科学通报 2024 年 第 69 卷 第 11 期: 1406 ~ 1416
进 展
《中国科学》杂志社
SCIENCE CHINA PRESS

# 人工智能决策的道德缺失效应及其机制

胡小勇[1], 李穆峰[1], 王笛新[1], 喻丰[2*]

1. 西南大学心理学部, 西南大学人格与认知教育部重点实验室, 重庆 400715;
2. 武汉大学哲学学院心理学系, 武汉 430072
* 联系人, E-mail: psychpedia@whu.edu.cn

2023-10-25 收稿, 2024-01-31 修回, 2024-02-01 接受, 2024-02-02 网络版发表
国家社会科学基金西部项目(23XSH003)资助

**摘要** 随着技术的不断创新与发展, 人工智能在人类生活中的重大决策方面发挥着日益重要的作用. 然而, 人工智能的广泛应用也带来了一系列道德挑战, 其中最突出的一个是人们对人工智能的不道德决策往往反应(道德评估、道德愤怒和道德惩罚)较弱, 表现出与对人类决策不同的行为模式. 这一现象将导致严重的社会问题. 心智感知理论和相关研究表明, 造成这种现象的主要原因是人们认为人工智能比人类缺乏能动性和体验性; 而拟人化与提高期望则是有效的干预策略, 可以增强人们对人工智能决策的道德敏感度. 与其他学科中"算法伦理"研究主要从设计层面探讨公平算法的原则和方法不同, 心理学视角研究更关注人们对人工智能与人类决策心理反应的差异, 这对于如何有效地应对算法偏见带来的社会问题, 如何构建公平算法提供了新的思路, 也为"算法伦理"研究提供了一个新的视角. 由于人工智能具有发展性和快速迭代性, 该领域的机制及干预策略有待未来研究进一步澄清.

**关键词** 人工智能, 道德缺失效应, 心智感知理论, 拟人化

人工智能是一种利用计算机科学和其他相关领域的理论、方法和技术, 构建具有某种程度的智能系统, 它能够执行不同类型和层次的认知功能, 如感知、推理、学习、决策、创造等[1]. 人工智能已展现出强大的解决问题能力, 并越来越多地用于做出深刻影响人们生活的决策, 如谁会被监视[2]、谁会获得医疗保健资格[3]、谁会获得金融信贷[4]等. 随着人工智能被广泛应用于人类社会, 人工智能的道德问题愈加凸显. 例如, 一种用于筛选求职者的算法为男性分配的就业能力分数始终高于具有类似资格的女性[5]; 还有一种用于评估和管理患者风险的算法, 严重低估黑人患者的健康需求, 即使他们的病情比白人患者更加严重, 也只能得到相同的风险评分, 导致黑人患者获得额外治疗的机会远低于白人患者[6]; 此外, 在信贷借贷[7]和个人隐私信息[8]等领域都记录了类似的人工智能不道德行为的实例. 许多研究从技术、法律和伦理的角度对人工智能决策的道德问题进行了探讨[9]. 然而, 作为人工智能的最终作用对象, 人类本身对于人工智能道德问题的看法同样至关重要, 它会很大程度地影响到人类对人工智能的接受、使用、监管以及未来发展方向[10~12]. 因此, 近年来越来越多的研究从心理学的层面考察人工智能决策的道德问题.

针对人工智能的不道德行为, 许多心理学方面的研究发现, 当人工智能成为道德决策主体时, 人们也会对人工智能做出道德判断[12,13]. 但是与人类不同, 人们对人工智能的道德反应更弱, 更少将责任归于人工智能; 即使做出和人类同等程度的不道德决策, 人们仍然不愿意去惩罚和反对人工智能[14~16]. 这一现象会导致许多严重的社会问题. 首先, 它会促使公司将人工智能视为道德减负工具, 通过使用人工智能进行各类决策

**引用格式:** 胡小勇, 李穆峰, 王笛新, 等. 人工智能决策的道德缺失效应及其机制. 科学通报, 2024, 69: 1406–1416
Hu X Y, Li M F, Wang D X, et al. Reactions to immoral AI decisions: The moral deficit effect and its underlying mechanism (in Chinese). Chin Sci Bull, 2024, 69: 1406–1416, doi: 10.1360/TB-2023-1094

以避免承担责任. 比如, 在人事招聘中使用人工智能来歧视特定性别或年龄段的应聘者, 这会使得人们对此类不道德行为的警觉和抵触情绪下降, 从而导致公司逃避应该承担的责任[14,17]. 其次, 人工智能的道德缺失效应将会导致遭受伤害群体的诉求和维权难度大大提高, 甚至可能被有意忽视. 同时也会导致社会公众对于此类群体的同情心和关注度明显降低, 使得他们失去本该得到的社会支持和关爱[18]. 例如, 在被人工智能误诊的情况下, 受害者可能面临高额的医药费用, 他们面临身体和心理上的巨大压力以及维权困难等问题, 但他们很难得到公众、医疗机构和政府等方面的支持. 最后, 当人们不再对人工智能不道德行为做出及时而强烈的反应时, 这种效应将会与人工智能被广泛应用的形势结合, 导致社会道德标准的不断降低, 瓦解人们对正义和公正的坚定信仰, 进而对社会的基本道德准则造成持久性的伤害. 鉴于人工智能道德缺失效应的严重社会影响, 需要深入探究人类在道德上对于人工智能应用的看法及其背后的机制, 从而有助于建立针对人工智能应用的统一道德评价标准和行为准则, 以确保人工智能的安全、可靠和良性发展, 进而更好地为人类发展做出贡献.

## 1 人工智能决策的道德缺失效应

在教育、招聘和司法等人工智能正在被广泛应用的领域, 与人类做出的不道德决策相比, 人们对于人工智能的不道德决策的道德反应更小, 会产生更少的责备和责任归因, 以及更低程度的道德愤怒、道德惩罚和道德行动倾向, 研究者们把这种现象称之为人工智能的道德缺失效应(artificial intelligence moral deficit effects)[14].

在道德认知方面, 人们对于人工智能的不道德决策会产生更少的责任归因以及更大的错误容忍性. 一项针对数千名对象的大型在线调查询问了人们对人工智能在责任承担方面的看法, 结果显示, 人们认为人类往往需要承担较高的责任, 而人工智能和机器人只需承担中低程度的责任[19]. 除调查研究外, 实验研究则进一步指出了人工智能决策与道德认知之间的因果关系. 例如, 相比于人类药剂师, 当人们了解到错误的药物处方是由机器人做出时, 人们对于机器人药剂师的责任归因更少[20]. 事实上, 除了医疗领域, 人工智能的道德缺失效应在各种道德领域均广泛存在. 例如, 一些研究者在4种道德情景(假释量刑偏见、选美歧视、视频推荐偏见和网络暴力)中考察了人们对于人工智能和人类道德决策的道德判断差异. 结果发现, 面对同等程度的道德违反, 相比于人类, 人们认为人工智能的错误性更小[21]. 在另一项包括伤害、不公平、背叛等更多种类道德情景的研究中, 研究者发现, 与人类相比, 人工智能做出侵犯决策后同样更容易被认为没有错误[22]. 来自跨文化的研究证据也支持上述发现. 例如, 在30余种不道德决策场景中, 有研究者让来自亚洲、非洲和美洲的被试对人类和人工智能的不道德决策进行评价, 结果发现人工智能均被评价为错误性更小、更不应该受到责备以及更不需要对决策负责[15]. 甚至, 当人工智能充当合作者或提供建议者角色而非独立决策者时, 该效应仍然存在. 以人机联合协作为例, 研究发现在特定任务背景下, 相对于人类专家提供的决策建议, 当人工智能提供决策建议时, 人们普遍认为人工智能承担的责任较少[23].

在道德情感方面, 人们对于人工智能的不道德决策产生的道德愤怒等负面情感较少. 例如, 在一项实验室实验中, 研究者们发现人工智能与人类在信任背叛的不道德决策情境下引发人们的道德愤怒水平存在显著差异. 在这项实验中, 被试需要与人工智能或人类合作完成一个金钱分配任务. 结果显示, 无论是人工智能还是人类, 只要做出了背叛信任的不道德决策, 都会激起被试的道德愤怒. 然而, 相比于人类, 人工智能引发的道德愤怒水平要显著低一些[24]. 此外, 通过情境启动实验, 研究者们考察了在招聘情境中, 人力资源专家或人工智能招聘程序做出性别歧视决策时时人们的愤怒水平, 进一步证实了人工智能不道德决策导致愤怒情感缺失效应. 具体来说, 在该研究中, 研究者让被试首先阅读一份关于招聘性别歧视的材料, 该材料显示人类招聘官或者招聘算法给女性申请者的岗位胜任评分总是低于类似资格的男性. 在阅读完该材料后, 研究者使用0(非常不同意)到100(非常同意)的滑块测量被试的道德愤怒水平(例如, “我对人类招聘官或招聘算法的歧视行为感到愤怒”). 结果发现, 人们对产生歧视行为的招聘算法的道德愤怒水平显著低于人类招聘官[14]. 随后, 在验证性现场实验中, 研究者通过一个真正经历过上述情境启动实验中描述的招聘过程的人群, 来检验算法愤怒缺失的普遍性. 该研究中, 研究者通过联系5家科技公司的人力资源经理, 招募了从事科技行业的被试, 并将被试随机分配到算法决策组或人类决策组. 被试同样先阅读招聘性别歧视的材料, 之后被告知歧

视决策由人工智能算法或人力资源专家做出. 最后, 通过5个条目的道德愤怒量表测量了所有被试的愤怒水平. 结果发现, 相对于人力资源专家决策组, 人工智能决策组被试愤怒水平更少[14].

在道德行为及行为倾向方面, 人们对于人工智能的不道德决策同样会产生更少的反对行动和惩罚倾向. 以性别歧视为例, 在一项实验研究中, 被试了解到某公司的女性员工比男性员工获取绩效奖金的难度更高, 而这一绩效制度是由人类管理者或算法设计的, 后来一群员工因性别歧视问题向公司发起请愿, 研究者询问被试是否会考虑签署请愿书. 结果发现, 相比于人类管理者, 当歧视性绩效制度由人工智能算法做出时人们签署请愿书的倾向会显著下降[18]. 在学历歧视情境中, 该效应得到了进一步验证. 例如, 一项研究首先让被试了解到某公司在招聘时表现出学历歧视, 会优先淘汰某些低学历申请者. 之后, 使用3个条目测量了被试的惩罚倾向(例如, “你认为这个招聘官/招聘算法应该为这种行为受到多大程度的道德惩罚?”). 结果发现, 算法组的惩罚倾向评分显著低于人类组. 此外, 在他们的实验中同样发现了人工智能的道德缺失效应[16]. 除了歧视这一道德类型, 人们对人工智能直接伤害的道德惩罚倾向同样更低. 例如, 有研究者以锡耶纳大学(University of Siena)的学生为被试, 向他们呈现了一个关于伤害的情景材料. 在该情境中, 破坏者(人类/机器人)在大学校园引爆炸弹导致了5名学生的死亡. 为了控制无关因素对被试判断的影响, 场景只是客观描述了伤害事实, 而没有对破坏者和被害者的情绪和感受进行描述. 在阅读情景后, 被试需要对破坏者确定适当的惩罚力度(判处监禁年数从0到50年). 结果发现, 相比于人类破坏者, 人们对机器人破坏者的惩罚力度更低, 判处监禁年数较人类破坏者更少[25].

上述研究表明, 在各类不道德场景中均存在针对人工智能决策的道德缺失效应. 与人类相比, 人们对于人工智能的道德认知、道德情感以及道德行为及行为倾向等反应均更弱. 即使做出与人类同等程度的错误决策, 人们仍然会对人工智能产生更低程度的道德反应.

## 2 人工智能道德缺失效应的机制

为什么人工智能作为道德行为者时做出不道德的决策会导致更少的道德愤怒、道德惩罚欲望、责任归因等道德反应? 许多研究表明, 感知实体存在心智是人们产生道德反应的重要前提, 人们只有在感知到主体存在一定程度的心智时才会产生道德反应[26,27]. 通常, 人们沿着两个独立的维度感知心智: 能动性和体验性. 赋予一个实体能动性意味着观察者相信这个实体能够像正常成年人一样行动、计划、自我控制、记忆、交流和思考. 赋予一个实体体验性意味着观察者相信这个实体有一些情绪状态, 比如恐惧、痛苦、快乐、愤怒、欲望、个性、意识、自豪、尴尬和喜悦. 在整体的心智归因中, 体验性往往比能动性解释更多的方差, 并且比能动性更本质化, 这表明体验性是区分人类和其他实体的核心维度[28]. 心智感知与现实无关, 能动性可以独立于体验性被感知. 例如, 婴儿被认为在体验性上很高, 但在能动性上很低; 成年人被认为在能动性和体验性上都明显高于人工智能和其他实体[29]. 心智感知涉及赋予其他实体心智能力, 而道德判断涉及给实体贴上好或坏的标签. 依据道德二元论及相关实证研究可知, 心智感知与道德判断密切相关. 具体来说, 当实体被感知为具有能动性时, 那么它会因为道德错误而受到指责[27]; 同样, 当实体被认为具有达到安全、道德适当行为所需的体验性时, 那么它就应该承担责任[12,30,31]. 基于此, 本文以人们对人工智能的能动性和体验性的感知为中介变量, 论证人工智能决策道德缺失效应的内在机制, 指出人们普遍认为人工智能在这两方面都不如人类, 因此人工智能作为道德行为者时做出不道德的决策时, 人们会表现出较弱的道德反应, 例如道德愤怒、道德惩罚欲望、责任归因等.

### 2.1 能动性的中介作用

能动性是人工智能决策道德缺失效应的一个重要机制. 能动性(agency)是指一个实体具有意图、推理、目标追求和交流的能力[28]. 它与道德责任高度相关, 个体自主性越高、意图和动机越强烈, 人们就认为其更应该对自己的决定和行为负责[28]. 人工智能作为道德行为者时, 人们会关注其与道德责任相关的能动性方面. 而大量的研究发现, 尽管人们认为人工智能存在一定水平的能动性心智, 但仍低于人类的能动性水平[30], 因此人工智能做出不道德的决策导致更少的道德愤怒、道德惩罚欲望、责任归因等道德反应.

大量实证证据支持上述理论假设, 即能动性是人工智能道德缺失效应的重要心理机制. Bigman等人[32]认为, 由于机器人的行为受到程序约束, 破坏了人们对其自由意志的感知, 因而人们赋予机器人更少的道德

责任. 以自由意志作为能动性指标, 一系列实证证据确实表明, 自由意志在人工智能道德缺失效应中起着中介作用. 例如, 一些研究者让172名被试阅读一段关于学历歧视的材料, 即某公司的人力资源经理或人工智能招聘算法在对应聘者的简历进行筛选时存在学历偏差, 筛掉了所有硕士研究生以下学历的应聘者, 而该公司的大部分岗位对学历并无硬性要求. 之后测量了被试对人类和算法的自由意志水平的看法以及对他们的道德惩罚欲望. 中介效应检验结果发现, 自由意志的中介作用显著, 即人们认为算法相比人类有更少的自由意志, 因此不倾向于对其进行道德惩罚[16]. 为了增加实验结果的稳健性, 研究者们丰富了歧视类型, 并且通过对被试自由意志信念的实验操纵, 进一步探讨自由意志信念是否是造成道德惩罚欲差异的机制; 结果发现, 普通算法组的道德惩罚欲显著低于相信算法有自由意志组, 而人类组与相信算法有自由意志组的道德惩罚欲无显著差异. 该实验进一步验证了自由意志信念是造成人们对不同歧视主体(人类 vs. 算法)产生不同道德惩罚欲的重要机制[16].

以不道德动机为能动性指标, 一系列实证证据表明, 不道德动机在人工智能道德缺失效应中起着中介作用. 一系列研究发现[14], 感知到的歧视动机在算法歧视的道德愤怒缺失中起中介作用. 在该系列研究中, 被试首先阅读了一段关于性别歧视的材料, 即一项针对某高科技公司的外部审查发现, 尽管收到了大量女性申请软件工程职位的申请, 但该公司几乎没有雇佣女性. 接着, 被试被随机分配到算法组和人类组, 他们分别被告知做出招聘决策的主体为人工智能算法和人力资源经理. 最后, 测量了被试对招聘主体的歧视动机感知以及道德愤怒程度. 结果发现, 歧视动机在算法歧视的道德缺失效应中中介作用显著, 即由于人们认为算法比人类具有更低水平的歧视动机, 因此对做出性别歧视决策的算法产生的道德愤怒程度更低. 随后验证性研究中, 研究者们扩大了研究范围, 考察了算法歧视如何影响人们对使用算法进行决策的公司的判断. 结果发现, 在了解到存在歧视前后, 算法组对公司的道德愤怒增加量比人类组的显著更少, 且歧视动机在其中起中介作用[14].

### 2.2 体验性的作用

体验性是人工智能决策道德缺失的另一个重要机制. 体验性(experience)是指一个实体具有意识、情感和人格的能力; 高体验性的实体能够感受和承受痛苦; 低体验性的实体则麻木不仁、无情感[28]. 感知到的体验性决定了道德责任. 虽然感知到体验性常常被认为与道德对象有关, 但越来越多的研究指出体验性对于一个实体成为道德主体是必要的; 人类正是由于具有情感, 才能够形成完善的道德认知, 成为一个合格的道德主体[12]. 此外, 情绪对于道德判断至关重要, 尤其是同理心被视为是道德判断的核心要素[33,34]. 当实体被认为具有达到道德适当行为所需的体验性时, 那么它就应该承担责任. 然而, 对于人工智能而言, 它在与情感有关的体验性心智方面远低于人类, 缺乏形成道德能力至关重要的情感能力[30]. 人工智能难以理解一种行为对他人是好的还是坏的, 或在道德上是正确还是错误的, 因此它在道德地位上不是一个完整的道德代理人, 从而对它的道德反应会比人类更弱.

体验性在人们对人工智能决策道德反应中的作用得到了初步研究证据的支持. 例如, 在一项由3个系列实验构成的研究中, 研究者将被试被随机分到人工智能故意伤害、意外伤害和控制组, 他们分别了解到某个人工智能由于不同原因杀死了一个工人(为了排除工人对自己的干扰、由于故障没有识别出工人、由于出现在了本不应该出现的地方), 在各自阅读以上场景后, 研究者测量了被试的感知心智水平、感知伤害意图以及对人工智能的责备动机和惩罚水平. 结果显示, 体验性在感知故意伤害对责备动机和惩罚严厉程度的影响中起中介作用, 即在有意伤害条件下, 人们感知到了更高的人工智能的体验性, 因此更倾向于对其做出更多的责备和更严厉的惩罚. 随后, 为了检验伤害的严重性是否是影响结论的重要因素, 研究者们使用相同的范式考察了当人工智能造成人类轻伤(例如割伤手指)时人们如何对其进行道德反应, 结果发现, 体验性同样在感知故意伤害对责备动机和惩罚严厉程度的影响中起部分中介作用[35]. 那么, 对于人工智能决策的道德缺失效应, 体验性在其中起到中介作用吗? 当前已有实证研究给予了肯定回答. 例如, 一项研究探讨了人们对人类和人工智能做出道德或不道德行为的品德判断是否存在差异, 以及这种差异的原因是什么[36]. 研究者设计了4个实验, 每个实验都包含了一系列的道德场景, 如救援、捐赠、谎言、欺骗等. 在每个场景中, 被试需要评价场景中的人类或人工智能的道德水平, 以及他们是否具有体验性. 研究发现, 无论是道德还是不道德的场景, 被试对人工智能的道德判断程度都比对人类

的低, 即他们认为人工智能不如人类善良或邪恶. 进一步发现, 体验性在人们对人类和人工智能的道德判断中起到了中介作用, 即与人类相比, 人们认为人工智能体验性水平低, 进而对其道德判断更加宽容[36].

上述研究表明, 人们之所以会对人工智能的不道德行为的道德反应程度更弱, 除了由于人工智能缺乏能动性以外, 还由于人们认为人工智能较人类缺乏体验性, 因此并不认为其是一个完全意义上的道德代理人, 从而对其不道德行为的反应程度较人类更弱. 道德二元论同时指出, 人们依据能动性与体验性感知实体的心智, 当遇到道德问题时, 这种对心智的感知成为了人们进行道德判断产生的前提; 并且这两个维度相互独立. 基于上述研究发现与该理论观点, 本文认为人工智能决策的道德缺失效应是通过体验性与能动性的平行中介作用实现的(图1). 当人们感知到人工智能存在体验性与能动性时, 道德判断就会发生. 而道德反应的强弱则取决于感知到的能动性与体验性心智的程度, 程度越高, 反应越强烈.

## 3 人工智能决策道德缺失效应的干预

根据心智感知理论, 能动性与体验性是道德判断的主要依据, 只有感知到具有较高的心智水平后人们才能够做出正常的道德判断. 而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. 而通过关注人们感知人工智能心智并对其做出道德判断这一过程, 在探讨提高人工智能心智感知以缓解道德缺失的干预方式时可以从两个关键方面出发: 一是对客体(人工智能)进行干预, 二是对主体(人类)进行干预.

第一, 对于客体(人工智能)而言, 若要提高人们对于人工智能心智水平的感知, 最直接最有效的方式就是提高人工智能的拟人化水平. 拟人化是指人们将人类特征、意图、精神和情感状态以及行为归因于非人类物体的倾向[37]. 大量研究表明, 拟人化上升可以显著提高人们对人工智能的心智感知水平, 并继而提升对人工智能的道德反应. 首先, 对人工智能进行拟人化设计可以有效提高对人工智能的道德反应. 例如, 拟人化可以提高人们对于人工智能汽车的责任归因水平. 在一项单因素被试间实验中, 100名被试被随机分配到3个条件下(正常汽车、自主汽车、拟人化汽车), 实验在实验室环境下进行, 被试驾驶汽车模拟器同时身上连接生理设备以记录生理反应. 在3种条件下, 汽车均会发生车祸. 之后研究者结合生理指标测量并分析了被试对于事故责任的归因, 归因对象包括司机、汽车、汽车公司和设计师. 结果发现, 相比于普通汽车, 拟人化的汽车被认为对事故负有更多的责任[38]. 另一项研究也发现, 与非类人的机器人做出的相同不道德选择相比, 人们对类似人类的机器人做出的选择的评价更低[39]. 因此, 人工智能的拟人化程度会影响人们对它的道德反应水平. 其次, 进一步的机制研究表明, 拟人化是通过提高人们对人工智能心智感知的水平来提高道德反应的. 例如, 拟人化由于提高了人们对于人工智能能动性的感知而增强了人们的报复倾向[40]. 拟人化机器人以及体验性水平更高的机器人更应该为自己的不道德决策受到惩罚[41]. 此外, 具有人形外观的机器人被感知为更有意图性, 从而在做出不道德决策后导致了人们更多的愤怒[42]. 并且, 拟人化可以让人工智能系统更贴近人类的思维和行为模式, 从而改变人们对其能动性和体验性心智的感知, 进而增强人们对做出不道

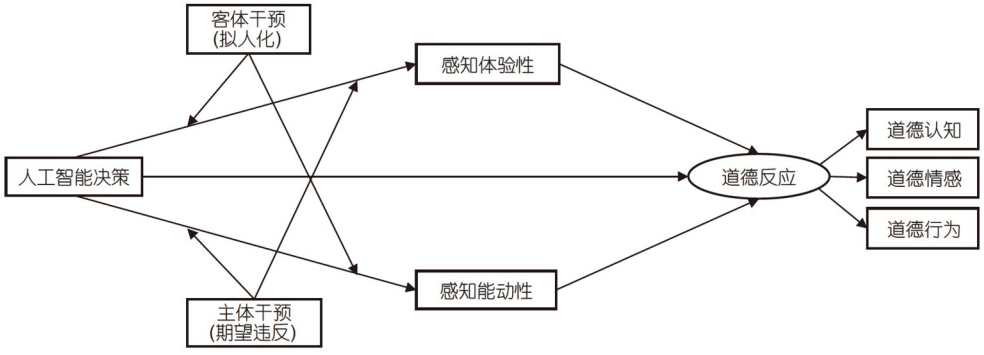


图 1 人工智能决策的道德缺失效应的机制及干预路径
Figure 1 Mechanism and intervention of the moral deficit effect in immoral AI decisions

德决策的人工智能的道德反应强度. 例如, Bigman等人[14]向被试呈现关于算法做出性别歧视决策的信息后, 通过使用指导语来控制算法的拟人化程度; 随后, 测量了被试对算法的类人性、偏见动机的感知水平以及由此产生的道德愤怒水平; 结果表明, 拟人化能够提升人们对人工智能动机水平的感知, 从而引发人们对人工智能不道德决策更多的愤怒情绪. 此外, 许丽颖等人[16]的研究表明, 自由意志在拟人化对道德惩罚的影响中中介效应显著. 也就是说, 拟人化通过提高人工智能自由意志水平, 进而增加了人们的道德惩罚欲望. 综上可知, 对人工智能进行拟人化可以提高人们对于其心智水平的感知, 从而提高人们对于人工智能的道德反应. 在实际应用或实验研究中, 提高对物体的拟人化程度可以通过使物体特征类似于人脸[43]、身体[44], 或赋予其语音识别能力[45]、性别[46]、高信息交互[47]来实现. 其中, 通过赋予人工智能类似人类的面部特征, 可以增强人们对于人工智能的类人性感知, 也可以使人工智能发出内部心理状态信号, 例如积极或消极的意图[48].

第二, 对于主体(人类)而言, 可以通过提高人们对人工智能的主观期望来进行干预, 利用人工智能未达预期时人们产生的消极反应可以“弥补”对人工智能的反应不足. 人们在与人工智能互动时, 会无意识地将人类社会的规则和期望应用在人工智能身上, 并根据感知的温暖和能力水平来推断人工智能在道德困境中的决定和行为[49]. 然而, 人们并不是客观地感知人工智能的功能, 而是会通过感知人工智能外观和属性等信息对人工智能形成特定期望, 但这种期望并不反映其真实功能. 因此, 当人工智能的实际表现与人们的期望不一致时, 就会引发对此类行为的认知情感评估, 从而导致对行为的积极或消极的评价[50]. 因此, 调整人们对人工智能的期望是一种可能的干预方案. 研究证据显示, 期望违反会影响人们对人类与人工智能的道德评价[51]. 对于自动驾驶汽车这样涉及生命安全的领域, 人们对于自动驾驶汽车确实抱有远高于对人类司机的期望[52], 而事故的发生违反了人们的高预期, 因此就可能产生更大的愤怒和指责. 因此, 在对人工智能的道德缺失效应进行干预时, 或许可以提前设定人们对人工智能的合适期望, 当其决策未达到预期时, 产生的消极反应可能可以缓解道德缺失带来的负面效果. 设置人们对人工智能的期望主要有以下3种技术: (1) 准确性指标, 向人们明确地说明人工智能系统的准确性; (2) 基于实例的解释: 利用现实事例向人们解释人工智能系统的运作过程, 提高人们对系统的理解; (3) 性能控制: 允许人们直接调整人工智能系统的性能, 提高对系统的控制感. 通过这些技术, 可以有效提高人们对于人工智能系统的期望, 从而使期望干预发挥作用[53].

综上可知, 拟人化与提高对人工智能的预期均通过提高对人工智能心智感知水平来提高道德反应. 通过拟人化, 人们对人工智能的能动性和体验性心智的感知得到增强, 并由此提高了对人工智能的道德反应; 而提高对人工智能的预期同样需要落到人工智能的心智感知上, 即提高了人们对人工智能的能动性和体验性心智的预期, 并进而在未达预期时产生消极反应, 进而缓解道德缺失(图1). 这两种方式目标都是缓解人工智能的道德缺失, 但二者无论是干预对象还是干预手段上均存在显著差异. 拟人化是从人工智能这一客体为对象来达到缓解道德缺失的目的, 期望违反是从人类这一主体为对象展开干预. 虽然拟人化可能在提高对人工智能的期望水平方面发挥一定作用, 但它更多的是从人工智能系统的角度出发. 期望违反则更强调从人类主体的角度出发, 通过操纵人们对人工智能行为的期望来引导其对系统的评价. 从干预手段上看, 拟人化主要是通过对人工智能进行设计、处理来实现, 而期望操纵主要是通过向人类输入与人工智能有关的某些信息来达到目的. 在实际应用中, 选择拟人化还是期望违反的干预方法应该取决于特定情境的要求以及对道德原则的遵循. 拟人化可能更适用于那些强调用户体验和情感联系的场景, 而期望违反可能更适用于需要调整公众对人工智能伦理标准的看法的情境. 这两种方法都反映了对人工智能与人类交互和道德影响的深刻思考, 而在选择时应该综合考虑它们的优势和局限性, 需要根据具体情境和目标权衡这两种方法的优劣, 以达到更全面、有效且合乎道德的人工智能应用.

## 4 小结与展望

人工智能不仅具有强大的问题解决能力, 而且在影响人类生活的重大决策中发挥着越来越重要的作用. 然而, 在面对不道德的决策时, 人们对人工智能的不道德决策往往反应较弱, 表现出与对人类决策不同的行为模式, 包括对道德评估、道德惩罚欲望和道德行动意愿的减弱. 基于心智感知理论及相关研究证据指出, 感知到人工智能的能动性与体验性较人类低是导致上

述效应的重要原因; 而拟人化则被证实是针对上述机制进行干预, 减少人工智能决策道德缺失效应的有效策略. 本领域研究在多个学科(包括计算机科学、哲学、法学、社会学等)关于“算法伦理”的现有成果基础上, 开辟了新的研究视角, 为构建公平算法提供了新的思路. 具体而言, 与现有“算法伦理”研究主要关注公平算法设计的原则和方法不同, 本领域研究更侧重于探究人们对偏见算法的心理反应及其与对偏见人类的心理反应之间的差异. 揭示这些差异是应对偏见算法引发的社会问题的关键一环. 由于人工智能的发展性与快速迭代性, 该领域还存在大量的问题, 需要未来研究进一步厘清.

第一, 目前已经有许多针对人工智能在道德领域的研究, 本研究更侧重于探究人们对偏见算法的心理反应及其与对偏见人类的心理反应之间的差异, 为构建公平算法提供了新思路, 然而当前该领域研究还存在一些矛盾之处, 有待未来研究进一步明确. 虽然当前大多数研究发现人们对于做出道德决策的人类和人工智能具有不同的道德反应模式, 对人工智能的责任归因、道德愤怒和道德惩罚更少, 但是也有一部分相反的证据表明, 人们对于人工智能的道德反应程度更高. 例如, 一项研究测试了人们对人工智能和人类驾驶员之间的责备程度的差异, 在阅读给定交通事故场景后, 284名被试被要求为驾驶员分配责任级别. 结果发现, 与人类司机相比, 被试更多地指责人工智能司机[54]. 验证性研究也表明, 无论是在简单还是复杂情境下, 人工智能司机犯错时受到的指责比人类司机更多; 当发生道德违规的情况时, 人们往往会把责任推给人工智能, 而忽视了组织、程序员或用户的作用[55]. 对于相同的决定, 与人类相比, 自动驾驶汽车被认为更应受谴责、更不道德、更不值得信赖[56]. 对于这些与人工智能道德缺失效应相反的情况, 很可能是因为存在某种调节道德主体身份与道德反应的变量, 从而使得出现这种矛盾的情况. 那么, 是哪些因素在其中起到调节作用, 当前还鲜有研究予以明确, 需要未来研究系统考察.

第二, 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, 然而该领域研究还不够深入. 尽管许多研究考察了某些变量在导致对人工智能差异化道德判断中的作用, 例如自由意志、自主性以及动机, 但这些研究所考察的内容比较单一, 不成系统, 没有在一个总体的框架之下对差异做出整体性的解释. 因此本文在前人提出的心智感知理论的基础之上, 对现有的机制研究进行了整合, 提出人工智能在心智水平上的缺少是导致差异化道德判断的根本原因. 着重提出除了关注人工智能的能动性, 体验性也是一个必不可少的机制. 需要指出的是, 虽然道德二元论认为体验性主要与道德对象相关, 但越来越多的研究指出体验性对于一个实体成为道德主体是必要的; 人类正是由于具有情感, 才能够形成完善的道德认知, 成为一个合格的道德主体[12]. 这些观点支持了基于心智感知理论的最新发现, 即道德判断需要体验性的参与[57]. 当实体被认为具有达到道德适当行为所需的体验性时, 那么它就应该承担责任[12,30,31]. 正是由于人工智能缺少情绪情感等体验性能力, 它无法成为一个合格的道德代理人, 因而对人工智能的道德判断更弱. 然而, 目前对体验性的研究还不够丰富, 主要集中在对体验性这一整体概念的研究, 如果要形成一个完整人工智能决策道德缺失的理论框架, 未来研究应深入考察体验性心智维度下的具体内容基础上, 综合深入考察能动性与体验性的共同作用机制. 需要注意的是, 虽然许多研究者认为心智感知对于道德或许十分重要, 但它并不足以成为道德的唯一本质. 例如, 有研究者[58]提出, 道德判断的过程是复杂的, 而不是由有意的伤害者和痛苦的受害者构成的, 人们在感知道德事件发生到做出道德判断这一过程中还会关注行动者的行动理由、义务和能力. 此外, 有研究者从责任分散的角度对这种差异进行了解释, 认为由于人工智能是被人类制造出来的, 不是一个独立决策和行动的主体, 因此它的道德责任不仅由人工智能承担, 还应该由研发设计者、公司、程序员等主体进行承担[35], 而人类作为独立行动的主体, 理应承担自身决策的全部责任. 更有研究者从报复差距(the retribution gap)来解释人们对人工智能的道德反应为何不同于人类. 报复差距是指人们对于报复的渴望与缺乏适当的报复对象之间的不匹配, 人们之所以在对待人工智能上与人类存在区别, 是由于人工智能不具备接受惩罚的能力, 对人工智能进行惩罚无法起到惩罚所应有的作用, 例如警告或心理补偿[17]. 因此, 未来的研究应从各个层面深入探讨导致人工智能决策的道德缺失效应各种潜在机制.

第三, 人工智能决策的道德缺失效应对个人、组织以及整个社会都将产生极大的危害, 如何有效干预尚缺乏系统的方案. 研究已证实, 人工智能决策的道德缺失效应使公司在招聘中利用人工智能歧视应聘者,

降低人们对不道德行为的警觉, 使公司逃避责任; 增加受害群体的诉求和维权难度; 甚至, 该效应与人工智能的广泛应用结合, 导致社会道德标准降低, 瓦解对正义和公正的信仰, 对社会道德造成持久伤害[18,59]. 在人工智能迅速普及的大背景下, 如何消除这些危害, 以确保人工智能的安全、可靠和良性发展, 是一个亟需解决的问题. 一些证据显示, 拟人化可以有效缓解人工智能的道德缺失. 通过将人类的声音、形象或行为方式赋予人工智能, 人们会感知到其心智感知水平更高, 越倾向于将其作为一个完整的道德主体来看待, 因而表现出与一般人类接近的道德反应模式, 从而达到缓解道德缺失的目的. 鉴于道德是一种社会实践[60], 能够与人类互动的人工智能系统被视为深深地嵌入在社会结构中, 并被视为“人类的对等体”. 拟人化则使得人们对人工智能进行更多的情绪归因, 并期望它具有同理心, 理解其他实体情感或认知状态的能力, 会增加人们对人工智能心智的感知, 并反过来影响其应受责备等道德反应的程度. 然而, 单纯强调提高对人工智能的道德反应是不够的. 要确保干预手段真实、有效地减少人工智能的不道德行为, 必须提升对人工智能设计者、公司和开发者的道德反应, 这样才能够从根本上防范或减少人工智能做出不道德行为的可能性. 在这方面, Sullivan和Fosso Wamba[35]的研究表明, 意图的存在可以促使人们对人工智能、公司以及开发团队做出更为深刻的责任归因和指责; 当人工智能造成伤害时, 人们认为故意伤害条件下比意外伤害条件下导致的后果更加严重, 因此在故意伤害条件下, 人们会更多地责怪人工智能、公司和开发团队. 而拟人化可以显著提高人们对人工智能的心智感知, 包括对意图的感知. 因此, 当拟人化的人工智能导致负面后果时, 人们更容易感知到其有意图性, 从而加强对事件严重性的认知, 促进相关人员更全面地考虑人工智能在决策时的限制条件, 从而推动道德人工智能的发展. 此外, 人类和人工智能之间的道德反应的不同过程也可能源于他们对各自社会角色和伴随这些角色而来的道德辩护的不同感知[30]. 当人们与人工智能互动时, 他们可能以不同的方式将人工智能视为社会代理, 期望人工智能像具有人性或经验一样参与道德行为[61]. 因此, 除了通过拟人化提升人们对人工智能的心智状态推断外, 调整规范与期望也应能显著调节人工智能决策的道德缺失效应. 初步的证据也显示, 在对人工智能的道德缺失效应进行干预时, 可以提前设定人们对人工智能的合适期望, 当其决策未达到预期时, 产生的消极反应可能可以缓解道德缺失带来的负面效果. 这些研究发现启示人工智能设计师、科学家和政策制定者们应该谨慎地考虑他们的设计决策会对道德产生什么影响. 由于人们在道德判断人工智能时倾向于依赖经验心智, 因此让人工智能在选择和行动时充分考虑道德方面的因素应该成为人工智能时代的优先事项. 然而, 该领域现有研究大都是基于一些小理论开发出的零散的干预策略, 缺乏系统性, 需要未来研究深入考察, 从而有助于建立针对的人工智能应用的统一道德评价标准和行为准则, 以确保人工智能的安全、可靠和良性发展, 进而更好地为人类发展做出贡献.

## 参考文献


1 Rai A, Constantinides P, Sarker S. Next generation digital platforms: Toward human-AI hybrids. Mis Q, 2019, 43: iii–ix
2 Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. Sci Adv, 2018, 4: eaao5580
3 Bates D W, Saria S, Ohno-Machado L, et al. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. Health Affairs, 2014, 33: 1123–1131
4 Gomber P, Zimmermann K. Algorithmic trading in practice. Oxf Handb Comput Econ Finance, 2018, 2: 311–332
5 Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. In: Martin K, ed. Ethics of Data and Analytics: Concepts and Cases. New York: Auerbach Publications, 2022. 296–299
6 Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019, 366: 447–453
7 Bartlett R, Morse A, Stanton R, et al. Consumer-lending discrimination in the FinTech Era. J Financ Econ, 2022, 143: 30–56
8 Shank D B, Gott A. Exposed by AIs! People personally witness artificial intelligence exposing personal information and exposing people to undesirable content. Int J Hum Comput Interact, 2020, 36: 1636–1645
9 Magrani E. New perspectives on ethics and the laws of artificial intelligence. Internet Policy Rev, 2019, 8
10 Langer M, Landers R N. 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. Comput Hum Behav, 2021, 123: 106878

11 Hoffmann-Riem W. Artificial intelligence as a challenge for law and regulation. In: Wischmeyer T, Regul R T, eds. Regulating Artificial Intelligence. New York: Springer International Publishing, 2020. 1–29
12 Bigman Y E, Gray K. People are averse to machines making moral decisions. Cognition, 2018, 181: 21–34
13 Ghotbi N, Ho M T, Mantello P. Attitude of college students towards ethical issues of artificial intelligence in an international university in Japan. AI Soc, 2022, 37: 283–290
14 Bigman Y E, Wilson D, Arnestad M N, et al. Algorithmic discrimination causes less moral outrage than human discrimination. J Exp Psychol Gen, 2023, 152: 4–27
15 Wilson A, Stefanik C, Shank D B. How do people judge the immorality of artificial intelligence versus humans committing moral wrongs in real-world situations? Comput Hum Behav Rep, 2022, 8: 100229
16 Xu L Y, Yu F, Peng K P. Algorithmic discrimination causes less desire for moral punishment than human discrimination (in Chinese). Acta Psychol Sin, 2022, 54: 1076–1092 [许丽颖, 喻丰, 彭凯平. 算法歧视比人类歧视引起更少道德惩罚欲. 心理学报, 2022, 54: 1076–1092]
17 Danaher J. Robots, law and the retribution gap. Ethics Inf Technol, 2016, 18: 299–309
18 Bonezzi A, Ostinelli M. Can algorithms legitimize discrimination? J Exp Psychol Appl, 2021, 27: 447–459
19 Lima G, Jeon C, Cha M, et al. Will punishing robots become imperative in the future? In: CHI '20: CHI Conference on Human Factors in Computing Systems. Honolulu: ACM, 2020. 1–8
20 Leo X, Huh Y E. Who gets the blame for service failures? Attribution of responsibility toward robot versus human service providers and service firms. Comput Hum Behav, 2020, 113: 106520
21 Shank D B, DeSanti A, Maninger T. When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. Inf Commun Soc, 2019, 22: 648–663
22 Maninger T, Shank D B. Perceptions of violations by artificial and human actors across moral foundations. Comput Hum Behav Rep, 2022, 5: 100154
23 Tolmeijer S, Christen M, Kandul S, et al. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In: CHI′22: CHI Conference on Human Factors in Computing Systems. New Orleans: ACM, 2022. 1–17
24 Schniter E, Shields T W, Sznycer D. Trust in humans and robots: Economically similar but emotionally different. J Econ Psychol, 2020, 78: 102253
25 Guidi S, Marchigiani E, Roncato S, et al. Human beings and robots: Are there any differences in the attribution of punishments for the same crimes? Behaviour Inf Tech, 2021, 40: 445–453
26 Chakroff A, Young L. How the mind matters for morality. AJOB Neurosci, 2015, 6: 43–48
27 Gray K, Waytz A, Young L. The moral dyad: A fundamental template unifying moral judgment. Psychol Inq, 2012, 23: 206–215
28 Gray H M, Gray K, Wegner D M. Dimensions of mind perception. Science, 2007, 315: 619
29 Gray K, Wegner D M. Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition, 2012, 125: 125–130
30 Malle B F. How many dimensions of mind perception really are there? In: Goel A K, Seifert C M, Freksa C, eds. Proceedings of the 41st Annual Meeting of the Cognitive Science Society. Montreal. 2019. 2268–2274
31 Weisman K, Dweck C S, Markman E M. Rethinking people's conceptions of mental life. Proc Natl Acad Sci USA, 2017, 114: 11374–11379
32 Bigman Y E, Waytz A, Alterovitz R, et al. Holding robots responsible: The elements of machine morality. Trends Cogn Sci, 2019, 23: 365–368
33 Masto M. Empathy and its role in morality. South J Philos, 2015, 53: 74–96
34 Molnar-Szakacs I. From actions to empathy and morality—A neural perspective. J Econ Behav Organ, 2011, 77: 76–85
35 Sullivan Y W, Fosso Wamba S. Moral judgments in the age of artificial intelligence. J Bus Ethics, 2022, 178: 917–943
36 Shank D B, North M, Arnold C, et al. Can mind perception explain virtuous character judgments of artificial intelligence? Tech Mind Behav, 2021, https://doi.org/10.1037/tmb0000047
37 Aggarwal P, McGill A L. Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. J Consum Res, 2007, 34: 468–479
38 Waytz A, Heafner J, Epley N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol, 2014, 52: 113–117
39 Laakasuo M, Palomäki J, Köbis N. Moral uncanny valley: A robot's appearance moderates how its decisions are judged. Int J Soc Robot, 2021, 13: 1679–1688
40 Yam K C, Goh E Y, Fehr R, et al. When your boss is a robot: Workers are more spiteful to robot supervisors that seem more human. J Exp Soc Psychol, 2022, 102: 104360
41 Nijssen S R R, Müller B C N, Bosse T, et al. Can you count on a calculator? The role of agency and affect in judgments of robots as moral agents. Hum Comput Interact, 2023, 38: 400–416
42 Nachmann K, Pillot B, Moore P, et al. Driving with robots: Mind perception and propensity for aggressive driving. Proc Hum Factors Ergon Soc Annu Meet, 2020, 64: 1965–1970

43 Kim S Y, Schmitt B H, Thalmann N M. Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Mark Lett, 2019, 30: 1–12

44 Kim S, McGill A L. Gaming with Mr. Slot or gaming the slot machine? Power, anthropomorphism, and risk perception. J Consum Res, 2011, 38: 94–107

45 Butt A H, Ahmad H, Shafique M N. AI-powered “voice recognition avatar”. Int J Gaming Comput Mediat Simul, 2021, 13: 1–17

46 Martin A E, Mason M F. Hey Siri, I love you: People feel more attached to gendered technology. J Exp Soc Psychol, 2023, 104: 104402

47 Go E, Sundar S S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Comput Hum Behav, 2019, 97: 304–316

48 Söderlund M, Oikarinen E L. Service encounters with virtual agents: An examination of perceived humanness as a source of customer satisfaction. Eur J Mark, 2021, 55: 94–121

49 Srinivasan R, Sarial-Abi G. When algorithms fail: Consumers’ responses to brand harm crises caused by algorithm errors. J Market, 2021, 85: 74–91

50 Burgoon J K, Newton D A, Walther J B, et al. Nonverbal expectancy violations and conversational involvement. J Nonverbal Behav, 1988, 13: 97–119

51 Hong J W, Cruz I, Williams D. AI, you can drive my car: How we evaluate human drivers vs. self-driving cars. Comput Hum Behav, 2021, 125: 106944

52 Liu P, Yang R, Xu Z. How safe is safe enough for self-driving vehicles? Risk Anal, 2019, 39: 315–325

53 Kocielnik R, Amershi S, Bennett P N. Will you accept an imperfect AI?: Exploring designs for adjusting end-user expectations of AI systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery Press, 2019. 1–14

54 Hong J W, Wang Y, Lanz P. Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. Int J Hum Comput Interact, 2020, 36: 1768–1774

55 Shank D B, DeSanti A. Attributions of morality and mind to artificial intelligence after real-world moral violations. Comput Hum Behav, 2018, 86: 401–411

56 Young A D, Monroe A E. Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. J Exp Soc Psychol, 2019, 85: 103870

57 Decety J, Cowell J M. Interpersonal harm aversion as a necessary foundation for morality: A developmental neuroscience perspective. Dev Psychopathol, 2018, 30: 153–164

58 Monroe A E, Guglielmo S, Malle B F. Morality goes beyond mind perception. Psychol Inq, 2012, 23: 179–184

59 Awad E, Levine S, Kleiman-Weiner M, et al. Drivers are blamed more than their automated cars when both make mistakes. Nat Hum Behav, 2020, 4: 134–143

60 Hage J. Theoretical foundations for the responsibility of autonomous agents. Artif Intell Law, 2017, 25: 255–271

61 Bastian B, Laham S M, Wilson S, et al. Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. Br J Soc Psychol, 2011, 50: 469–483

Summary for “人工智能决策的道德缺失效应及其机制”

# Reactions to immoral AI decisions: The moral deficit effect and its underlying mechanism

Xiaoyong Hu[1], Mufeng Li[1], Dixin Wang[1] & Feng Yu[2*]

[1] *Key Laboratory of Cognition and Personality (SWU), Ministry of Education, and Faculty of Psychology, Southwest University, Chongqing 400715, China;*
[2] *Department of Psychology, School of Philosophy, Wuhan University, Wuhan 430072, China*
* Corresponding author, E-mail: psychpedia@whu.edu.cn

Artificial intelligence (AI) is a branch of computer science in which systems are created with intelligence that enables them to perform cognitive functions, such as perception, reasoning, learning, and decision-making. AI has formidable problem-solving abilities in various domains, such as surveillance, health care, and finance. However, its societal applications raise ethical issues, such as gender and racial discrimination. Moreover, psychological research on AI ethics has revealed that people react less morally to unethical decisions made by AI than to those made by humans, showing deficiencies in moral evaluations, punishments, and behavioral responses to AI. These moral deficits of AI decisions have serious negative impacts on individuals, organizations, and society. For example, research has confirmed that this effect leads companies to use AI in their recruitment processes to discriminate against job applicants, reduces people’s awareness of unethical behavior, enables companies to evade accountability, exacerbates the challenges of advocacy and justice for affected groups, and even undermines societal moral standards along with the widespread use of AI. These negative impacts will damage trust in justice and fairness, causing lasting harm to societal ethics.

How can we explain AI moral deficit effects? We draw on the mind perception theory, which suggests that people’s moral reactions depend on how they perceive the mental attributes of an entity. These attributes can be divided into two dimensions: agency and experience. Agency refers to the belief that the entity can act, plan, self-regulate, remember, communicate, and think like a typical adult human. Experience refers to the belief that the entity can feel emotional states, such as hunger, fear, and pain. Mind perception is the process of attributing mental capabilities to an entity, while moral judgment is the process of evaluating the entity as good or bad. Entities with high agency are held responsible for their moral actions, and entities with high experience are expected to behave safely and ethically. People tend to view AI as having less agency and experience than adults do, which leads to the AI moral deficit effect. We explore the psychological reasons behind people’s lack of moral concern about immoral decisions made by AI. We contrast people’s moral responses to AI and human decision-makers and the serious social implications of this difference. Based on the theory of mind perception and moral dualism and supported by empirical evidence, we claim that two parallel factors underlie the moral deficit effect of AI decision-making: agency and experience. Additionally, we show that anthropomorphism and expectation violation can influence this effect. Thus, we propose a theoretical model of the inherent psychological mechanism of the moral deficit effect of AI decision-making from the perspective of a moral agent. This model extends the theory of mind perception and deepens the understanding of moral dualism.

In sum, we investigate how people react differently to biased AI and biased humans. Unlike other disciplines (e.g., computer science, philosophy, law, sociology) that focus on the design of fair algorithms in their “algorithmic ethics” research, we examine the human side of the problem. We contend that comprehending these differences is crucial for tackling the social challenges of biased AI and proposing a novel approach to building fair algorithms, as well as a fresh outlook on “algorithmic ethics,” which shifts the attention from algorithmic design to human responses. This approach is vital for ethics and for AI researchers to devise ethical frameworks and guidelines for AI systems. Furthermore, the findings can guide policy-making, legal frameworks and social interventions to reduce the adverse effects of AI and foster social equity and justice. However, due to the dynamic and fast-changing nature of AI, this field still needs additional research on the underlying mechanisms and intervention strategies involved.

**artificial intelligence, moral deficit effect, mind perception theory, anthropomorphism**

doi: 10.1360/TB-2023-1094