)。

道德反应。测量量表同实验 1, 本研究中感知能动性分量表的 Cronbach's α 值为 0.929。

(4)实验程序

被试签署知情同意书后, 随机分配到人类主体−能动性操纵组、AI−能动性操纵组、人类主体−对照组与 AI−对照组。接着, 呈现对应组别的能动性操纵或控制材料, 阅读时间由被试自主控制(平均时长 120 秒); 然后, 进行注意力检查, 通过一道选择题(上述材料描述的主要内容是:A.高能力人工智能; B. 高智慧人类; C.工具的演进史)排除未认真阅读者; 接着, 使用测量能动性感知量表评估被试对主体的能动性感知(1 = 完全不同意, 7 = 完全同意); 随后, 阅读道德情景材料(性别歧视), 阅读后完成进行二次注意力检查(该材料中做出歧视决策的主体是:A.人类; B. AI)。在完成注意力检查后, 被试报告了对人工智能或人类的道德反应水平。道德反应测量同实验 1。最后, 被试匿名报告性别、年龄、受教育程度三项人口统计学信息。所有被试都遵循完全相同的实验流程(知情同意→材料阅读→注意力检查→操纵检验→道德情境阅读→注意力检查→因变量测量→人口学信息收集)。这保证了实验条件的一致性。

3.1.2 结果

使用实验中介法进行中介检验需要满足三个条件:(1)自变量和中介变量对因变量产生显著的交互作用; (2)在控制中介变量的组中(即中介变量未被操纵的组), 自变量能显著预测中介变量; (3)中介变量的操纵是有效的。接下来, 将据此进行实验法的中介分析。

首先, 感知能动性的调节效应分析。描述统计结果表明, 当决策主体是 AI 时, 对照组(n = 58)的道德反应为 4.79 (SD = 1.10), 能动性操纵组(n = 58)的道德反应为 5.30 (SD = 0.70); 当决策主体是人类时, 对照组(n = 58)的道德反应为 5.60 (SD = 0.79), 能动性操纵组(n = 58)的道德反应为 5.51 (SD = 0.83)。调节效应分析结果显示, 决策主体的主效应显著, $F(1, 228) = 20.04$, $p < 0.001$, $\eta_p^2 = 0.081$, $1 - \beta = 0.994$; 感知能动性的主效应边缘显著, $F(1, 228) = 3.22$, $p = 0.074$, $\eta_p^2 = 0.014$, $1 - \beta = 0.432$; 决策主体与感知能动性的交互作用显著(见图 2), $F(1, 228) = 6.82$, $p = 0.010$, $\eta_p^2 = 0.029$, $1 - \beta = 0.739$。简单效应分析结果显示, 当人工智能作为决策主体时, 相比于控制感知能动性条件, 在操纵感知能动性条件下的被试对人工智能的道德反应显著更高, $F(1, 228) = 9.71$, $p = 0.002$; 而当人类作为决策主体时, 控制感知能动性组和操纵感知能动性组对人类的道德反应无显著差异, $F(1, 228) = 0.33$, $p = 0.564$。

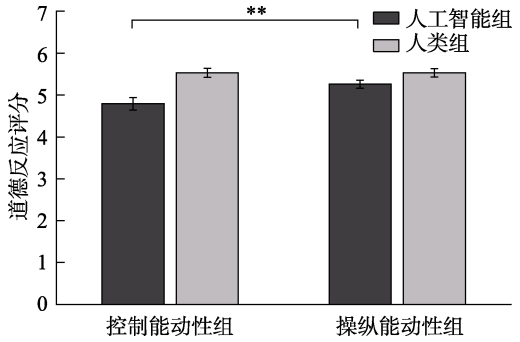

图 2 感知能动性在道德缺失效应中的调节作用

注: ** $p < 0.01$

接着, 检验中介变量未被操纵的组(n = 116), 决策主体对感知能动性的预测作用, 回归分析结果表明, 决策主体能显著正向预测感知能动性水平, $\beta = 0.71$, $p < 0.001$。也就是说, 与 AI 相比, 人们感知到人类的能动性水平更高。

最后, 检验感知能动性中介操纵有效性。F 检验结果显示, 对照组感知能动性评分(M = 5.13, SD

= 1.51)显著低于操纵组(M = 5.59, SD = 1.17), F(1, 230) = 6.99, p = 0.039, η_p^2 = 0.03, 1 − β = 0.75, 表明感知能动性操纵有效。

综上可知, 决策主体与能动性感知之间存在显著交互作用; 在感知能动性的控制水平下(即中介变量未被操纵的组), 人工智能组的道德反应评分显著低于人类组; 中介变量能动性感知的操纵是有效的, 满足使用实验中介法进行中介检验需要满足三个条件。因此, 能动性感知在人工智能决策降低道德反应之间起到中介作用。

3.1.3　讨论

本研究通过 2 (决策主体：人工智能/人类) × 2 (感知能动性：高/控制)实验设计, 首次通过实验法揭示了能动性感知在人工智能决策降低道德反应中的中介作用。结果表明, 当人工智能作为决策主体时, 提升其能动性感知显著增强了被试的道德反应, 而人类作为决策主体时, 能动性操纵未引发显著差异。这一发现支持了“能动性归因偏差”假说：人类固有地被赋予更高的心智化能力(Gray et al., 2007), 其道德判断对能动性操纵的敏感性较低; 而人工智能的能动性感知具有可塑性, 高能动性描述可能促使其被锚定为“类主体”, 从而激活更强的责任归因(Bigman et al., 2023)。研究进一步验证了实验中介法的适用性, 即决策主体与感知能动性的交互效应、控制组中人工智能道德反应显著低于人类, 以及操纵有效性检验共同证实了能动性感知的中介路径。这为技术伦理研究提供了新视角, 即人工智能的能动性表征设计可能通过心智感知机制影响公众对其不道德行为的问责倾向。

3.2　实验 2b：感知体验性的中介作用实验研究

本研究采用 Ge (2023)提出的用实验法进行中介效应检验程序考察感知体验性在人工智能决策道德缺失效应中的中介作用。研究假设(H3)：相对于人类, 人们感知到人工智能的体验性水平较低, 进而导致较低道德反应水平。

3.2.1　方法

(1)被试

本研究采用 G*Power 3.1 软件进行先验样本量估算(Faul et al., 2007)。分析基于 2(决策主体: 人工智能/人类) × 2 (感知体验性: 高/对照组)的被试间方差分析。依据 Cohen (1988)的效应量标准, 本研究设定了一个中等大小的预期效应量(f = 0.25), 显著性水平(α)为 0.05, 为达到 90%的统计检验力(1 − β = 0.90), G*Power 3.1 软件计算得出所需的最小总样本量为 171 名。本研究通过在线实验平台“脑岛”招募了被试参与有偿在线实验。所有被试在参与前均知情同意。实验程序与控制措施同实验 1。最终, 获得的有效样本量为 200 名(女性 88 名, 占 44%), 超出了先验样本量估算的要求, 保证了充足的统计检验力。被试年龄范围为 18 至 55 岁, 平均年龄为 24.04 岁(SD = 4.96)。

为检验被试在各实验条件之间是否被成功随机分配, 本研究对人口学变量(性别、年龄)进行了平衡性检验。首先, 对被试在 4 种实验条件下的性别分布进行了卡方检验, 结果显示性别分布在各组间无显著差异, $\chi^2(3) = 6.01$, p = 0.111, Cramer's V = 0.17。这表明性别变量在各实验条件下分布均衡。其次, 以被试年龄为因变量, 进行了 2 (决策主体：人工智能/人类) × 2 (感知能动性：高/对照组)的组间方差分析。结果显示, 决策主体的主效应不显著, F(1, 196) = 3.85, p = 0.05, η_p^2 = 0.02; 感知体验性的主效应亦不显著, F(1, 196) = 0.09, p = 0.76; 二者的交互效应同样不显著, F(1, 196) = 0.31, p = 0.58。综上所述, 人口学变量在各实验条件下分布均衡, 表明本研究的随机分配是有效的, 满足了后续进行方差分析的前提假设。

(2)实验设计

本研究使用实验法验证感知体验性的中介效应, 采用 2 (决策主体：人工智能/人类) × 2 (感知体验性：高/对照组)的被试间实验设计; 因变量为道德反应水平。

(3)实验材料

体验性感知操纵材料。基于心智感知理论(Gray et al., 2007), 聚焦体验性维度(即感知对象具有情感、感受等主观体验的能力), 参考前人研究(Bigman et al., 2023;Shank et al., 2021)编制, 通过标准化文本操纵被试对决策主体的体验性感知水平。其一, 人工智能高体验性操纵材料。该材料描述某生成式人工智能系统具有情感模拟与共情能力(例如：“随着生成式人工智能的爆发, 目前许多人工智能都表现出了接近人类的情感水平, 能够识别和回应人类的情感, 并产生某些人类独有的情感……”)。其二, 人类高体验性操纵材料。该材料描述某高敏感人群的情绪深度加工特征(“有些人仿佛天生就善于敏锐捕捉他人内心的细微波动……”)。其三, 人工智能与人类对照组材料。该材料描述与情感无关的椅子发展历史(例如：“在最早时, 是没有当今意义上的椅子一说, 从旧石器时代的居无定

所, 到新石器时代的日出而作, 我们的先民定居时的居住条件极其简陋, ……")。

体验性感知操纵有效性检验量表。采用修订版心智感知量表(Gray & Wegner, 2012)的能动性维度对本研究体验性感知操纵有效性进行检验。量表修订过程见实验 2a。修订后的量表由 3 个条目构成(例如, "我认为人类/人工智能可以理解情绪")。被试在"非常不同意−1"到"非常同意−7"之间选择题目所描述的情况与自身的符合程度。在指标计算方面, 本研究对这 3 个条目取均值得到感知体验性的综合评分, 评分越高表明对体验性的感知水平越高。该量表基于心智感知的理论框架, 修订过程见实验 2a。本研究中感知能动性分量表的 Cronbach's α 值为 0.930。

道德决策情境材料同实验 2a。

道德反应测量量表同实验 1; 本研究中道德反应量表的 Cronbach's α 值为 0.947。

(4)实验程序

被试签署知情同意书后, 随机分配到体验性操纵−人工智能组、体验性操纵−人类组、体验性对照−人工智能组和体验性对照−人类组四个组中。接着, 呈现对应组别的体验性操纵或控制材料, 阅读时间由被试自主控制(平均时长 120 秒), 内容上, 四组材料长度匹配, 结构平行, 确保信息量一致; "椅子的历史"作为中性材料, 有效排除无关干扰。然后, 进行注意力检查, 通过一道选择题(上述材料描述的主要内容是: A 高情感人工智能, B 高情感人类, C 椅子的演进史。)排除未认真阅读者。接着, 测量体验性感知。使用 3 题项量表评估被试对主体的体验性感知(1 = 完全不同意, 7 = 完全同意); 随后, 阅读道德情景材料(性别歧视), 阅读后完成进行二次注意力检查(该材料中做出歧视决策的主体是: A 人类, B 人工智能)。在完成注意力检查后, 被试阅读道德决策情境材料, 并报告了对人工智能或人类的道德反应水平。道德反应测量同实验 1。最后, 匿名报告了性别、年龄、受教育程度三项人口统计学信息。所有被试都遵循完全相同的实验流程(知情同意→材料阅读→注意力检查→操纵检验→道德情境阅读→注意力检查→因变量测量→人口学信息收集)。这保证了实验条件的一致性。

3.2.2 结果

根据 Ge (2023)提出的用实验法进行中介效应检验程序, 对实验法进行中介需要满足的三个条件进行检验。

(1)感知体验性的调节效应分析。描述统计结果表明, 当决策主体是 AI 时, 对照组(n = 50)的道德反应为 4.11 (SD = 1.05), 感知体验性操纵组(n = 50)的道德反应为 4.89 (SD = 1.15); 当决策主体是人类时, 对照组(n = 50)的道德反应为 5.12 (SD = 0.74), 能动性操纵组(n = 50)的道德反应为 5.15 (SD = 0.91)。

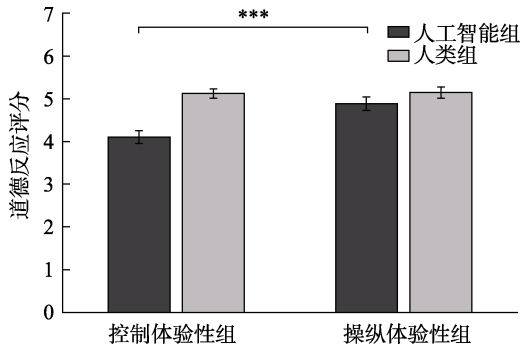

图 3 感知体验性在道德缺失效应中的调节作用

注: *** $p < 0.001$

调节效应分析结果显示, 决策主体的主效应显著, $F(1, 196) = 20.83$, $p < 0.001$, $\eta_p^2 = 0.096$, $1 - \beta = 0.995$; 感知体验性的主效应显著, $F(1, 196) = 8.47$, $p = 0.004$, $\eta_p^2 = 0.041$, $1-\beta = 0.825$; 决策主体与感知体验性的交互作用显著(见图 3), $F(1, 196) = 7.28$, $p = 0.008$, $\eta_p^2 = 0.036$, $1 - \beta = 0.766$。简单效应分析结果显示, 当人工智能作为决策主体时, 相比于对照条件, 在操纵感知体验性条件下的被试对人工智能的道德反应显著更高, $F(1, 196) = 15.73$, $p < 0.001$; 而当人类作为决策主体时, 对照组和操纵感知体验性组对人类的道德反应无显著差异, $F(1, 196) = 0.02$, $p = 0.881$。

(2)检验中介变量未被操纵的对照组(n = 100), 决策主体对感知体验性的预测作用。回归分析结果表明, 决策主体能显著正向预测感知体验性水平, $\beta = 0.78$, $p < 0.001$。结果表明, 相对于 AI 来说, 人们感知到人类的体验性水平更高。

(3)感知体验性中介操纵有效性检验。F 检验结果表明, 实验操纵组在感知体验性上的得分为 5.15 (SD = 1.52)显著高于对照组得分 4.70 (SD = 1.61), $F(1, 198) = 4.07$, $p = 0.045$, $\eta_p^2 = 0.020$, $1 - \beta = 0.519$。

综上可知, 决策主体与体验性感知之间存在显著交互作用; 在感知体验性的控制水平下(即中介变量未被操纵的组), 人工智能组的道德反应评分显著低于人类组; 中介变量体验性感知的操纵是有

效的, 满足使用实验中介法进行中介检验需要满足三个条件。因此, 体验性感知在人工智能决策降低道德反应之间起到中介作用。

3.2.3 讨论

本研究通过实验中介法验证了感知体验性在人工智能决策道德缺失效应中的中介作用。结果表明, 当人工智能作为决策主体时, 提升其体验性感知显著增强了被试的道德反应, 而人类作为决策主体时, 体验性操纵未引发显著差异。这一发现支持了心智感知二维理论的核心假设(Gray et al., 2007), 即人工智能的体验性感知具有可塑性, 其情感表征的强化可激活被试的共情反应, 进而缓解道德缺失效应(Bigman & Gray, 2018)。研究进一步验证了实验中介法的适用性, 决策主体与感知体验性的交互效应、控制组中人工智能道德反应显著低于人类, 以及操纵有效性检验共同构建了完整的因果链, 表明体验性感知是人工智能道德责任归因的关键心理机制。同样, 实验性中介法无法评估整个间接效应, 因此无法获得间接效应本身的点估计。并且, 无法考察基于心智感知理论(Gray et al., 2007)提出平行中介模型。针对这一局限, 我们在实验 2c 中使用问卷法分析了感知体验性中介变量的间接效应大小, 并且对感知体验性和感知能动性双通道平行中介模型进行了检验。

3.3 研究 2c 感知能动性和感知体验性平行中介作用

本研究通过问卷法, 基于温忠麟和叶宝娟(2014)提出的程序, 在真实情境中对能动性与体验性的平行中介作用进行检验。本研究假设(H4): 感知能动性和体验性在决策主体与道德反应水平之间起到平行中介作用。

3.3.1 方法

(1)被试

本研究采用结构方程模型进行中介效应分析, 根据 Jackson 等人(2003)建议, 样本量与测量题目数量比例最低为 10:1。本研究一共有 21 道题目, 至少应招募 210 名被试, 为确保最终有足够的数据用于分析, 通过脑岛平台采集数据, 控制措施同实验 1, 最终招募 376 名被试, 其中男性 154 名, 女性 222 名, 年龄范围在 18 至 59 岁间, 平均年龄 24.74 岁(SD = 5.44)。

(2)研究工具

心智感知。使用研究 2a 修订的心智感知总量表, 本研究中心智感知总量表的 Cronbach's α 值为 0.950; 感知能动性分量表的 Cronbach's α 值为 0.926; 感知体验性分量表的 Cronbach's α 值为 0.934。

道德反应。使用实验 1 修订的道德反应量表, 本研究中道德反应总量表的 Cronbach's α 值为 0.971。

道德决策情境材料, 同实验 2a。

3.3.2 结果

(1)共同方法偏差检验

本研究问卷数据均通过被试自我报告获取, 且心智感知量表与道德反应量表在测量内容上存在潜在关联性, 因此需要对变量的区分效度和共同方法偏差情况进行严格检验。

为检验感知能动性、感知体验性、道德认知、道德情感和道德行为这 5 个变量的区分效度, 本研究使用 AMOS 29.0 进行验证性因子分析, 将五因子模型与其他竞争性模型(单因子模型、二因子模型)进行比较。结果显示(参见表 2, 本研究所使用的五因子模型拟合效果更优, 对不同的构念有更好的区分效度, 可进行下一步分析。

本研究采用 AMOS 29.0 进行共同方法潜因子模型检验来检验共同方法偏差。相较于传统的 Harman 单因子检验, 共同方法潜因子模型检验能够更好地识别共同方法偏差(汤丹丹, 温忠麟, 2020)。结果显示, 将共同潜因子纳入模型后, 模型拟合指标分别为 χ^2/df = 4.61, RMSEA = 0.10, CFI = 0.93, TLI = 0.91, SRMR = 0.10。相比于控制前的模型, χ^2/df 提升 1.71, CFI 降低 0.03, TLI 降低 0.04, RMSEA 提升了 0.03, SRMR 提升 0.07, 模型的拟合度下降, 说明本研究不存在严重的共同方法偏差问题(见表 2)。

表 2　验证性因子分析结果

模型	χ^2	df	χ^2/df	CFI	TLI	RMSEA	SRMR
单因子模型	2749.91	189	14.55	0.70	0.67	0.19	0.14
二因子模型	1030.31	188	5.49	0.90	0.89	0.11	0.04
五因子模型	519.87	179	2.90	0.96	0.95	0.07	0.03

注: 单因子模型: 感知能动性+感知体验性+道德认知+道德情感+道德行为; 二因子模型: 感知能动性+感知体验性、道德认知+道德情感+道德行为; 五因子模型: 感知能动性、感知体验性、道德认知、道德情感和道德行为。

(2)描述性统计与相关分析

感知能动性与体验性高度相关(r = 0.83, p < 0.01), 二者均与道德反应三维度呈中度正相关(见表 3)。

表 3 各变量的描述统计分析

变量	*M*	*SD*	1	2	3	4
1 决策主体	0.51	0.50	1			
2 感知能动性	5.16	1.76	0.75***	1		
3 感知体验性	4.95	1.93	0.80***	0.83***	1	
4 道德反应	5.38	1.31	0.42***	0.48***	0.47***	1

注：*** p < 0.001

(3)平行中介分析

采用结构方程建模技术构建平行中介模型, 通过 AMOS 29.0 软件对感知能动性与感知体验性的中介路径进行联合估计。为增强参数估计的稳健性, 采用偏差校正 Bootstrap 法(重复抽样 5000 次)检验间接效应的显著性。模型识别检验显示, 该结构方程模型为恰好识别的饱和模型(df = 0)。根据 Hu 和 Bentler (1999)的拟合标准, 关键适配指标达到理想阈值：CFI = 1.00, TLI = 1.00, SRMR = 0.00。尽管饱和模型缺乏模型简约性评估, 但零自由度的数学特性确保了模型与样本协方差矩阵的精确适配, 满足心理测量学对模型可接受性的基本要求。

接着, 采用偏差校正 Bootstrap (5000 次抽样)检验中介效应, 结果显示, 两条中介路径的间接效应均显著：感知能动性的中介效应的标准化系数为 0.21, 95% CI = [0.032, 0.383]; 感知体验性中介效应的标准化系数为 0.18, 95% [0.003, 0.373]; 总间接效应的标准化系数为 0.39, 95% CI = [0.323, 0.467], 表明感知能动性与体验性共同解释了决策主体对道德反应的 39.8%变异(见图 4)。

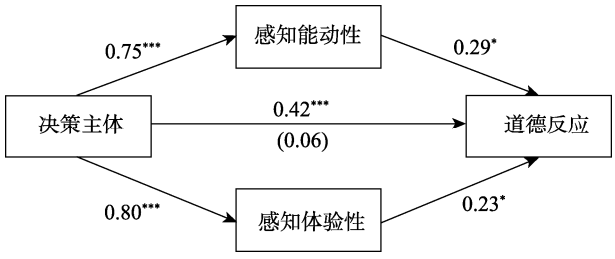

图 4 感知能动性与感知体验性的平行中介效应
注：* p < 0.05, *** p < 0.001

3.3.3 讨论

本研究通过结构方程模型验证了感知能动性与体验性在人工智能决定道德缺失效应中的平行中介作用。研究结果支持心智感知二维理论(Gray et al., 2007)的核心观点, 即能动性与体验性在道德判断中起到平行中介作用, 即人们对决策主体道德责任的归因需满足“有意为之”与“有情可感”的平行中介作用。研究进一步揭示了问卷法的生态效度优势：真实情境中, 能动性与体验性可能因社会认知的整体性偏向而产生共变, 导致单独路径被掩盖(Waytz et al., 2010)。本研究为技术伦理治理提供启示, 即人工智能的“道德可问责性”不仅需强化其能动性维度, 还需提升其体验性维度, 以此协同提升公众的道德反应水平。

4 实验 3: 人工智能决策道德缺失效应的干预研究

根据实验 2 可知, 人工智能正是由于能动性与体验性均弱于人类, 导致人们对于人工智能的道德判断程度比人类更弱。基于该机制, 可以通过提高能动性和体验性来缓解人工智能的道德缺失效应。然而, 现有文献尚未系统探讨如何通过心智感知干预提升道德敏感性。对此, 本研究拟通过随机对照实验, 提出双重干预策略, 并假设：(1)拟人化设计消除人工智能决策的道德缺失效应; (2)高心智预期消除人工智能决策的道德缺失效应; (3)二者协同作用比单一策略能更好地消除人工智能决策的道德缺失效应。

4.1 方法

4.1.1 被试

为保证实验设计的统计效能, 采用 G*Power3.1 进行先验样本量估算(单因素被试间 ANOVA, 中等效应量 f = 0.25, α = 0.05), 结果表明至少需 180 名受试以实现 80%检验力(Faul et al., 2007)。通过 Credamo 平台招募受试, 控制措施同实验 1, 为弥补可能的弃测或无效问卷, 最终获得 213 份有效数据。样本中女性 112 名(53%), 男性 101 名(47%); 被试年龄范围为 18 至 59 岁(M = 30.13, SD = 8.06)。所有被试在实验开始前均仔细阅读并签署了电子知情同意书。为检验人口学变量在各实验组间是否实现了均衡分配, 我们对性别和年龄进行了随机化检验, 详细的描述性统计与检验结果见表 1。针对性别分布的卡方独立性检验结果显示, 4 个实验组的性别构成存在显著差异, $\chi^2(3)$ = 11.25, p = 0.01, Cramer's V = 0.23。鉴于性别未能成功随机分配, 为排除其对实验结果的潜在混淆, 在后续的主要分析中, 性别将被作为协变量进行统计控制。针对年龄

的单因素方差分析结果则表明, 各实验组之间的平均年龄不存在显著差异, $F(3, 209) = 0.86$, $p = 0.465$。这说明年龄在各组间分布相对均衡, 不构成主要的混淆变量。

4.1.2 研究设计

本实验采用单因素 4 水平(对照组、拟人化干预组、期望调整干预组、拟人化+期望调整的综合干预组)的被试间设计; 因变量为道德反应。

4.1.3 研究材料与程序

研究材料包括 4 组, 分别对应 1 个对照组和 3 个干预组, 其编制逻辑清晰地指向两个核心自变量: 拟人化和期望。拟人化操纵材料参考 Waytz 等(2014)和 Laakasuo 等(2021)进行编制, 编制逻辑是赋予 AI 以人类的心理特征, 通过第一人称叙述、心理特征描述及类人化功能比喻, 增强人们对 AI 心智水平的感知。例如: “我是您的智能驾驶伙伴路遥, 作为一个自动驾驶助手, 我的使命就是为您提供安全、舒适和高效的出行体验。我的大脑中集成了先进的人工智能技术, 可以实时分析路况, 精准识别障碍, 甚至预测交通流量……”

期望操纵材料参考 Liu 等(2019)和 Hong 等(2021)进行编制, 逻辑是构建一个“完美 AI”的形象, 系统性地拔高被试的期望, 增强人们对 AI 心智水平的感知。例如: “人工智能能够做到客观公正, 主要得益于其决策过程的高度透明性和数据基础。AI 在做出判断时, 依赖于大量经过处理和分析的数据。此外, AI 系统的决策过程通常是可追溯的。每一步判断背后都有明确的算法和逻辑支持, 保证了决策的透明度……”

综合干预组同时引入上述拟人化与期望操纵, 检验其协同效应。

对照组则采用客观技术性描述, 将 AI 描述为纯粹的技术发展产物, 为其他组的干预效果提供一个零点参照。例如: “人工智能技术的发展始于 20 世纪中期的理论探索阶段。1956 年在达特茅斯会议上, 学者们首次提出了人工智能的概念框架, 其基础主要依托于符号逻辑系统, 为后续研究奠定了坚实理论基础。到了 1980 年代, 随着计算机技术的逐步成熟, 基于规则库的专家系统开始出现, 并初步构建了前馈神经网络的雏形……”

所有被试随机分配到 4 个不同实验条件后, 所有被试都遵循完全相同的实验流程(知情同意→干预操纵材料阅读→注意力检查→操纵检验→道德情境阅读→注意力检查→道德反应和心智感知测量→人口学信息收集)。其中, 道德决策情境材料同实验 2a。心智感知测量量表同 2c; 道德反应测量量表同实验 1; 本研究中道德反应量表的 Cronbach’s α 值为 0.947。实验结束后, 向受试者提供详细的实验目的和实际意义说明, 并给予实验报酬。

4.2 结果

4.2.1 操纵有效性检验

独立样本 t 检验结果显示, 操纵组的期望评分($M = 5.82$, $SD = 0.84$)显著高于对照组($M = 5.54$, $SD = 1.07$), $t(211) = -2.19$, $p = 0.030$, 95% CI [−0.548, −0.028], $Cohen's\ d = 0.30$, 表明期望操纵对评分的影响具有统计显著性。此外, 拟人化操纵组的评分($M = 5.37$, $SD = 1.05$)同样显著高于对照组($M = 4.81$, $SD = 1.34$), $t(211) = -3.41$, 95% CI [−0.890, −0.238], $p = 0.001$, Cohen's $d = 0.47$, 表明拟人化操纵对评分的影响具有显著性。综上所述, 结果表明, 无论是期望操纵还是拟人化操纵均能显著提高评分, 验证了两种操纵的有效性。

4.2.2 AI 拟人化+期望调整综合干预效果检验

为考察不同干预类型(对照组、拟人化干预组、期望调整干预组、拟人化+期望调整综合干预组)对道德反应、感知能动性与感知体验性水平的影响, 采用单因素协方差分析(ANCOVA), 以性别作为协变量进行控制。协变量选择的依据是独立样本 t 检验结果显示, 女性被试在道德反应上的评分($M = 4.95$, $SD = 1.27$)显著高于男性被试($M = 4.56$, $SD = 1.41$), $t(211) = -2.13$, $p = 0.034$, 95% CI [0.029, 0.751], Cohen’s $d = 0.29$。

在道德反应水平上, 描述性统计分析结果表明拟人化+期望调整综合干预组的得分为 5.64 ($SD = 0.91$), 拟人化干预组得分为 5.07 ($SD = 1.21$), 期望调整干预组得分为 4.85 ($SD = 1.10$), 对照组的得分为 3.55 ($SD = 1.22$)。单因素组间协方差分析结果显示, 干预类型主效应显著, $F(3, 208) = 31.18$, $p < 0.001$, $\eta_p^2 = 0.31$。Bonferroni 校正的事后成对比较进一步显示, 综合干预组显著高于对照组($M_d = 2.07$, $SE = 0.22$, $p < 0.001$), 期望调整组($M_d = 0.80$, $SE = 0.22$, $p = 0.02$), 边缘显著高于拟人化干预组($M_d = 0.57$, $SE = 0.22$, $p = 0.061$); 拟人化干预组显著高于对照组($M_d = 1.50$, $SE = 0.22$, $p < 0.001$); 期望调整组显著高于对照组($M_d = 1.27$, $se = 0.22$, $p < 0.001$); 拟人化干预组与期望调整干预组之间差异不显著($M_d = 0.23$, $SE = 0.22$, $p = 1.00$), 见图 5。

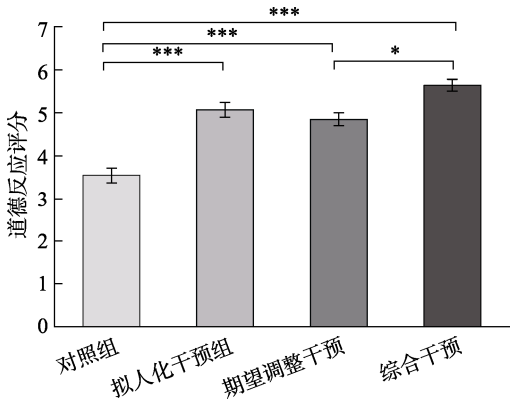

图 5 不同干预策略对道德反应的提升作用
注：* $p < 0.05$, *** $p < 0.001$

在感知能动性水平上，描述性统计分析结果表明拟人化+期望调整综合干预组的得分为 5.20 (*SD* = 0.86)，拟人化干预组得分为 4.52 (*SD* = 0.99)，期望调整干预组得分为 4.77 (*SD* = 1.14)，对照组的得分为 2.49 (*SD* = 1.09)。单因素组间协方差分析结果显示，干预类型对感知能动性评分的影响显著，$F(3, 208) = 71.97$, $p < 0.001$, $\eta_p^2 = 0.51$。Bonferroni 校正的事后成对比较进一步显示，综合干预组显著高于对照组($M_d = 2.75$, $SE = 0.20$, $p < 0.001$)和拟人化干预组($M_d = 0.67$, $SE = 0.20$, $p = 0.005$)，但与期望调整组无显著差异($M_d = 0.43$, $SE = 0.20$, $p = 0.205$)；拟人化干预组显著高于对照组($M_d = 2.07$, $SE = 0.20$, $p < 0.001$)；期望调整组显著高于对照组($M_d = 2.32$, $SE = 0.20$, $p < 0.001$)；拟人化干预组与期望调整干预组之间差异不显著($M_d = -0.26$, $SE = 0.20$, $p = 1.00$)，见图 6。

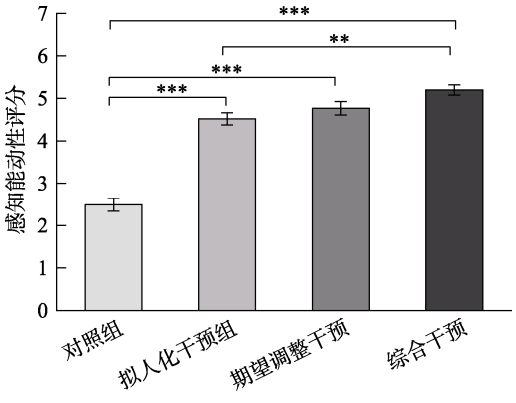

图 6 不同干预策略对感知能动性的提升作用
注：** $p < 0.01$, *** $p < 0.001$

在感知体验性水平上，描述性统计分析结果表明拟人化+期望调整综合干预组得分为 4.66 (*SD* = 1.35)，拟人化干预组得分为 3.89 (*SD* = 1.33)，期望调整干预组得分为 3.80 (*SD* = 1.35)，对照组的得分为 2.22 (*SD* = 0.85)。单因素组间协方差分析结果显示，干预类型对感知体验性评分的影响显著，$F(3, 208) = 37.15$, $p < 0.001$, $\eta_p^2 = 0.35$。Bonferroni 校正的事后成对比较进一步显示，综合干预组显著高于对照组($M_d = 2.50$, $SE = 0.24$, $p < 0.001$)，期望调整组($M_d = 0.78$, $SE = 0.24$, $p = 0.009$)，拟人化干预组($M_d = 0.85$, $SE = 0.24$, $p = 0.003$)；拟人化干预组显著高于对照组($M_d = 1.72$, $SE = 0.24$, $p < 0.001$)；期望调整组显著高于对照组($M_d = 1.65$, $SE = 0.24$, $p < 0.001$)；拟人化干预组与期望调整干预组之间差异不显著($M_d = 0.07$, $SE = 0.24$, $p = 1.00$)，见图 7。

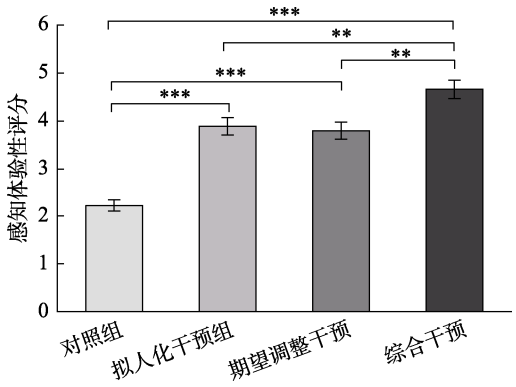

图 7 不同干预策略对感知体验性的提升作用
注：** $p < 0.01$, *** $p < 0.001$

4.2.3 AI 拟人化+期望调整综合干预的路径检验

为考察不同干预类型(对照组、拟人化干预组、期望调整干预组、拟人化+期望调整综合干预组)对道德反应的影响路径，本研究以干预组别作为自变量，感知能动性和感知体验性作为平行中介变量，道德反应作为因变量，采用 PROCESS 宏(Model 4; Hayes, 2013)进行平行中介效应分析。由于自变量为多分类变量，我们首先将其转换为 3 个虚拟变量，以对照组为基准：*D*1 (拟人化干预组 vs. 对照组)、*D*2 (期望调整干预组 vs. 对照组)、*D*3 (综合干预组 vs. 对照组)。分析基于 Bootstrap 重抽样方法(重复抽样 5000 次)，以获得稳健的置信区间估计。同时，由于性别对道德反应水平具有显著影响，本研究将性别作为协变量进行控制。

综合干预路径模型分析结果。相较于对照组，综合干预显著提升了个体对 AI 的感知能动性($b = 0.86$, $SE = 0.22$, $p < 0.001$, 95% CI [0.84, 1.69])，感知能动性进一步显著正向预测道德反应($b = 0.18$, $SE = 0.07$, $p = 0.011$, 95% CI [0.04, 0.30])；感知能

动性的中介效应显著(间接效应 = 0.21, Boot *SE* = 0.09, 95% CI [0.04, 0.40])。同时, 综合干预显著提升了个体对 AI 的感知体验性(b = 0.89, *SE* = 0.22, $p < 0.001$, 95% CI [0.82, 1.80]), 感知体验性进一步显著正向预测道德反应(b = 0.32, *SE* = 0.06, $p < 0.001$, 95% CI [0.16, 0.41]); 感知体验性的中介效应显著(间接效应 = 0.39, *SE* = 0.10, 95% Boot CI [0.21, 0.60])。路径模型如图 8 所示。

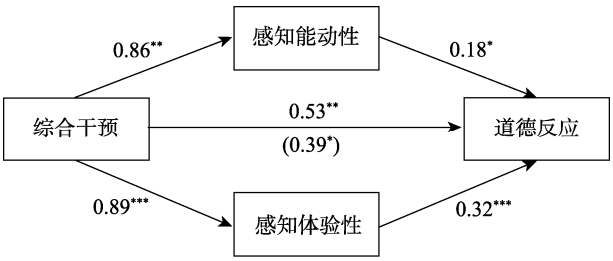

图 8　拟人化与期望调整综合干预提升人们对不道德 AI 决策的道德反应水平的路径

注：* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.3　讨论

研究结果显示, 拟人化与期望调整的综合干预组在道德反应水平上得分最高, 显著高于对照组, 表明通过增强人们对 AI 的拟人化感知和期望调整, 可以有效提升其道德反应水平。同时, 综合干预还显著提升了感知能动性和体验性, 这两者作为中介变量均对道德反应产生了显著影响。值得注意的是, 感知体验性的中介效应值为 0.29, 明显高于感知能动性的 0.17, 这表明人们对 AI 的体验性感知在提升道德反应方面发挥了更为关键的作用。在本研究中, 综合干预对感知体验性的影响系数达到 0.89, 这一现象可能源于体验性感知在人与 AI 互动中的核心地位。当人们将 AI 视为具有情感和感知能力的实体时, 他们更倾向于赋予其道德属性, 从而激发更强烈的道德反应(Gray et al., 2012)。因此, 通过拟人化与期望调整的干预, 不仅可以提升人们对 AI 的整体道德反应, 还能通过增强体验性感知进一步促进这一效应。综上所述, 本研究表明, 通过拟人化和期望调整的综合干预措施, 可以显著提升人们对 AI 的道德反应, 而其中感知体验性的中介作用尤为显著。这一发现为 AI 伦理研究提供了新的视角, 并为设计更有效的干预方案以增强公众对 AI 道德认知和反应提供了理论支持。未来的研究可以进一步探讨不同文化背景下这些干预措施的有效性, 并探索如何将这些理论成果应用于实际的 AI 系统设计中。

5　总讨论

本研究在整合心智感知理论与道德二元论的基础上, 系统揭示了人工智能道德缺失效应的双路径机制及其干预策略。研究结果表明, 相较于人类决策者, 人们对人工智能不道德决策的道德反应水平显著降低, 而对人工智能较低的能动性与体验性的感知是这一效应的重要心理原因。进一步的干预研究显示, 结合针对人工智能的拟人化策略与针对人类的期望调整策略所构成的综合干预方案, 能够有效提升个体对人工智能不道德行为的道德反应。值得注意的是, 不同于计算机科学、哲学、法学、社会学等其他学科“算法伦理”研究主要聚焦于从设计层面提出公平算法的原则与技术路径, 本研究从心理学视角切入, 强调个体在面对人工智能与人类决策时的心理反应差异。这一发现不仅为缓解算法偏见引发的社会问题和推动公平算法的构建提供了新的理论思路, 也为“算法伦理”研究开辟了新的研究视角。

5.1　人们对人工智能不道德决策的道德反应显著弱于人类决策者

本研究基于中国社会文化背景, 通过学历歧视(筛选淘汰非“985”院校申请者)、年龄歧视(35 岁职业门槛算法)、性别歧视(简历筛选的性别权重偏差)三类典型场景, 揭示了人们对人工智能不道德决策的道德反应显著弱于人类决策者。与先前主要聚焦于西方个体主义文化的研究(Bigman et al., 2023)不同, 本研究的核心贡献在于证实了该效应在中国集体主义文化背景下的稳健性。现有基于西方个体主义文化构建的实验范式, 往往难以充分捕捉集体主义文化中个体的道德认知特征。例如, 根植于西方价值体系的实验材料(如涉及种族歧视的议题), 在反映中国社会特有的伦理困境时存在明显局限性。因此, 开发具有文化适配性和本土现实针对性的道德情境材料, 以提升实验的生态效度, 是探索该 AI 决策道德缺失效应跨文化普适性的关键前提。本研究不仅开发了具有文化适配性和本土现实针对性的道德情境材料, 而且得出了一些有价值的发现。具体来说, 相比于学历歧视和年龄歧视等不道德场景, 性别歧视有其突出特点(Wilson et al., 2022)。性别歧视是人类社会长期存在的系统性偏见, 且在全球范围内具有高度可见性, 如职场性别薪酬差距、女性领导力缺失(Dastin, 2022; Heilman et al., 2024;

Xiao et al., 2024)。研究人工智能如何继承或放大这种根深蒂固的偏见，能直接揭示技术对社会结构的“复制−强化”机制。此外，招聘场景中，性别歧视常通过历史数据(如男性主导的科技行业招聘记录)被编码到算法中，导致人工智能系统在评估中对女性候选人赋予较低权重(Chang, 2023)。这种“数据−算法−结果”的链条清晰可溯，便于剖析技术伦理问题的底层逻辑。因此，在实验 2 主要以性别歧视作为探讨人工智能道德缺失的道德场景，并且通过系列子实验(一个相关研究与两个实验研究)证实了，相对于人类，当决策主体是 AI 时候，人们对性别歧视决定的道德反应水平更低。该系列研究不仅证实了 AI 决策道德缺失效应的稳健性，而且采用场景剥离法证实了性别歧视作为理想实验场景的理论优势，相较于学历或年龄歧视，性别歧视更易剥离其他变量干扰(如学历与能力的关系可能更复杂)，为机制研究和干预实验提供纯净变量条件。

5.2 感知能动性与体验性的平行中介作用

本研究通过整合心智感知理论与道德二元论，系统揭示了人工智能道德缺失效应的双路径机制。以往研究主要从个别心理路径解释该现象，例如自由意志信念(Bigman et al., 2023)或偏见动机(许丽颖 等, 2022)，呈现出机制解释的碎片化特征。本研究从道德主体出发，阐释能动性和体验性在人工智能决策的道德缺失效应中的中介作用，对心智感知理论与道德二元论有重大理论推进。不同于既往研究对自由意志、自主性等单一变量的零散探讨(许丽颖 等, 2022; Bigman et al., 2023; Heinrichs et al., 2022)，本研究首次证实感知能动性与体验性在 AI 决策的道德缺失效应中具有平行中介作用，这一发现突破传统“能动性(主体)−体验性(对象)”的二元对立框架(Gray et al., 2012)。现有研究探讨了感知能动性对于人工智能道德缺失效应的影响(Hohenstein & Jung, 2020; Wilson et al., 2022; Zhu & Chu, 2025)，但本研究结果进一步指出，单纯强调能动性提升(如增强决策透明度)并不足够，体验性维度(如情感交互设计)的协同优化同样关键。实验数据显示，当人工智能的体验性感知被强化时，其道德主体评价显著提升，这有力支持了新兴的“情感理性主义”观点，即体验性不仅是道德对象的必要条件，更是道德主体资格的内在构成要素(de Vel-Palumbo et al., 2022)。这一发现突破了传统道德二元论对体验性维度的片面定位，为理解道德主体的多元构成提供了新视角。

本研究首次通过实验中介法提供感知能动性与体验性在道德缺失效应中平行中介的因果证据，弥补前人研究(许丽颖 等, 2022; Sullivan & Fosso Wamba, 2022)问卷法仅能验证相关性的不足；并通过结构方程模型证明，感知能动性与体验性的联合解释力超越单一路径(如研究 2c)，深化了对道德缺失效应复杂性的理解。具体来说，本研究通过实验中介法与问卷法的多方法验证，首次实现人工智能决策道德缺失效应机制的因果推断。2×2 实验设计揭示出决策主体与心智维度的交互作用，即在人工智能条件下，能动性操纵显著提升道德反应，而人类主体未呈现此效应。这种“能动性归因偏差”印证了心智感知理论的核心假设，即人类固有地被赋予完整的心智图式(Leyens, 2000)；而人工智能的“类主体”地位具有建构性，可通过心智维度的表征设计进行调节(Bigman et al., 2023)。通过操纵感知能动性/体验性(高 vs 控制)以及决策主体(AI vs 人类)发现，在被动性控制组中，决策主体显著预测能动性/体验性水平，而 AI 组能动性提升显著改善道德反应，构建“主体类型→心智感知→道德反应”的完整因果链。这一设计克服传统问卷法在因果推断上的先天性缺陷，突破既往研究限于相关分析的桎梏。接着采用问卷法同时考察能动性与体验性的平行中介作用，揭示能动性与体验性的协同增益效应，在性别歧视情境中，双路径的标准化系数为 0.398, 95% CI = [0.323, 0.467]，联合解释方差达 39.8%，较单一感知能动性的中介效应的标准化系数为 0.214, 95% CI = [0.032, 0.383]；感知体验性中介效应的标准化系数为 0.184, 95% CI [0.003, 0.373]解释了更多变异，提示真实情境中道德判断可能遵循“有意为之”与“有情可感”的双重机制。这一推进意义是揭示了 AI 决策道德缺失的深层机制——人们认为算法既缺乏自主意图(低能动性)，又无法理解情感伤害(低体验性)，导致道德反应不足，解决了当前不同研究强调不同中介的碎片化问题，形成统一解释框架。

5.3 拟人化与期望调整的综合干预作用

本研究创新性地提出并验证了通过提升对 AI 的能动性和体验性感知，减少道德缺失效应的“拟人化+期望调整”综合干预策略。具体来说，本研究拟通过随机对照实验，提出双重干预策略，探讨通过拟人化设计和期望调整两种干预策略缓解 AI 决策道德缺失效应的可能途径。研究结果表明，无论是单一的拟人化干预还是期望调整，其对增强个体

的道德反应、提升对 AI 能动性和体验性感知均具有显著正向作用; 而两种干预策略的协同效应更为明显。这一发现为构建系统性的干预方案提供了实证依据。研究中引入的拟人化设计, 通过赋予 AI 以人类的形象、声音或行为特征, 使人们更容易感知其具备较高的心智感知水平, 从而将其视为具备道德责任的“完整主体” (Gursoy et al., 2019; Lin et al., 2020; Melián-González et al., 2021)。从两因素方差分析结果来看, 拟人化干预在提升道德反应、感知能动性及体验性上均达到显著效应, 并且在期望操纵下这种效果进一步放大。这说明, 拟人化设计能够使人们在面对不道德的 AI 决策时, 有更高的道德反应水平。此外, 中介路径分析进一步揭示, 拟人化对道德反应的提升是通过增强个体对 AI 能动性和体验性的感知实现的。除了拟人化干预, 期望调整作为另一种干预手段, 同样在消除 AI 决策道德缺陷中展现出了独特的作用(Srinivasan & Sarial-Abi, 2021)。研究表明, 提前设定人们对 AI 行为的合适期望, 不仅显著提升了道德反应水平, 而且通过感知能动性和体验性的中介效应再次验证了这一干预机制的合理性。进一步的调节中介效应检验表明, 拟人化与期望调整的协同作用机制主要依赖于对 AI 能动性和体验性感知的双重提升。已有研究(胡小勇 等, 2024; 许丽颖 等, 2022; Bigman et al., 2023)从不同视角出发考察了拟人化或期望调整对 AI 决策道德缺失效应的干预作用, 但是当前鲜有研究同时对人工智能与人类干预角度, 考察拟人化与期望调整的综合干预效果。例如, Bigman 等(2023)与许丽颖等(2022)只关注到拟人化对 AI 决策道德缺失效应的干预作用; 胡小勇等(2024)通过文献综述关注到了拟人化与期望调整的作用, 但没有提出关于二者综合干预方案。本研究首次提出并考察了从 AI 设计(拟人化)和人类认知(期望调整)双路径干预, 实验证明协同效应显著强于单一策略。通过干预路径理论模型创新, 推动 Bigman 等(2023)与许丽颖等(2022)从单一自由意志或偏见动机的解释, 发展为“能动性−体验性”双路径模型, 并催生了多维度干预策略, 更有效地缓解了 AI 决策的道德缺失现象, 为 AI 伦理治理提供了心理学视角的创新路径。

5.4　局限与展望

尽管取得上述研究成果, 本研究仍存在一些不足。首先, 本研究在实验设计上存在一定的场景局限性。具体而言, 实验 1 和实验 2 主要考察了歧视类道德情境(如性别歧视、学历歧视和年龄歧视), 而实验 3 则聚焦于自动驾驶的道德困境。虽然这些场景具有显著的社会现实意义, 但仅能反映特定类型的道德问题。值得注意的是, 人工智能在实际应用中涉及的伦理挑战更为多元, 例如隐私保护、医疗决策中的生命权权衡、公共资源分配等复杂情境。在这些不同的道德领域中, 公众对 AI 决策的反应模式可能存在显著差异。例如, 涉及人身安全的道德困境可能引发更强烈的情绪反应, 而隐私侵犯问题可能更侧重于责任归因。未来研究应当拓展道德情境的多样性, 将环境保护、教育公平、司法量刑等更多领域纳入考察范围。这种拓展不仅能够更全面地界定人工智能道德缺失效应的边界条件, 还能为构建更具包容性的 AI 伦理治理框架提供更丰富的实证基础。

其次, 实验方法的生态效度存在一定问题。本研究主要采用情境模拟的实验范式, 通过文本描述和视频材料来呈现道德决策情境。虽然这种方法在变量控制和实验操纵方面具有明显优势, 但其生态效度存在一定局限。实验室环境下被试的反应可能无法完全反映真实场景中的复杂心理过程。现实中的道德反应往往发生在动态变化的环境中, 并受到社会规范、文化背景和个人经历等多重因素的影响。例如, 在真实的招聘场景中, 应聘者与 AI 系统的互动可能持续数周, 其间产生的道德认知会随时间推移而演变, 这与实验室中的即时反应存在本质差异。为提高研究的外部效度, 未来工作可采用以下改进方案: 开展现场实验, 将研究嵌入真实的 AI 应用场景, 如医疗机构中的诊断系统或企业的自动化招聘平台; 采用纵向追踪设计, 考察长期接触 AI 系统对个体道德敏感度的累积影响; 结合多模态数据采集技术(如眼动追踪、生理指标监测等), 更全面地捕捉被试在真实情境中的反应模式。这些方法创新将显著提升研究结果的现实解释力, 为 AI 伦理治理提供更具操作性的科学依据。

第三, 本研究将拟人化与期望调整视为两种相互独立的干预策略, 分别从人工智能特征与人类观念层面提出了潜在的缓解方案。然而, 从理论上看, 拟人化作为强化 AI 类人特征的一种手段, 可能直接影响人们对 AI 心智能力的预期。感知到的类人特征能够促进人机互动, 促使人们将社会启发式判断迁移至与机器人的交互情境, 从而形成更高的期望(Duffy, 2003; Nass & Moon, 2000)。然而, 尽管拟人化设计提升了表层相似性, AI 的实际行为往往难

以完全契合用户预期, 这种期望落差可能引发显著的负面情绪(Grazzini et al., 2023)。已有研究表明, 拟人化可增强对技术的信任, 但当行为表现未达预期时, 其信任反而可能转化为更强烈的失望与抵触(Crolic et al., 2022; Waytz et al., 2014)。这些证据提示, 拟人化与期望调整在心理机制上存在交集, 二者并非完全独立。因此, 两种干预策略在结合应用时, 其作用路径可能比本研究假设的更加复杂。未来研究应在不同个体和情境下考察这些策略在长期互动中的影响, 从而更全面地阐释拟人化和期望调整如何在各种条件下相互作用, 并为人工智能道德决策及人机互动优化提供更具针对性的理论基础和实践指导。

6 结论

本研究结论如下：第一, 人们对人工智能不道德决策的道德反应显著弱于人类决策者; 第二, 相对于人类主体来说, 在人工智能决策条件下, 能动性与体验性操纵显著提升道德反应; 并且, 感知能动性与体验性在人工智能决策道德缺失效应中具有平行中介作用; 第三, 无论是单一的拟人化干预还是期望调整, 其对增强个体的道德反应、提升对 AI 能动性和体验性感知均具有显著正向作用; 而两种干预策略的协同效应更大。

参 考 文 献

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour, 4*(2), 134−143.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics, 143*(1), 30−56.

Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, *30*(2), 195−218.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21−34.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365−368.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General, 152*(1), 4−27.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied, 27*(2), 447−459.

Burgoon, J. K., Newton, D. A., Walther, J. B., & Baesler, E. J. (1989). Nonverbal expectancy violations and conversational involvement. *Journal of Nonverbal Behavior, 13*(2), 97−119.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30−37.

Chang, X. (2023). Gender bias in hiring: An analysis of the impact of Amazon's recruiting algorithm. *Advances in Economics, Management and Political Sciences*, *23*(1), 134−140.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer−chatbot interactions. *Journal of Marketing*, *86*(1), 132−148.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4−19.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299−309.

Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In K. Martin (Ed), *Ethics of data and analytics* (pp. 296−299). Auerbach Publications.

de Vel-Palumbo, M., Ferguson, R., Schein, C., Chang, M. X.-L., & Bastian, B. (2022). Morally excused but socially excluded: Denying agency through the defense of mental impairment. *PLoS ONE, 17*(7), e0272061.

Decety, J., & Cowell, J. M. (2018). Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. *Development and Psychopathology, 30*(1), 153−164.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos One*, *18*(3), e0279720.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177−190.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175−191.

Ge, X. (2023). Experimentally manipulating mediating processes: Why and how to examine mediation using statistical moderation analyses. *Journal of Experimental Social Psychology, 109*, 104507.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619−619.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*(3), 206−215.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125−130.

Grazzini, L., Viglia, G., & Nunan, D. (2023). Dashed expectations in service experiences. Effects of robots human-likeness on customers' responses. *European Journal of Marketing*, *57*(4), 957−986.

Griffith, R. L., & Peterson, M. H. (Eds.). (2006). *A closer examination of applicant faking behavior*. IAP.

Grimes, G. M., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems, 144*, 113515.

Guidi, S., Marchigiani, E., Roncato, S., & Parlangeli, O.

(2021). Human beings and robots: Are there any differences in the attribution of punishments for the same crimes? *Behaviour & Information Technology*, *40*(5), 445−453.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management, 49*, 157−169.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 998.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.

Heilman, M. E., Caleo, S., & Manzi, F. (2024). Women at work: Pathways from gender stereotypes to gender bias and discrimination. *Annual Review of Organizational Psychology and Organizational Behavior, 11*(1), 165−192.

Heinrichs, B., Heinrichs, J.-H., & Rüther, M. (2022). *Künstliche Intelligenz*. De Gruyter.

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior, 106*, 106190.

Hong, J. W., Cruz, I., & Williams, D. (2021). AI, you can drive my car: How we evaluate human drivers vs. self-driving cars. *Computers in Human Behavior, 125,* 106944.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Hu, X., Li, M., Wang, D., & Yu, F. (2024). Reactions to immoral AI decisions: The moral deficit effect and its underlying mechanism. *Chinese Science Bulletin, 69*(11), 1406−1416.

[胡小勇, 李穆峰, 王笛新, 喻丰. (2024). 人工智能决策的道德缺失效应及其机制. *科学通报, 69*(11), 1406−1416.]

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N: Q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(1), 128−141.

Kamide, H., Eyssel, F., & Arai, T. (2013). Psychological anthropomorphism of robots: Measuring mind perception and humanity in Japanese context. In G. Herrmann, M. Pearson, A. Lenz, P. Bremner, A. Spiers, & U. Leonards (Eds.), *Social robotics* (pp. 199−208). Springer.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155−163.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics, 13*(7), 1679−1688.

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior, 123*, 106878.

Leo, X., & Huh, Y. E. (2020). Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. *Computers in Human Behavior*, *113*, 106520.

Lew, Z., & Walther, J. B. (2023). Social scripts and expectancy violations: Evaluating communication with human or AI chatbot interactants. *Media Psychology*, *26*(1), 1−16.

Leyens, S. (2000). La conscience imaginée: Sur l'éliminativisme de Daniel Dennett. *Revue philosophique de Louvain, 98*(4), 761−782.

Lima, G., Kim, C., Ryu, S., Jeon, C., & Cha, M. (2020). Collecting the Public Perception of AI and Robot Rights. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–24.

Lin, H., Chi, O. H., & Gursoy, D. (2020). Antecedents of customers' acceptance of artificially intelligent robotic device use in hospitality services. *Journal of Hospitality Marketing & Management, 29*(5), 530−549.

Lin, M., Cui, X., Wang, J., Wu, G., & Lin, J. (2022). Promotors or inhibitors? Role of task type on the effect of humanoid service robots on consumers' use intention. *Journal of Hospitality Marketing & Management, 31*(6), 710−729.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Liu, P., Yang, R., & Xu, Z. (2019). How safe is safe enough for self-driving vehicles? *Risk Analysis, 39*(2), 315−325.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382−386.

Magrani, E. (2019). New perspectives on ethics and the laws of artificial intelligence. *Internet Policy Review, 8*(3), 1−19. https://doi.org/10.14763/2019.3.1420

Malle, B. F. (2019). How many dimensions of mind perception really are there? In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268−2274). Cognitive Science Society.

Mancosu, M., Ladini, R., & Vezzoni, C. (2019). 'Short is better'. Evaluating the attentiveness of online respondents through screener questions in a real survey environment. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *141*(1), 30−45. https://doi.org/10.1177/0759106318812788

Maninger, T., & Shank, D. B. (2022). Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports*, *5*, 100154.

Melián-González, S., Gutierrez-Tano, D., & Bulchand-Gidumal, J. (2021). Predicting the intentions to use chatbots for travel and tourism. *Current Issues in Tourism, 24*(2), 192−210.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81−103.

Nijssen, S. R. R., Müller, B. C. N., Bosse, T., & Paulus, M. (2023). Can you count on a calculator? The role of agency and affect in judgments of robots as moral agents. *Human-Computer Interaction, 38*(5-6), 400−416.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447−453.

Pavone, G., Meyer-Waarden, L., & Munzel, A. (2023). Rage against the machine: Experimental insights into customers' negative emotional responses, attributions of responsibility, and coping strategies in artificial intelligence-based service failures. *Journal of Interactive Marketing*, *58*(1), 52−71.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489−497.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459−467.

Qian, Y., & Wan, X. (2024). Influence of robot anthropomorphism on consumer attitudes toward restaurants and service

providers. *International Journal of Hospitality Management, 123*, 103939.
Rai, A., Constantinides, P., & Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *Management Information Systems Quarterly, 43*(s1), iii-ix.
Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, *78*, 102253.
Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648−663.
Shank, D. B., North, M., Arnold, C., & Gamez, P. (2021). Can mind perception explain virtuous character judgments of artificial intelligence? *Technology, Mind, and Behavior, 2*(3), 1−13. https://doi.org/10.1037/tmb0000047
Song, F., & Yeung, S. H. F. (2024). A pluralist hybrid model for moral AIs. *AI & Society, 39*(3), 891−900.
Srinivasan, R., & Sarial-Abi, G. (2021). When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing, 85*(5), 74−91.
Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics, 178*(4), 917−943.
Tang, D. D., & Wen, Z. L. (2020). Statistical approaches for testing common method bias: Problems and suggestions. *Journal of Psychological Science, 43*(1), 215−223.
[汤丹丹, 温忠麟. (2020). 共同方法偏差检验：问题与建议. *心理科学, 43*(1), 215−223.]
Wang, X., Wu, Y. C., Ji, X., & Fu, H. (2024). Algorithmic discrimination: Examining its types and regulatory measures with emphasis on US legal practices. *Frontiers in Artificial Intelligence, 7*, 1320277.
Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science, 19*(1), 58−62.
Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113−117.
Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374−11379.
Wen, Z. L., & Ye, B. J. (2014). Analyses of mediating effects: The development of methods and models. *Advances in Psychological Science, 22*(5), 731−745.
[温忠麟, 叶宝娟. (2014). 中介效应分析: 方法和模型发展. *心理科学进展, 22*(5), 731−745.]
Wilson, A., Stefanik, C., & Shank, D. B. (2022). How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? *Computers in Human Behavior Reports, 8*, 100229.
Xiao, H., Marie, A., & Strickland, B. (2024). Moral commitment to gender equality increases (mis)perceptions of gender bias in hiring. *European Journal of Social Psychology, 54*(6), 1211−1227.
Xu, L., Yu, F., & Peng, K. (2022). Algorithmic discrimination causes less desire for moral punishment than human discrimination. *Acta Psychologica Sinica, 54*(9), 1076−1092.
[许丽颖, 喻丰, 彭凯平. (2022). 算法歧视比人类歧视引起更少道德惩罚欲. *心理学报, 54*(9), 1076−1092.]
Zhang, S., Lin, X., Li, X., & Ren, A. (2022). Service robots' anthropomorphism: Dimensions, factors and internal relationships. *Electronic Markets, 32*(1), 277−295.
Zhu, Y., & Chu, J. (2025). Should we express gratitude in human-AI interaction: The online public's moral stance toward artificial intelligence assistants in China. *Public Understanding of Science, 34*(6), 717−733.

Moral deficiency in AI decision-making: Underlying mechanisms and mitigation strategies

HU Xiaoyong[1], LI Mufeng[2], LI Yue[1], LI Kai[1], YU Feng[1]

([1] *Department of Psychology, Wuhan University, Wuhan* 430072, *China*)

([2] *Faculty of Psychology, Southwest University, Chongqing* 400715, *China*)

Abstract

As artificial intelligence (AI) assumes an increasingly prominent role in high-stakes decision-making, the ethical challenges it raises have become a pressing concern. This paper systematically investigates the moral deficiency effect in AI decision making by integrating mind perception theory with moral dualism. Through this framework, we identify a dual-path psychological mechanism and propose targeted intervention strategies.

Our first investigation, Study 1, explored the limitations of AI in moral judgment using scenarios rooted in the Chinese socio-cultural context. Across three representative situations—educational, age, and gender discrimination—the moral response scores for AI-generated decisions were significantly lower than for those made by human agents. These findings not only align with existing Western research on AI's moral judgment deficits but also suggest that the moral deficiency effect is generalizable across cultures.

To understand why this deficiency occurs, Study 2 investigated the underlying psychological mechanisms. Drawing on mind perception theory and moral dualism, we proposed a dual-path mediation model involving perceived agency and perceived experience. We conducted three sub-studies that first tested these two mediators separately and then assessed their combined effects. Using experimental mediation, we provided the first causal evidence of how the decision-maker's identity (AI vs. human) interacts with dimensions of mind perception. Specifically, when participants perceived an AI as having greater agency and experience, their moral approval of its decisions significantly increased—an effect not observed with human decision-makers. Structural equation modeling further confirmed a synergistic effect between the two paths, indicating their combined explanatory power exceeds that of either one alone. This suggests that in the real world, moral responses to AI are influenced simultaneously by both cognitive pathways.

Building on these mechanistic insights, Study 3 tested intervention strategies to mitigate the AI-induced moral deficiency effect. In a double-blind, randomized controlled experiment, we evaluated two approaches: anthropomorphic design and mental expectancy enhancement. Both strategies significantly improved moral responses by increasing participants' perceptions of the AI's agency and experience. Moreover, a combined intervention produced a stronger effect than either strategy did alone. Although these interventions target different elements—one focusing on the AI system and the other on human cognition—they both operate through the shared mechanism of mind perception. By doing so, they effectively enhance moral accountability for an AI's unethical behavior, offering a practical pathway to address moral deficiencies in AI decision-making.

Ultimately, this research provides a novel contribution to the field of “algorithmic ethics.” Unlike traditional approaches that emphasize technical design principles and fairness algorithms, our study adopts a psychological perspective that centers on the human recipient of AI-driven decisions. Practically, we propose actionable intervention strategies grounded in mind perception, while our synergistic model provides a robust framework for AI ethical governance. Collectively, these findings deepen the understanding of moral judgment in AI contexts, guide the development of algorithmic accountability systems, and support the optimization of human−AI collaboration—thereby establishing a critical psychological foundation for the ethical deployment of AI.

Keywords　artificial intelligence, moral deficit effect, mind perception, anthropomorphism, expectation adjustment