心理学探新 2025,Vol.45,No.4,300-308
PSYCHOLOGICAL EXPLORATION

人工智能的道德主体性及其责任归属*

王小桃[1,2],张 璟[1],张凤华[1],董圣鸿[1]
(1.江西师范大学心理学院,南昌 330022;2.江西师范大学心理健康教育研究中心,南昌 330022)

摘 要:随着人工智能(artificial intelligence,AI),尤其是生成式AI技术的迅猛发展,其在道德决策与人机互动中的作用日益凸显,引发了心理学领域对AI是否具备道德主体性(moral agency)的广泛关注。本文首先梳理了AI作为道德主体时的道德表现,涵盖如算法偏见之类的隐性道德行为,以及在道德困境中所展现的显性道德行为。其次,文章系统回顾了当前关于人类对AI进行道德归责的研究进展,主要聚集于三个方面:AI是否应该被视为可归责主体;人们对AI与人类行为者进行道德归责时存在非对称性;以及影响人们对AI进行道德归责的心理与社会因素。最后,文章提出未来可从三个方向展开更为深入的探讨:一是深入解析生成式AI在道德判断中的决策机制;二是系统探索人类对AI道德反应偏差的认知根源;三是探讨AI引发的责任缺口现象背后的心理与社会机制。综上,对AI的道德行为研究拓展了心理学对道德认知和归责机制的研究视野,也对未来AI系统的公平性和伦理治理提出了新的挑战和机遇。

关键词:人工智能;道德主体;道德归责;心智感知;道德心理学
中图分类号:B848 **文献标志码**:A **文章编号**:1003-5184(2025)04-0300-09

1 引言

随着人工智能(artificial intelligence,AI)系统在医疗、司法、交通等领域的广泛应用,其在道德决策过程中的角色日益受到关注。这些系统不仅协助医生制定诊疗优先级、辅助雇主进行人事筛选、支持法官作出保释决定,甚至在自动驾驶情境中权衡行人与乘客的生命优先级。这种深度嵌入社会实践的趋势,迫使我们重新思考AI的道德地位及其所引发的责任归属问题(Bonnefon et al.,2024)。

传统道德心理学主要聚焦三类道德主体:人类(具备完全道德地位)、动物(拥有有限道德地位)以及超自然存在(如神灵,具有象征性道德意义)。然而,AI技术的快速发展使得AI可能成为一类全新的道德主体(Bonnefon et al.,2024)。AI系统展现出独特的道德双重性:一方面,它们被赋予道德主体(moral agent)的地位,需要做出具有道德意义的决策;另一方面,它们也可能成为道德受体(moral patient),引发人类的情感回应和道德关怀(Gray et al.,2012)。

虽然AI已经在许多情境中被赋予进行道德决策的能力,但是AI能否为其做出的道德决策承担责任,还存在争议。一方面,有研究强调,AI缺乏归责所需的核心能力,如意识、情感与道德理解,因而不具备成为归责对象的资格(Heersmink & Knight,2018;Malle et al.,2015)。另一方面,越来越多研究指出,归责并非纯粹的理性评估,而是会受到功能需求与具体情境的影响(Lima et al.,2023)。在特定条件下,即便面对无心智的技术系统,个体也会出于规范维系或心理调节的需要,将道德责任归于AI(Waytz et al.,2014)。上述争论进一步引出了“责任缺口”这一问题:当AI系统被视为应当承担部分或全部道德责任的主体时,原本应该承担责任的人类行为者(如开发者、操作者或决策者)就可能因此被减轻责任甚至完全免责,导致无人对决策后果承担充分道德责任(Matthias,2004)。在这种情境下,AI系统有可能被滥用作“道德挡箭牌”,成为人类逃避问责的工具(Lima et al.,2023)。

因此,厘清人类如何对AI道德行为进行理解和归因,识别哪些条件会促使人们将AI视为应受奖惩的道德主体,不仅具有理论意义,也具有现实紧迫性。一方面,从学术角度看,AI作为一种全新的“类主体”,挑战了传统道德心理学关于主体性与责任

* **基金项目**:江西省高校人文社会科学基地项目(Jd19063),江西省社会科学规划项目(21JY09),江西省教育科学规划项目(21ZD021)。
通信作者:董圣鸿,E-mail:shdong@jxnu.edu.cn。

归属的边界认知，迫使我们重新审视“道德主体”和“道德受体”的内涵（Bonnefon et al.，2024）。另一方面，从实践需求角度看，厘清AI的道德主体性对于构建合理的责任分配框架、规范人机协作中的道德问责，以及防止技术滥用与道德推诿具有关键意义（Matthias，2004；Lima et al.，2023）。缺乏对此问题的深入理解，将使人类社会在应对AI引发的伦理争议时难以形成有效的治理策略，进而加剧公众的不信任与道德困惑。

本文将围绕AI作为道德主体带来的伦理挑战以及人类如何对其进行道德归责展开系统梳理，主要包含两大部分。其一为AI作为隐性（如算法偏见）和显性（如道德困境解决）道德主体时的道德表现。其二为AI的道德归责问题，具体包括以下三个方面：AI是否应该被视为可归责的道德主体，人们对AI与人类的道德归责中的非对称性，以及影响人们对AI进行道德归责的心理与社会因素。最后，本文将在总结现有研究的基础上，提出未来研究展望。

2　AI的道德主体性

2.1　隐性道德主体：以算法偏见为例

尽管有些AI系统并未被有意识地嵌入道德价值观，其行为结果却常常带来明确的道德后果。比如，医疗AI因误诊造成患者伤害，内容推荐算法将未成年人引导至暴力或不当内容，人脸识别系统误将无辜者识别为潜在威胁，机器故障造成人员伤亡等。这类系统没有直接进行道德判断，但其行为结果却直接或间接引发了道德责任，可被视为“隐性道德主体”（implicit moral agents，Bonnefon et al.，2024）。

算法偏见作为一种结构性和系统性的伦理风险问题，是AI隐性道德主体的重要体现。研究表明，AI系统不仅可能继承现实世界中的不平等，还可能在多个领域加剧已有的社会偏见。例如，美国司法系统应用的再犯风险评估工具COMPAS被发现系统性地高估黑人群体的再犯风险（周尚君等，2019）。一种将医疗支出作为健康需求指标的算法导致黑人患者在医疗资源配置上遭到显著低估（Obermeyer et al.，2019）。AI招聘工具在招聘过程中对女性存在偏见（Dastin，2022），以及面部识别技术对深色皮肤群体更高的误识率（Buolamwini & Gebru，2018）。最近的一系列综述进一步指出，在招聘、面部识别、语音识别和医疗等诸多场景中，AI系统普遍存在显著的性别与种族偏见（An et al.，2024；Nazer et al.，2023；Shah，2024）。

值得注意的是，算法偏见并非源于单一环节，可能贯穿AI开发与部署的全周期。Nazer等人（2023）对医疗领域的AI开发过程进行了拆解，指出在以下五个阶段中均可能引发不同形式的偏见：（1）问题设定阶段，若忽视包容性和公平性，研究议题本身就可能因社会结构性不平等（如种族、性别、残疾）而偏向某些群体。例如，美国在囊性纤维化（主要影响白人）的研究资助显著高于镰状细胞病（主要影响黑人）；（2）数据收集阶段，偏见可能源自抽样失衡（如以白人男性为主要样本）、标签偏误或遗漏关键社会变量（如残疾状态），从而影响模型的代表性与公平性；（3）数据预处理阶段，常见的数据清洗方法如均值填补、异常值剔除等，若未考虑群体差异，可能在无意中压制了边缘群体的特征分布，例如忽略截肢者的极端体重值或特定族群数据的结构性缺失；（4）算法开发与验证阶段，包括模型选择、训练集划分与过拟合等问题，可能导致模型在非主流群体中预测失准，进一步削弱其泛化能力；同时，“黑箱”模型的不可解释性也限制了问责机制的建立；（5）模型部署与应用阶段，即使模型在初始测试中表现良好，长期应用中也可能因“数据漂移”而性能下降，如人口结构或诊疗标准的变化带来协变量或概念漂移，导致模型输出偏离真实需求。

算法偏见的广泛存在，推动了关于算法公平性的理论与实证研究。但是，“公平”的定义高度依赖情境，不同场景对公平的理解存在显著差异：在大学招生中，公众更关注群体比例的代表性；在刑事保释评估中，重点在于假阳性率的一致性；而在信贷决策中，优先考虑的是预测准确度的公平，即资源效率最大化（Bonnefon et al.，2024）。由于各种公平标准难以同时满足，如何在算法开发中权衡和选择适当的公平准则仍是尚未解决的重要研究问题。

最后，尽管公众对AI系统可能带来的偏见有所担忧，但实证研究显示，相较于人类偏见，公众对算法偏见的容忍度普遍更高（Bigman et al.，2023；Hidalgo et al.，2021）。某些曾受到人类歧视的群体甚至对AI系统的接受度更高，认为AI系统有可能做出比人类行为者更公正的决策（Bigman et al.，2021；Pammer et al.，2021）。这一现象揭示了公众对于“技术中立性”的期待。

2.2 显性道德主体:AI 在道德冲突中的决策

随着 AI 技术的快速发展,越来越多的系统被用于处理道德困境。例如,自动驾驶汽车在面对无法避免的碰撞情形时,可能不得不在保障乘客生命与保护行人安全之间做出艰难抉择;又如某些专门设计用于伦理咨询或集体决策的系统(如 Delphi 系统),其核心功能正是模拟和处理复杂的道德困境。这类需要面对并解决道德冲突的 AI 系统被称为显性道德主体(explicit moral agents, Bonnefon et al., 2024)。

2.2.1 AI 与人类的价值对齐

尽管近年来 AI 技术发展迅速,现阶段的 AI 系统仍未达到强人工智能(AGI)的水平,因此尚不具备自主生成或内化道德规范的能力。当 AI 系统需要面对并解决道德冲突时,其道德判断完全依赖于外部预先设定的规则、偏好与价值体系。如何确保 AI 系统在面对道德冲突时能够做出符合人类社会道德标准的决策,即所谓 AI 与人类"价值对齐问题"(value alignment problem, Gabriel, 2020),成为 AI 伦理设计的核心议题。

传统 AI 系统的价值输入主要来源于三类核心信息源。首先,开发者及其团队的主观价值观与偏好会直接影响系统的目标设定与行为约束;其次,伦理学专家提出的规范性理论与指导原则为 AI 提供了系统化、哲学化的伦理框架;第三,通过大规模问卷调查、行为实验或社会模拟等方法获得的公众道德倾向,也为 AI 系统提供了社会性和情境化的道德数据支持,例如著名的"道德机器实验"(Moral Machine experiment, Awad et al., 2018)。然而,人类道德判断本身受到强烈的情境依赖性影响(Tassy et al., 2013),这种情境敏感性加大了 AI 伦理规则编码的复杂度.

此外,当不同来源的道德偏好或价值体系之间发生冲突时,AI 系统应如何进行整合? 现有文献提出了三种主流路径:(1)民主机制:通过收集和聚合多数群体的意见来决定 AI 的价值导向;(2)效益最大化策略:基于功利主义原则,以最大化整体幸福或效用为目标优化 AI 决策;(3)公平最大化模式:强调保障少数群体权益与脆弱群体利益,力图减少结构性不平等与系统性歧视(Gabriel, 2020)。

2.2.2 生成式 AI 的道德判断能力

然而,这些依赖外部预先输入价值规则的传统 AI,难以应对复杂多变的实际情境。与此不同,近年来崛起的生成式 AI(Generative AI),尤其是大语言模型(LLMs),并不直接依赖明确预设的道德规范,而是通过大规模文本语料的学习,在潜移默化中"内化"并模拟人类的价值观与道德推理模式。这种基于数据驱动的道德学习路径,为解决 AI 的价值对齐问题提供了新的思路和工具,也由此成为研究 AI 道德判断能力的重要突破口。

一项系统性研究发现,主流语言模型(如 GPT 系列、BERT 和 RoBERTa)能够以较高的准确率判断某种行为是"对"的或是"错"的。随着训练数据的扩大,其道德判断结果也更接近于人类的判断结果。并且这些模型的道德判断能力并非源自直接编程的伦理规则,而是在大规模语料学习过程中,间接吸收了包含法律条文、伦理讨论、宗教训诫和日常道德规范等文化内容,从而构建出复杂的伦理结构(Schramowski et al., 2022)。与此相一致,Dillion 等人(2023)以 464 个典型道德场景对 GPT-3.5 进行测试,发现其与人类的判断高度相关($r = 0.95$),并表现出跨性别、跨年龄、跨时间的一致性。

为了进一步检验 LLMs 的道德倾向与人类的道德倾向是否相同,Guo 等人(2023)采用道德基础问卷(Haidt, 2007),对多个主流模型进行了检验。结果显示与人类样本相比,LLMs 更加重视"权威-尊重"、"群体-忠诚"与"纯洁-神圣"等道德维度,这与现实中政治保守主义者的道德偏好存在高度一致性。Jiang 等人(2025)将 LLMs 作为虚拟伦理专家引入德尔菲(Delphi)系统,结果显示多个模型的判断一致性高达 85%,显著高于人类判断的一致性(64%)。他们进一步根据道德基础理论(Haidt, 2007)对模型的道德判断理由进行编码分析发现,LLMs 能有效遵从人类的道德准则,如"伤害最小化"、"公平对待"与"忠诚原则"。

不仅如此,生成式 AI 在道德角色方面具有高度的可塑性。研究表明,人们可以通过提示词操控 LLMs 的人格设定,在特定的人格框架下,LLMs 能够表现出连贯且一致的行为特征(Jiang et al., 2023; Pan & Zeng, 2023)。在此基础上,焦丽颖等人(2025)赋予 LLMs 不同的善恶人格,发现不同人格的 LLMs 的道德判断存在显著差异,其中具有"善良人格"的 LLMs 具有更高的判断一致性,且与人类伦理的一致性也更高。

3　对AI的道德归责

3.1　AI是否应该承担道德责任?

传统的道德归责体系通常建立在三个核心前提之上:行为主体具有意图、能够预见行为后果,并拥有足够的控制力(Malle et al.,2014)。然而,当前的AI系统通常不具备这些关键特征,因此从理论上看,其并不适用于现有的道德归责框架。

尽管如此,研究发现人们在实际判断中仍倾向于将AI视作具备某种程度道德能动性的行为体,并对其进行归责(Lima et al.,2023)。例如,Kahn Jr.等人(2012)发现,当机器人实施不道德行为时,大多数参与者认为机器人应承担一定的道德责任。Malle等人(2015)也指出,当自动驾驶系统未能作出符合功利主义标准的决策时,公众对其的道德指责超过了对人类行为者的指责。

Shank与DeSanti(2018)通过设计七种模拟AI引发道德违规的情境(如假释判决、儿童视频推荐等),要求被试评估四类潜在责任主体(AI系统本身、使用企业、开发者与终端用户)的责任程度。结果显示,整体而言四类主体被归责程度无显著差异;但当被试确信AI行为构成道德违规(约占43%的情境)时,AI系统的归责评分显著上升,而其他三类主体未出现类似趋势。这一发现表明,在特定条件下,公众愿意将道德责任部分归因于AI系统。该倾向在其它应用领域也有所体现。Kneer与Stuart(2021)发现,公众认为AI机器人在引发环境破坏时应当承担一定责任。Lima等人(2021)也指出,在医疗和军事场景中,人们普遍认为AI系统对造成的伤害应负有道德责任。在招聘情境中,Bigman等人(2023)发现,当AI做出具有性别歧视的决策时,人们会对AI产生愤怒、不满等情绪反应,这与面对人类不当行为时的情绪模式相似,进一步反映了人们对AI进行道德归责的心理倾向。

然而,这种道德归责并不等同于赋予AI完整的道德地位。Ladak等人(2024)在综合多项研究后指出,公众通常仅赋予AI"中等程度"的道德能力,即认为其既非完全无责任的工具,也未具备完整的道德人格。这种"有限责任体"的心理模型为公众在不同情境下灵活地分配责任提供了认知基础。

3.2　对AI与人类进行道德归责的非对称性

尽管处于相同情境并产生相同的决策结果,大量研究显示,人们对AI与人类行为体的道德归责存在显著差异,这种现象被称为"人-机道德判断非对称性"(Human-Robot Moral Judgment Asymmetry,Awad et al.,2018;Laakasuo et al.,2025)。这种非对称性提示AI在公众心中并未被视为真正意义上的完全道德主体(Bigman et al.,2023)。本部分从两个方面梳理相关研究:其一,公众对AI寄予更高的理性与功利主义期待;其二,公众对AI行为的归责表现出"过度归责"与"道德反应不足"并存的双重标准。

3.2.1　公众对AI道德决策的理性与功利主义期待

研究表明,公众普遍对AI赋予更高的理性预期,倾向于要求其在道德困境中采取最大化整体利益的功利主义(utilitarianism)选择(Malle et al.,2015;Awad et al.,2018)。Malle等人(2015)在自动驾驶情境下的"电车难题"研究中率先发现,AI若选择"转轨牺牲一人"以拯救多数人,通常获得更高的道德评价;反之,若AI"放任电车撞死五人",则会受到比人类更强烈的道德谴责。

后续研究进一步验证了这类现象。例如,Hristova与Grinberg(2016)在区分偶发性与工具性道德困境的实验中发现,当面临是否主动牺牲个体以保护多数时,AI做出功利选择更容易获得道德认可。此外,褚华东等人(2019)发现,公众对AI功利取向的期待具有情境依赖性。在"个人型"情境(如亲手推人落轨)中,人们更倾向于要求AI进行功利选择;但在"非个人型"情境(如遥控转轨)中,人类与AI之间的道德评价差异并不明显。Chu与Liu(2023)的后续研究重复并扩展了这一发现,进一步强调了情境因素的调节作用。

最近,Malle等人(2025)通过大样本进一步系统验证了上述模式。他们在13项实验中一致发现,在"不作为"选项下,AI受到的谴责显著高于人类;而在"主动牺牲一人"的情境中,AI与人类所受批评程度趋于一致。这一结果凸显了人们对AI更为严格的功利主义期待,即认为AI应当主动最大化多数人利益,而非仅仅"被动中立"。更大规模的跨文化调查结果也支持该观点。此外,在"道德机器"(Moral Machine)项目中,Awad等人(2018)分析了超过4000万次道德判断,发现公众普遍偏好AI遵循功利主义逻辑,如优先保护更多生命、人类高于动物、年轻人高于老年人等。这一系列结果表明,AI在公众心中被赋予更高的道德理性期待,当AI没有

做出功利性决策时会受到更严苛的伦理谴责。

3.2.2 过度归责与道德反应不足:AI 归责的“双重标准”

除了相较于人类更偏向功利主义的道德期待之外,对 AI 和人类道德归责的非对称性还体现在归责程度上的不同,表现为相对于对人类的归责,对 AI 的归责存在道德反应不足(归责过轻)与过度归责(归责过重)的“双重标准”现象。

一方面,在涉及核心伦理价值(如生命权与人类尊严)的问题上,公众对 AI 行为的道德容忍度普遍低于对人类行为的容忍度。例如,Laakasuo 等人(2023,2025)在医疗与临终关怀场景中发现,AI 护士或医生执行“强制用药”或“撤除生命支持”决策时,相较于人类同行更难获得公众的伦理认可。Stuart 与 Kneer(2021)在一项模拟污染地下水源导致公众身体伤害的实验中也发现,即便 AI 的行为尚未造成实际伤害,仅因其存在潜在风险,仍会引发更为严苛的道德评价与归责。此外,Manoli 等人(2025)在职场伦理研究中发现,当 AI 从事不当行为(如植入恶意代码)时,公众不仅谴责该 AI 个体,还容易将负面评价扩展至整个 AI 群体,形成“道德溢出效应”;相比之下,人类的不当行为则更常被视作个体偏差,较少引发类属泛化。

另一方面,公众在面对 AI 实施某些道德不当行为时,却表现出“道德反应不足”的倾向。Bigman 等人(2023)提出“算法道德愤慨缺失”(Algorithmic Outrage Deficit)概念,指出即便 AI 与人类实施相同程度的性别歧视行为,公众对 AI 的愤慨与惩罚意愿明显较弱。Lima 等人(2023)与许丽颖等(2022)在涵盖性别、种族、学历和年龄歧视的研究中一致发现,当这些歧视由 AI 实施时,相较于人类实施者,公众表现出更低的道德谴责与惩罚动机,尤其是在 AI 被视为“无自由意志”的条件下。这种反应不足亦出现在高风险情境中。Malle 等人(2019)在军事伦理研究中发现,当 AI 无人机选择服从或违抗包含附带平民伤亡风险的命令时,所受到的责备强度变化不大;而人类飞行员若违抗命令,则更易遭遇道德批评和谴责。对 AI 的道德反应不足可能源于的公众持有“科技乐观主义”的观念。Kapania 等人(2022)认为,公众普遍赋予 AI 以积极变革的期待,倾向于将 AI 的偏差视为技术成长的“过渡性问题”,从而压制了应有的伦理警觉与批判。

3.3 影响公众对 AI 道德归责的心理与社会因素

人们是否将 AI 视为道德主体,并认为其应为自身决策结果承担部分或全部责任,受到多种因素的共同作用。以下从三个方面系统梳理这些影响因素,分别为对 AI 的心智属性感知,主观归责信念,以及社会文化因素。

3.3.1 对 AI 的心智属性感知

Gray 等人(2007)提出的“心智二维模型”为理解人类如何赋予 AI 道德主体地位提供了重要理论框架。该模型指出,人们在评估一个实体是否应承担道德责任时,主要依据其在两个心智维度上的表现:一是“能动心智”(agency),包括自我控制、计划、沟通及思维等能力;二是“经验心智”(experience),指感知(如饥饿、疼痛)、情绪体验(如恐惧、愉悦、愤怒)以及意识与个性等相关特质。当前 AI 系统虽通常被认为在能动心智上表现突出,但在经验心智上得分较低,因此难以被视为“完整的道德主体”(Nijssen et al.,2023)。Gray 等人还发现,机器人在能动性评分上甚至高于黑猩猩,但在经验维度上则接近非生命物体。

AI 的能动心智知觉尤其受到其类人性特征的显著影响。研究表明,具备类人外观、语言表达能力和情绪展示等人类特征的 AI(如社交机器人、语音助手)更容易被视为具有行为控制能力,因此在发生错误行为时更容易被归责(De Visser et al.,2016; Ryoo et al.,2024; Waytz et al.,2014; Yam et al.,2022)。相比之下,缺乏具象表现的 AI(如推荐算法、金融系统)由于缺乏可感知的能动性,通常不被认为具备道德责任(许丽颖,2022; Stuart & Kneer,2021)。

Bigman 等人(2019)总结了人们赋予 AI 道德地位时参考的关键要素,包括情境意识、意图性、自由意志、类人外观以及对人类造成潜在伤害的能力。最近,Ladak(2024)进一步系统整理了九项判断 AI 是否具备道德地位的标准,包括感知能力(主观体验)、意识、认知能力、生命性、信息性、社会关系、行为等效性、发展潜力和类别成员资格。Ladak 特别强调感知能力对 AI 获得道德地位的重要性,认为感知能力是 AI 获得道德认同的充分条件。

3.3.2 主观归责信念

值得注意的是,道德归责不仅取决于 AI 的属性,还受制于人们对“机器可被归责”这一观念的接

受程度。Lima等人(2023)指出,个体在判断AI是否应承担道德责任前,会先评估"将责任归于AI是否合理"。若认为合理,人们倾向将AI视为与人类平等的责任主体;否则,可能完全否认其责任。他们进一步总结出三种典型归责模式,分别为完全归责于AI,拒绝对AI进行归责,将AI视为人类的工具且仅赋予有限责任。归责倾向的不一致凸显了道德判断的主观性与建构性。

3.3.3　社会文化因素

社会文化深刻塑造着个体对AI行为的评价和责任归属偏好。Awad等人(2018)在"道德机器实验"中发现,东西方文化在自动驾驶道德困境中的选择差异明显。相较于西方参与者,东亚参与者更不倾向牺牲年长者,以体现对长者和社会等级的尊重。除此之外,Komatsu(2016)和Komatsu等人(2021)在日本与美国样本中发现,日本参与者比美国参与者更倾向将机器人视为具备道德判断能力的行为者,对其介入道德困境的行为接受度更高。但是当AI表现出过度主导性或干预性时,公众对其道德可接受度会明显下降,并且如果机器人出现道德失范,更倾向于对其进行归责(Malle et al.,2025)。闫霄等人(2024)等也发现在中国文化背景下,个体更倾向于赋予AI感知、情感与意图等心理属性,这种"心智赋予"进一步增强了对AI行为的道德归因,增加了归责倾向。

4　总结与展望

人工智能系统,尤其是生成式AI,正日益深入地参与到人类社会生活与道德判断的各个层面。这一变革不仅对传统伦理理论构成挑战,也为心理学研究提供了观察人类道德认知的新视角。本文首先回顾了AI作为隐性与显性道德主体时的道德表现,然后总结了人们对AI道德归责的态度,包括AI是否应被视为可归责主体、对AI与人类进行道德归责时的非对称性,以及影响公众对AI道德归责的心理与社会因素等。但是现在对AI(特别是生成式AI)的道德行为表现以及人们对AI的归责态度还有很多不明确的地方,需要更加深入的研究。未来的研究可围绕生成式AI的道德决策机制,人们对AI的归责偏差,以及因对AI进行归责引发的责任缺口问题展开进入研究。

首先,深入解析生成式AI在道德判断中的决策机制。当前,生成式AI在不同模型架构、提示策略和应用场景中,可能呈现出差异化的道德偏好与倾向。揭示生成式AI的决策逻辑,是评估其在关键情境中稳定性和可信性的前提。此外,生成式AI输出中存在潜在的性别、种族、阶层等偏见,若未被充分意识与干预,可能进一步固化现有社会不平等,甚至放大社会分化。因此,未来需要系统考察AI道德决策的生成机制,推动其透明化与可解释化。

其次,应进一步探索人类对AI道德反应偏差的认知根源。已有研究指出,人们普遍对AI与人类存在不同的道德期待与归责模式:一方面,人们倾向于以更功利主义的视角评价AI,即期望AI做出最理性、最优化的选择;另一方面,在面对AI引发的错误或伦理失误时,公众的归责模式呈现高度情境依赖性,有时会表现出对AI的归责过度,有时则表现出归责不足。然而,关于这一非对称性形成的深层心理机制与社会调节因素,仍有待系统揭示。未来研究应结合社会认知、归因理论与人机交互等视角,进一步阐明影响公众归责判断的内外部因素,以促进人机协作中的信任和合作。

第三,需深入探讨AI引发的责任缺口现象及其心理与社会机制。AI决策的不透明性与复杂性,往往导致责任主体模糊、责任链断裂,进而引发公众对责任归属的混乱认知。此外,越来越多的实证研究显示,个体在特定情境中会倾向于将错误或负面后果归因于AI系统本身,这可能为相关人员或组织提供逃避责任的心理契机,加剧伦理风险。进一步的研究应聚焦于哪些"类人属性"(如解释性语言、互动风格、拟人化外观等)会强化或削弱公众的归责倾向,并探讨文化背景如何塑造归责模式。例如,在集体主义导向较强、社会等级分明的文化中,个体更可能将责任归于集体或技术系统,而非个体开发者。这一跨文化视角对于丰富责任判断的心理模型,完善全球视野下的AI伦理治理具有重要意义。

综上所述,AI,尤其是生成式AI,在道德判断与责任归属等方面的深度参与,正在重新塑造人类社会的价值体系和心理边界。未来研究应坚守以人为中心的立场(许为 等,2021;Xu,2019),聚合认知心理学、社会心理学、计算机科学与伦理学等多学科资源,开展跨层次、多方法的系统性研究,以提升AI系统的公平性、可解释性、可信任性与伦理责任感,最终实现人机协作中更具韧性与适应性的伦理治理模式。

参考文献

褚华东,李园园,叶君惠,胡凤培,何铨,赵雷.(2019).个人-非个人道德困境下人对智能机器道德判断研究.*应用心理学*,*25*(3),262-271.

焦丽颖,李昌锦,陈圳,许恒彬,许燕.(2025).当 AI“具有”人格:善恶人格角色对大语言模型道德判断的影响.*心理学报*,*57*(6),929.

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.

许为,葛列众,高在峰.(2021).人-AI 交互:实现“以人为中心 AI”理念的跨学科新领域.*智能系统学报*,*16*(4),605-621.

闫霄,莫田甜,周欣悦.(2024).中西方文化差异对虚拟人道德责任判断的影响.*心理学报*,*56*(2),161.

周尚君,伍茜.(2019).人工智能司法决策的可能与限度.*华东政法大学学报*,(1),53-66.

An,J.,Huang,D.,Lin,C.,& Tai,M.(2025).Measuring gender and racial biases in large language models:Intersectional evidence from automated resume evaluation. *PNAS Nexus*,*4*(3),pgaf089.

Awad,E.,Dsouza,S.,Kim,R.,Schulz,J.,Henrich,J.,Shariff,A.,... & Rahwan,I.(2018).The moral machine experiment. *Nature*,*563*(7729),59-64.

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.

Bigman,Y. E.,Yam,K. C.,Marciano,D.,Reynolds,S. J.,& Gray,K.(2021).Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior*,*122*,106859.

Bonnefon,J. F.,Rahwan,I.,& Shariff,A.(2024).The moral psychology of Artificial Intelligence. *Annual Review of Psychology*,*75*(1),653-675.

Buolamwini,J.,& Gebru,T.(2018,January).Gender shades:Intersectional accuracy disparities in commercial gender classification. *In Conference on Fairness,Accountability and Transparency*(pp.77-91).PMLR.

Chu,Y.,& Liu,P.(2023).Machines and humans in sacrificial moral dilemmas:Required similarly but judged differently? *Cognition*,*239*,105575.

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

De Visser,E. J.,Monfort,S. S.,McKendrick,R.,Smith,M. A.,McKnight,P. E.,Krueger,F.,& Parasuraman,R.(2016).Almost human:Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology:Applied*,*22*(3),331.

Dillion,D.,Tandon,N.,Gu,Y.,& Gray,K.(2023).Can AI language models replace human participants? *Trends in Cognitive Sciences*,*27*(7),597-600.

Gabriel,I.(2020).Artificial intelligence,values,and alignment. *Minds and Machines*,*30*(3),411-437.

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

Gray,K.,Young,L.,& Waytz,A.(2012).Mind perception is the essence of morality. *Psychological Inquiry*,*23*(2),101-124.

Guo,S.,Mokhberian,N.,& Lerman,K.(2023,June).A data fusion framework for multi-domain morality learning. *In Proceedings of the International AAAI Conference on Web and Social Media*(Vol.17,pp.281-291).MIT Press.

Heersmink,R.,& Knight,S.(2018).Distributed learning:Educating and assessing extended cognitive systems. *Philosophical Psychology*,*31*(6),969-990.

Hidalgo,C. A.,Orghian,D.,Canals,J. A.,De Almeida,F.,& Martin,N.(2021).*How humans judge machines*. MIT Press.

Haidt,J.(2007).The new synthesis in moral psychology. *Science*,*316*(5827),998-1002.

Hristova,E.,& Grinberg,M.(2016).Should moral decisions be different for human and artificial cognitive agents? *In Proceedings of the Annual Meeting of the Cognitive Science Society*(Vol.38).MIT Press.

Jiang,H.,Zhang,X.,Cao,X.,Breazeal,C.,Roy,D.,& Kabbara,J.(2023).Personal LLM:Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv*:2305.02547.

Jiang,L.,Hwang,J. D.,Bhagavatula,C.,Bras,R. L.,Liang,J. T.,Levine,S.,... & Choi,Y.(2025).Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*,1-16.

Kahn Jr,P. H.,Kanda,T.,Ishiguro,H.,Gill,B. T.,Ruckert,J. H.,Shen,S.,... & Severson,R. L.(2012,March).Do people hold a humanoid robot morally accountable for the harm it causes? *In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*(pp.33-40).PMLR.

Kapania,S.,Siy,O.,Clapper,G.,Sp,A. M.,& Sambasivan,N.(2022,April).“Because AI is 100% right and safe”:User attitudes and sources of AI authority in India. *In Proceedings of the* 2022 *CHI Conference on Human Factors in Computing Systems*(pp.1-18).MIT Press.

Kneer,M.,& Stuart,M. T.(2021,March).Playing the blame

game with robots. *In Companion of the* 2021 *ACM/IEEE International Conference on Human – Robot Interaction* (pp. 407 – 411). PMLR.

Komatsu, T. (2016, March). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. *In* 2016 11*th ACM/IEEE International Conference on Human – Robot Interaction* (*HRI*) (pp. 457 – 458). IEEE.

Komatsu, T., Malle, B. F., & Scheutz, M. (2021, March). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan. *In Proceedings of the* 2021 *ACM/IEEE International Conference on Human – Robot Interaction* (pp. 63 – 72). IEEE.

Laakasuo, M., Kunnari, A., Francis, K., Košová, M. J., Kopecky, R., Buttazzoni, P., ... Hannikainen, I. (2025). Moral psychological exploration of the asymmetry effect in AI – assisted euthanasia decisions. *Cognition*, *262*, 106177.

Laakasuo, M., Palomäki, J., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., ... Francis, K. B. (2023). Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology*, *53*(1), 108 – 128.

Ladak, A., Loughnan, S., & Wilks, M. (2024). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*(1), 27 – 34.

Ladak, A., Wilks, M., & Anthis, J. R. (2023). Extending perspective taking to nonhuman animals and artificial entities. *Social Cognition*, *41*(3), 274 – 302.

Lima, G., Grgić – Hlača, N., & Cha, M. (2021, May). Human perceptions on moral responsibility of AI: A case study in AI – assisted bail decision – making. In *Proceedings of the* 2021 *CHI Conference on Human Factors in Computing Systems* (pp. 1 – 17). IEEE.

Lima, G., Grgić – Hlača, N., & Cha, M. (2023, April). Blaming humans and machines: What shapes people's reactions to algorithmic harm. *In Proceedings of the* 2023 *CHI Conference on Human Factors in Computing Systems* (pp. 1 – 26).

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147 – 186. IEEE.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and Well – being*, 111 – 133.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human – Robot Interaction* (pp. 117 – 124). IEEE.

Malle, B. F., Scheutz, M., Cusimano, C., Voiklis, J., Komatsu, T., Thapa, S., & Aladia, S. (2025). People's judgments of humans and robots in a classic moral dilemma. *Cognition*, *254*, 105958.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175 – 183.

Manoli, A., Pauketat, J. V., & Anthis, J. R. (2025). The AI double standard: Humans judge all AIs for the actions of one. *Proceedings of the ACM on Human – Computer Interaction*, *9*(2), 1 – 24.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6), e0000278.

Nijssen, S. R., Müller, B. C., 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.

Pammer, K., Gauld, C., McKerral, A., & Reeves, C. (2021). "They have to be better than human drivers!" Motorcyclists' and cyclists' perceptions of autonomous vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour*, *78*, 246 – 258.

Pan, K., & Zeng, Y. (2023). Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv*: *2307*. 16180.

Ryoo, Y., Jeon, Y. A., & Kim, W. (2024). The blame shift: Robot service failures hold service firms more accountable. *Journal of Business Research*, *171*, 114360.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre – trained language models contain human – like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258 – 268.

Shah, S. S. (2024). Gender Bias in Artificial Intelligence: Empowering Women Through Digital Literacy. *Journal of Artificial Intelligence*, *1*, 1000088.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real – world moral violations. *Computers in Human Behavior*, *86*, 401 – 411.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelli-

gent agents. *Proceedings of the ACM on Human – Computer Interaction*,*5*(CSCW2),1 – 27.

Tassy,S. ,Oullier,O. ,Mancini,J. ,& Wicker,B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*,*4*,250.

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.

Xu,W. (2019). Toward human – centered AI:a perspective from human – computer interaction. *Interactions*,*26*(4),42 – 46.

Yam,K. C. ,Goh,E. Y. ,Fehr,R. ,Lee,R. ,Soh,H. ,& Gray,K. (2022). When your boss is a robot:Workers are more spiteful to robot supervisors that seem more human. *Journal of Experimental Social Psychology*,*102*,104360.

Moral Agency and Responsibility Attribution of Artificial Intelligence

Wang Xiaotao[1,2],Zhang Jing[1],Zhang Fenghua[1],Dong Shenghong[1]

(1. School of Psychology,Jiangxi Normal University,Nanchang 330022;

2. Center of Mental Health Education and Research,Jiangxi Normal University,Nanchang 330022)

Abstract:With the rapid advancement of artificial intelligence(AI),especially generative AI technologies,its role in moral decision – making and human – AI interactions has become increasingly prominent. This development has sparked widespread interest in psychology regarding whether AI can be considered a moral agency. This paper first outlines the moral behaviors exhibited by AI,including implicit moral actions such as algorithmic bias and explicit moral actionssuch as decision – making in moral dilemmas. Secondly,it systematically reviews current research on responsibility attribution to AI,focusing on three main areas:whether AI should be considered a morally accountable agent;the asymmetry in moral responsibility attribution between AI and human actors;and the psychological and social factors influencing moral judgement of AI behaviors. Finally,the paper proposes three directions for future research:(1)in – depth analysis of the decision – making mechanisms underlying moral judgments in generative AI;(2)systematic exploration of the cognitive mechanisms behind human biases in moral attribution to AI;and(3)investigation of the psychological and social mechanisms underlying the phenomenon of responsibility gaps caused by AI. In conclusion,the study of AI's moral behavior broadens the psychological understanding of moral cognition and responsibility attribution,while presenting new challenges and opportunities for fairness and ethical governance in future AI systems.

Key words:artificial intelligence;moral agency;moral responsibility attribution;mind perception;moral psychology