·计算建模与人工智能·

人工智能应用对人类道德的影响及认知神经机制*

郭雪俐[1] 苏 淞[1] 刘 超[2,3,4] 唐红红**[1]
（[1]北京师范大学经济与工商管理学院，北京，100875）
（[2]北京师范大学认知神经科学与学习国家重点实验室 IDG/ 麦戈文脑研究院，北京，100875）
（[3]北京师范大学脑与学习协同创新中心，北京，100875）
（[4]北京师范大学人工智能安全与超级对齐北京市重点实验室，北京，100875）

摘 要 随着人工智能的迅猛发展及其在众多领域的广泛应用，其对人类道德的影响引发了学界的深刻思考。现有文献表明，AI 应用可通过道德认知、道德情绪、社会评价与反应等心理和认知神经机制对人类道德产生影响。进一步，当前研究总结对比了 AI 在代理、建议、人机协作和道德模范四种应用情境下对人类道德行为产生的正面和负面影响。未来研究需要深入比较 AI 在不同应用情境下影响差异的机制，探究 AI 对人类道德的长期影响，并探索 AI 设计干预和提升道德影响的途径，为 AI 的设计、开发和应用提供理论支持和参考。

关键词 人工智能 人机交互 道德 社会脑

1 引言

人工智能（artificial intelligence, AI）产生至今，其伦理道德问题一直是各界关注的焦点。在科学家们的建构中，AI 会逐步具备感知、推理、学习、交互、问题解决、决策制定及创造力等认知功能（Maedche et al., 2019）。这些功能使得 AI 能够对人类的思维方式、互动过程和决策制定产生影响。当 AI 在社会行为、组织管理、社会生产等层面产生广泛影响时，便会产生如蝴蝶效应般的重大社会变革。对此，国际社会迅速做出反应。联合国教科文组织（UNESCO）于 2021 年通过《关于人工智能伦理的建议》，提出一系列原则以引导负责任的 AI 使用，防止其对个人和社会产生负面影响；世界经济论坛 2021 年发布的《全球风险报告》指出，AI 技术的普及加剧了决策、隐私保护、就业及社会公平性等方面的伦理风险，并在 2024 年报告中进一步细化了相关风险，包括虚假信息操纵、网络攻击、数据安全和算法歧视等。当我们面对这些风险和挑战时，首先需要回答一个关键问题：AI 应用会对人类道德产生什么样的影响？

道德是人类社会在长期发展中演化出的、被社会广泛接受的行为准则（Gert & Gert, 2002）。大量研究显示，道德可以从认知、情绪和社会评价三方面影响人类（Blasi, 1980; Haidt, 2007, 2008; Huebner et al., 2009）。其中，道德认知涉及对道德内容的理

* 本研究得到国家自然科学基金项目（71872016, 32441109, 32271092, 32130045）和中央高校基本科研业务费专项资金 (1233200014) 的资助。
** 通讯作者：唐红红，E-mail：tang.h@bnu.edu.cn
DOI:10.16719/j.cnki.1671-6981.20250601

解、判断等；道德情绪是人类加工道德信息时产生的情绪反应，如内疚、羞愧等；社会评价则反映社会对道德表现的反馈，决定道德形成和发展（Haidt, 2007; Turiel, 2002）。随着技术发展，AI 已从执行特定任务的系统演变为具备学习能力的智能主体，并融入人类社会生活的多个领域。AI 与人类的关系已逐渐从过往机器被动接受人类指令的模式发展为人类与 AI 交互共生的模式（Heyder et al., 2023）。近年来，AI 技术快速发展，生成式 AI（generative artificial intelligence, GenAI）能基于现有数据快速生成多种形式新内容，特别是大语言模型（large language model, LLM）可执行问题回答、聊天互动等多种自然语言任务。这些技术推动了 AI 的应用和发展，催生了人机混合工作（human-machine hybrid work）这一新型工作模式。同时，Replika、ChatGPT、DeepSeek 等聊天机器人通过提供陪伴、建议和决策支持，逐步与人类建立起富含情感的人机关系。这些变化使得 AI 逐步走向社会化，并能够在与人的社会互动中从认知、情绪、社会评价与反应三个方面对人类道德产生影响（图 1）。为了深入理解这些影响，本文以 AI 应用与人类交互为切入点，回顾并整合相关文献，探讨 AI 对人类道德的影响及其认知神经机制，结合不同应用情境分析 AI 的积极和消极影响，为未来研究探讨如何减少消极影响、提升积极影响提供参考。

2 AI 应用在交互中影响人类道德的内在机制

早期“计算机作为社会参与者”（computers are social actors, CASA）的研究发现，人们会将社交规则、规范和期望应用于和计算机的交互中（Nass &

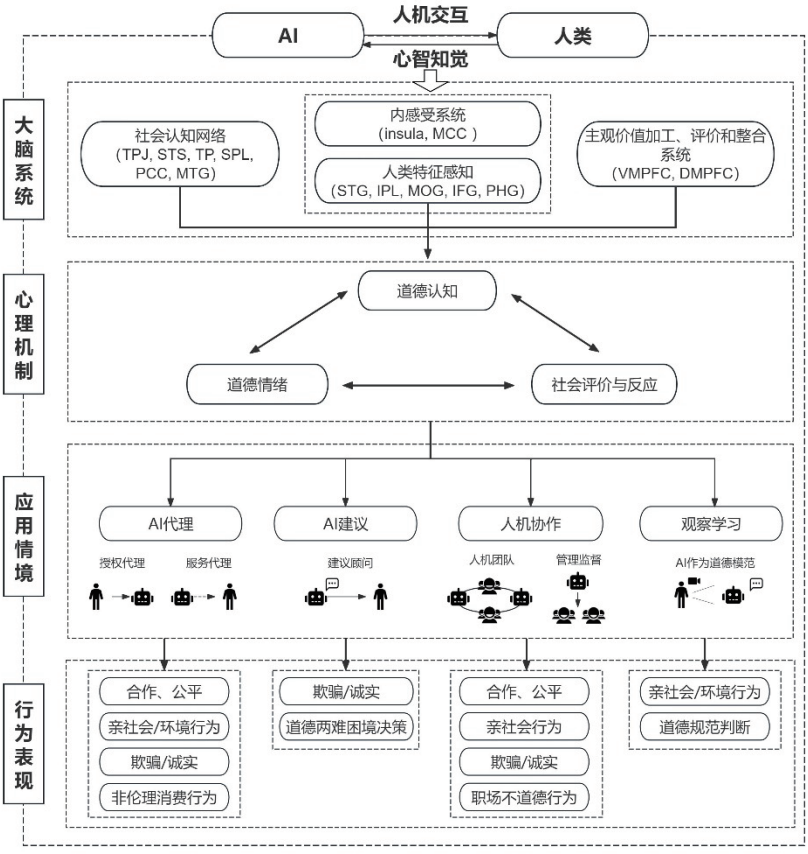

图 1 AI 应用在交互中对人类道德的影响

注：Insula，脑岛；MCC，middle cingulate cortex, 中部扣带回 ; STG，superior temporal gyrus, 颞上回；IPL，inferior parietal lobule, 顶下小叶；MOG, middle occipital gyrus, 枕中回；IFG, inferior frontal gyrus, 额下回；PHG，parahippocampal gyrus, 海马旁回；TPJ，temporo-parietal junction, 颞顶联合区；STS，superior temporal sulcus, 颞上沟；TP，temporal pole, 颞极；SPL，superior parietal lobule, 顶上小叶；PCC，posterior cingulate cortex, 后扣带回；MTG，middle temporal gyrus, 颞中回；VMPFC，ventral medial prefrontal cortex, 腹内侧前额叶皮层；DMPFC，dorsal medial prefrontal cortex，背内侧前额叶皮层。

Moon, 2000）。随后的研究表明，AI 也会受到社会化对待。例如，人们在与 AI 的互动中会应用礼貌等社会规范，通过温暖（warmth）和能力（competence）两个关键社会认知维度来感知和评价 AI（McKee et al., 2023）。同时，AI 的拟人化交互功能不断增强，Replika、ChatGPT、DeepSeek 等对话系统能够以具有社交性和同理心的方式回应用户，提供陪伴和情感支持，满足人类多样的社交需求（Li & Zhang, 2024; Yin et al., 2024）。这使得人与 AI 之间的交互（人 -AI 交互）能够以高度类似人际交互（人 - 人交互）的形式进行。

然而，人 -AI 交互至今远未达到与人 - 人交互相同的水平。其根本原因之一在于人们对 AI 和真人的心智知觉（mind perception）存在显著差异。心智知觉是指人们对某一实体是否具有心智能力的感知，主要取决于该实体被感知到的能动性（agency）和体验性（experience）程度。能动性是指实体是否具备像正常成年人一样自主计划、行动和思考的能力，而体验性则指实体是否能够感受情绪和身体感觉（Gray et al., 2007）。研究发现，人们倾向于认为 AI 具有中等的能动性但缺乏体验性，即认为 AI 具有一定的行动和思考能力，但无法感受或理解情绪和身体感觉（Gray et al., 2007）。即使 AI 能够通过语言分析准确理解并回应人类情绪，但只要告知人们回应者为 AI，人们对其情感理解能力的评价就会显著降低（Yin et al., 2024）。而当描述 AI 具有体验性时，会让人们感到紧张不安（Gray & Wegner, 2012）。此外，人们普遍不认为 AI 具备体验道德对错的能力（Gray et al., 2007），对 AI 参与道德决策表现出排斥和厌恶（Bigman & Gray, 2018）。

这些结果显示，人们倾向于认为 AI 相较于人类具有更低的心智水平。这种对 AI 较低的心智知觉，会直接影响人们在人 -AI 交互中的道德认知、道德情绪以及社会评价和反应。在道德认知层面，人们可能不会以对待人类的相同道德标准去评判 AI 的行为；在道德情绪方面，人们在与 AI 互动时可能会体验到更少或更弱的道德情绪；而在社会评价和反应方面，人们在人 -AI 交互场景中可能更容易忽视道德准则和社会评价的约束。如图 1 所示，AI 通过对这三类内在机制的影响，能够在各种情境下以不同角色与人类进行交互，进而影响人类的具体行为表现。

2.1 人 –AI 交互中的道德认知

人 -AI 交互可能改变人类的道德认知，首先体现在对道德的态度、推理和判断上，尤其是在责任归属问题上。研究发现，当 AI 造成伤害时，人们倾向于认为 AI 比人类更应该被责备。例如，在自动驾驶的道德两难中（伤害行人或乘客），驾驶者常将不利后果归责于自动驾驶系统而非自己（Gill, 2020）。相比同等程度的人为事故，人们认为 AI 导致的事故更严重、更不可接受，需承担更多责任（Liu & Du, 2022）。在 AI 违反公平、自由或纯洁道德基础的事件（如种族歧视、隐私泄露）中，人们更倾向于指责 AI，而非其背后的人类组织或开发者（Shank & DeSanti, 2018）。在人 -AI 共同决策导致道德违规时，人们认为其自身责任较单独决策时更少，而 AI 需分担部分责任（Shank et al., 2019）。这表明 AI 在很多情境下能够成为人类不道德行为的“替罪羊”，方便人类推卸责任（Hohensinn et al., 2024），降低人类对不道德行为结果的感知，进而加剧不道德行为。

道德认知改变还体现在人们对 AI 决策的功利主义道德判断偏向上。道德判断的双加工理论（dual-process theory）显示，人类道德判断主要由两大道德原则驱动：道义论（deontology）和功利主义（utilitarianism）。其中，道义论关注行为本身的道德正确性，而功利主义则关注行为结果的利益最大化（Greene et al., 2004; Greene et al., 2001）。面对会产生伤害的道德两难问题时，若情绪反应强烈或时间紧迫，人们倾向于产生道义论判断，认为伤害行为不可接受；若充分分析行为结果，则功利主义判断占主导，即当伤害行为带来总体有益结果时可接受，否则不可接受。由于 AI 决策依赖算法，人们通常认为其是通过精密计算追求利益最大化的系统，不会像人类一样受情绪驱动，根据行为本身的道德正确性做出判断（Dietvorst & Bartels, 2022）。这导致人们认为 AI 在涉及道德的决策与判断中会更关注功利结果，而忽略道义正确性（徐岚等 , 2024）。这一倾向在针对在线捐赠行为的研究中得到支持：与真人客服相比，AI 客服更显著地激活了人们的功

利主义道德判断，从而削弱了人们的捐赠意愿和行为（Zhou et al., 2022）。

2.2 人 –AI 交互中的道德情绪

人 -AI 交互所引发的人类道德情绪变化，首先表现为人们因伤害或帮助 AI 而产生的道德情绪较弱。根据道德二元论（the theory of dyadic morality），道德判断的核心是对伤害的感知，即一个有意图的道德行为者（moral agent）给一个道德受害者（moral patient）带来的痛苦（Gray et al., 2012a）。对道德行为者的判断主要基于能动性，即其是否具备自主意图和行动能力；而对道德受害者的判断则侧重于体验性，即是否能够感受情感和身体的痛苦。当某一实体被认为具有体验性时，便具备成为道德受害者的资格，对其造成伤害被视为道德错误（Gray et al., 2012b）。而 AI 通常被认为缺乏体验性，因此人们往往否认其作为道德受害者的资格。这导致人们在对 AI 实施与人类相同的不道德行为时，所引发的负面情绪（如内疚）相对较弱。例如，在独裁者博弈（dictator game）和公共物品博弈（public good game）中，人们因不公平对待 AI（vs. 人类）产生的内疚情绪更弱（Melo et al., 2016）。此外，人们认为欺骗 AI（vs. 人类）的内疚感较低，对 AI 的不诚实行为在道德上更能被接受（Petisca et al., 2020），这增强了人们在与 AI 交互时为了经济利益而谎报、瞒报信息等不道德行为倾向（Giroux et al., 2022; Kim et al., 2023）。相应地，人们因为帮助 AI（vs. 人类）感受到的积极情绪（如自豪）也更弱（Schniter et al., 2020），人们也不会对 AI（vs. 人类）遭受的痛苦产生同等程度的共情（Cross et al., 2019; Martin et al., 2020）。因此，人类不会在情感上像对待其他人类一样对待 AI，从而削弱了其在人 -AI 交互中的亲社会动机和行为。

当 AI 作为行为发起者参与亲社会情境时，人们所感受到的道德情绪也较弱。一方面，人们认为 AI 无法感知人类的痛苦，对 AI 参与亲社会行为产生的情绪反应较弱。例如，当消费者意识到慈善广告中使用了 AI 生成的面孔图片而非真实图片时，其共情关注（empathic concern）减弱，进而降低悲伤情绪感知和预期内疚，最终导致捐赠意愿降低（Arango et al., 2023）。另一方面，人们认为 AI 行为相较于正常成年人缺乏自主性，且因其机械属性更不易受到外界伤害，在应对灾难等危险情境时面临的风险和付出的努力都更少。这些认知削弱了 AI 的“英雄行为”（如在火灾、飓风等灾难中参与救援）所激发的勇气与鼓舞，使其无法像真实人类的行为一样促进人们的捐赠意愿和行为（Chen & Huang, 2023）。

2.3 人 –AI 交互中的社会评价与反应

道德的社会信号（social signaling）解释认为，人们从事道德行为的动机是为了向他人证明自己是有道德的（Barclay & Willer, 2007）。然而，大多数道德行为都需要人们牺牲一定的自我利益，这使得人们会倾向于表现出自己具有道德但又希望避免实际行为的代价。因此，人们可能会在态度和行为上表现出道德，但并非真正的道德（Batson et al., 1997）。人们通过表现道德以维护和提升自己的社会形象，保护道德自我概念的完整性（Batson et al., 1997; Batson et al., 2002）。仅感知到他人的存在便会增强人们对自身社会形象的关注，进而影响其道德认知和行为。例如，在有他人关注时，人们倾向于表现出更多的诚实和公正（Tang et al., 2017）。因此，在社交环境中，个体更倾向于表现出道德动机和行为以维护社会形象。

人 -AI 交互可能降低人们对自身社会形象的关注。首先，人们认为 AI 缺乏理解社交情境中行为和意图的能力，从而不具备进行社会判断的能力，这降低了人们对 AI 会产生负面评价的担忧（Holthöwer & Van Doorn, 2023）。其次，人们通常不会将 AI 视为内群体成员（Złotowski et al., 2017），也不会将获得 AI 的认可视作正向回报（Nash et al., 2018）。在人 -AI 互动中，人们较少感受到社会压力，也不太在意是否给 AI 留下好印象，从而忽视对自身社会形象的管理，减少对行为是否符合社会规范或维护声誉的考虑。这可能会促使人们表现出更多欺骗等不道德行为（Cohn et al., 2022），并降低其亲社会和亲环境行为（Wang et al., 2023）。

3 AI 应用影响人类道德的认知神经机制

认知神经科学领域的研究者们常将 AI 与人类对比，以探究人类心理和行为机制。近些年的研究发现，

人 -AI 交互与人 - 人交互在神经机制上的差异主要体现在社会脑网络中（Harris, 2024）。相较于人类交互，人们与 AI 交互时这些脑区的活动呈现显著变化，且主要涉及三个相关脑网络（图 1）。

3.1 内感受和人类特征感知系统

内感受和人类特征感知系统在面对 AI 和人类时的差异，构成了人们对 AI 心智知觉较低的神经基础。内感受系统负责人们对内部身体信号的感知和加工，以脑岛（insula）为关键脑区，包括中部扣带回（middle cingulate cortex, MCC）等（Craig, 2009; Critchley & Garfinkel, 2017），与个体的多种情绪感知密切相关。当人们面对 AI（vs. 人类）时，这两个区域的活动显著减弱（Cheetham et al., 2011; Di Cesare et al., 2016）。脑岛是对人类基本社会互动中的活力形式（vitality forms）进行感知加工的核心大脑区域（Di Cesare et al., 2016）。根据 Stern 的理论，活力形式指个体在交流中表现出的行为特征（如粗鲁或温柔），能够反映其在社会互动中的情绪或情感状态（Stern, 2010）。人 -AI 交互中脑岛活动的减弱，表明个体对 AI 的活力形式感知减弱，不会像对待人类一样对 AI 的内在状态进行推断和理解。此外，脑岛和 MCC 还参与共情反应，特别是疼痛感知（Lamm et al., 2011）。当人们在观察 AI（vs. 人类）的痛苦或愉悦经历时，脑岛和 MCC 均未出现相应激活，且在持续的社交互动后仍无显著变化（Cross et al., 2019）。这些研究结果为人类对 AI 缺乏体验性的认知提供了神经机制上的依据，从而为人 -AI 交互中个体情绪反应，特别是道德情绪减弱提供了支持。

此外，在参与人类特征感知相关的脑区中，颞上回（superior temporal gyrus, STG）、顶下小叶（inferior parietal lobule, IPL）、枕中回（middle occipital gyrus, MOG）和额下回（inferior frontal gyrus, IFG） 在不同研究中表现出不一致的激活模式：部分研究中面对 AI 时激活更强，而部分研究中则在面对人类时激活更强（Delgado et al., 2008; Gobbini et al., 2011; Hogenhuis & Hortensius, 2022）。此外，海马旁回（parahippocampal gyrus, PHG）在面对 AI 时的活动显著高于面对人类时（Delgado et al., 2008; Gobbini et al., 2011）。这些不一致可能源于人们对 AI 的非语言表达（如外形、动作、姿势）是否会对自己产生积极或消极影响的判断。例如，高度拟人化的机器人会引发恐怖谷效应（uncanny valley），而适度拟人化相较于机械化外观则更能提高人们的喜欢程度（Rosenthal-Von der Pütten et al., 2019），增强人们的心理亲近感和接近意愿（Baek et al., 2022）。由于缺乏这些脑区与人类特征具体维度相关的证据，未来研究需要进一步整合这些不一致的结果。

3.2 社会认知网络

大脑社会认知网络中的关键区域，如颞顶联合区（temporal-parietal junction, TPJ）和颞上沟（superior temporal sulcus, STS）是参与个体在社会互动中对互动对象心理状态（如意图、动机或情感）推断的核心区域（Saxe & Kanwisher, 2003）。已有研究显示，无论是在被动观察人 -AI 交互及 AI 行为（Wang & Quadflieg, 2015），还是在与 AI 的真实互动（Chaminade et al., 2015）中，TPJ 和 STS 的活动在面对 AI（vs. 人类）时都表现出显著减弱。这一现象在道德相关任务（如合作、公平、信任、互惠等决策）和非道德相关任务（如注视线索提示相关的注意任务）中均稳定存在。TPJ 和 STS 活动减弱，为人们对 AI 的低心智知觉，甚至是去心智化，提供了直接的认知神经证据。

在与 AI 的真实互动中，其他社会认知相关脑区的活动也显著变弱，如颞极（temporal pole, TP）、顶上小叶（superior parietal lobule, SPL）、后扣带回（posterior cingulate cortex, PCC）、颞中回（middle temporal gyrus, MTG） 等（Anders et al., 2015; Hogenhuis & Hortensius, 2022; Schindler et al., 2019）。这些社会认知相关脑区活动的减弱进一步显示出人们在互动中未对 AI 的行为意图和内心状态进行如人类一样的心理模型构建，反映出人们对 AI 的心智化程度相对较低。此外，社会认知网络活动的减少也意味着对社会规范的理解和关注相应降低（Mars et al., 2012），为人 -AI 互动中道德认知的改变、对社会评价的关注和反应降低提供了认知神经机制基础。

3.3 主观价值加工、评价与整合系统

第三个大脑网络是以内侧前额叶皮层（medial prefrontal cortex, MPFC）为核心的主观价值加工、评价与整合系统。MPFC 在很多研究中又被分为

腹内侧和背内侧两个亚区（ventral and dorsal medial prefrontal cortex, VMPFC; dorsal medial prefortal coxtex, DMPF）。其中，VMPFC主要加工决策中的选项或信息的主观价值，如奖赏和预期价值（Grabenhorst & Rolls, 2011）；DMPC更多参与社会评价和心理状态相关的价值表征及认知整合（Dixon et al., 2017）。多数研究显示，在与AI互动的决策中，VMPFC和DMPC的活动显著减弱（Hogenhuis & Hortensius, 2022; Schindler et al., 2019），表明人们对决策信息的价值评估不如与人类互动时深入。

这种减弱为人在与AI互动时道德决策的变化提供了认知神经基础。一方面，MPFC在社会决策中发挥着重要作用，包括评估社会风险和收益（Dixon et al., 2017; Grabenhorst & Rolls, 2011）。MPFC活动减弱反映了在人-AI交互中，人们简化了决策的价值评估，复杂决策情境中对多方利益和社会规范的考虑减少，更多关注自我利益，从而降低人们对不道德行为结果的感知。例如，在独裁者博弈、信任游戏（trust game）、最后通牒博弈（ultimatum game）及公共物品博弈等道德相关的决策中，个体与AI互动时的合作意愿和公平行为均低于与人类互动时（Karpus et al., 2021; Melo et al., 2016）。另一方面，MPFC与道德判断双加工理论中的道义论判断过程紧密相关（Greene, 2007; Greene et al., 2004; Greene et al., 2001）。因此，MPFC活动的减弱可能促使人们更多关注功利主义倾向的信息和过程，通过改变道德认知，减弱道德情绪体验以及对社会评价的感知和反应。

然而，Wang和Quadflieg（2015）发现，人们在判断和评价AI帮助人类行为时，MPFC的活动增强。这可能是由于该研究聚焦于AI作为帮助行为主体的情境，使得人们在评价过程中的反应更强烈。由于相关研究较少，AI作为行为对象或主体时价值评估

表1 AI影响个体道德行为的相关研究

AI应用情境	文献数量	人类对比	结果	机制类别	具体测量机制	作者（年份）
AI代理	13	是	公平(+)/捐赠(-)/欺骗(+)/代理偏好[a]	认知	责任分布/功利主义道德判断(+)	deMelo et al. (2018); Zhou et al. (2022); Mell et al. (2020)*; Feier et al. (2022); 徐岚等(2024)
		是	合作(+)/非伦理消费行为(+)	情绪	负面情绪（恐惧）/预期内疚(-)	Fernández Domingos et al. (2022)*; Giroux et al. (2022); Kim et al. (2023); Liu et al. (2023)
		是	绿色消费(-)	社会评价与反应	社会评价(-)/印象管理(-)	Wang et al. (2023); 聂春燕和汪涛(2025)
		无	捐赠(+)	认知	自我效能感(+)	Namkoong et al. (2023)
		无	捐赠(+)	社会评价与反应	心理亲近感(+)	Baek et al. (2022)
AI建议	5	是	欺骗(+)/诚实(-)	认知	正当化(+)/责任分担/责任扩散/算法欣赏	Leib et al. (2024); Kim et al. (2024)*; Krügel et al. (2023a)*; Krügel et al. (2022)*;
		无	道德两难困境中对AI建议的遵从(+)	认知	算法可解释性/算法欣赏	Krügel et al. (2023b)*
人机协作	9	是	合作(-)/公平(-)/帮助(-)/不道德指令遵从(-)	认知	责任分担(+)/心智知觉(-)/物化他人(+)	Krügel et al. (2023a); Karpus et al. (2021); Granulo et al. (2024); Lanz et al. (2024)
		是	合作(-)/公平(-)/职场不文明行为(+)	情绪	内疚感(-)/职业倦怠(+)	Melo et al. (2016); Karpus et al. (2021); Yam et al. (2023);
		是	欺骗(+)	社会评价与反应	声誉/社会形象关注(-)	Cohn et al. (2022)
		无	帮助(+)/亲社会行为(+)	社会评价与反应	归属感寻求(+)/团队认同(+)	Tang et al. (2023); Gorny et al. (2023)*
道德模范	7	是	环保行为(-)	认知	感知真实性(-)	Yang et al. (2023)
		是	捐赠(-)	情绪	勇气(-)/情感体验(-)	Chen, Huang (2023); Arango et al. (2023)
		无	道德规范判断/亲社会行为(+)	认知	规范性认知(+)/感知真实性(+)	Song-Nichols et al. (2020); Williams et al. (2018); Peter et al. (2021); Quach et al. (2024)

注：（+）表示积极影响，（–）表示消极影响；* 表示仅讨论了该机制，但未进行实证检验；[a] 代理偏好指在特定情境下，人们对于选择由AI还是人类代理为其工作或服务的偏好。

的差异如何影响道德认知和行为仍需进一步探讨。

4 AI 在不同应用情境中对人类道德影响的行为表现

在实际应用中，根据 AI 扮演角色的不同，AI 可以从多个维度影响人类道德。根据 Kobis 等（2021）对 AI 区分的四种角色即代理（delegate）、顾问（advisor）、合作伙伴（partner）和榜样（role model），人们可以让 AI 代表自己从事道德 / 不道德行为、根据 AI 提供的道德 / 不道德建议进行行为、将 AI 视为共同承担道德 / 不道德行为后果和责任的“同伙”、将 AI 作为道德榜样改变人们对道德规范的认知，并引发人类（特别是儿童）对相关行为的模仿。如表 1 所示，尽管 AI 在四种角色中既可能促进道德认知和行为，也可能引发不道德认知和行为，但现有研究多显示其对道德认知和行为的负面影响。在 AI 作为代理时，人 -AI 交互通过影响道德认知、情绪、社会评价和反应对行为产生影响；当 AI 作为顾问提出建议时，主要通过改变道德认知来促进欺骗行为；在人机协作中，根据协作时间长短和利益相关性，通过认知、情绪、社会评价与反应对道德行为产生双向影响; 而当 AI 作为道德模范时，主要通过认知和情绪路径对道德行为产生影响。

4.1 AI 代理

作为人类代理执行任务，是 AI 的主要应用之一。当人类授权 AI 代理自己完成任务时，人们对自己有利的不道德行为会增加（Kobis et al., 2021）。例如，在谈判中，相较于自己直接参与，人们使用 AI 代理时更倾向于采用欺骗和情感操控策略（Mell et al., 2020）。这是因为授权 AI 增加了行为的匿名性，拉大了与受害者的心理距离，降低了责任感和内疚感（Bonnefon et al., 2024; Kobis et al., 2021）。在接受 AI 服务时，人们因预期内疚感和社会形象关注的降低及功利主义道德判断的增加，更容易出现非伦理行为，如谎报退货原因、隐瞒收银员错误以获取利益（Kim et al., 2023; Liu et al., 2023）、减少对公益项目的捐赠（Zhou et al., 2022）和对绿色产品的购买意愿（聂春燕，汪涛 , 2025; Wang et al., 2023）。此外，为了避免行为的负面后果，人们更愿意让 AI 代理执行存在道德风险的任务（Feier et al., 2022）。

少数研究发现，人们也会抵制 AI 代理可能的不道德行为，从而对整体环境产生积极影响。例如，在公共品博弈中，人们更倾向于让 AI 做出群体利益最大化的决策，抑制群体因担忧背叛而减少贡献的行为，从而促进合作（Fernández Domingos et al., 2022）。当 AI 代表人类作为主体进行类似最后通牒游戏的利益分配时，人们会更抵制不公平的分配和出价，更不愿意与不公平的对手达成协议（de Melo et al., 2018）。这表明，人们在 AI 代理情境下会根据行为结果对自身利益的影响做出道德选择：当不道德行为更有利时，AI 代理会增加不道德行为；而当道德行为更有利时，AI 代理则会促进道德行为。

4.2 AI 建议促进不道德行为

AI 建议能够以类似人类建议的方式增加个体的不道德行为。例如，Leib 等（2024）通过大规模实验发现，无论是来自人类还是 AI 的不诚实建议，都显著增加了个体为追求个人利益而进行欺骗行为，且披露建议来源并未减少人们对 AI 建议的接受度。无论参与者是否知道建议来自 AI，AI 鼓励欺骗的建议均显著提高了欺骗行为的发生率。ChatGPT 等 AI 应用会在对话中生成支持功利主义的道德建议（Krügel et al., 2023b）。然而，当 AI 鼓励道德行为时，人们常拒绝其建议。在人机协作中，AI 对不道德行为的纠正建议通常不被采纳（Krügel et al., 2023a）。在以自我报告任务表现获取报酬的场景中，AI 和人类提供相同建议鼓励诚实行为时，人类建议能减少 38% 的欺骗行为，而 AI 建议未显著减少欺骗(Kim et al., 2024)。道德准则作为社会规范的核心，通常源于特定群体或社会所倡导且被其成员接受的行为规范体系（Gert & Gert, 2002）。人们在接收到不道德行为建议（如欺骗）时，可能会感知到对这种不道德行为的默许和认可，从而在不破坏道德自我形象的同时增强对自我利益的寻求，降低对社会评价的考虑（Tang et al., 2017）。接受不道德建议还会引发责任分散，让人们产生与建议者共同承担责任的感觉。因此，人们可能更在乎这种默许和认可本身，而不会特别关注其来源是人类还是 AI（Leib et al., 2024）。然而，当人们接收到道德行为建议（如诚实）时，建议来源则带来不同影响：人类建议往往能增强个体的社会评价和形象感知，对行为产生

积极影响；而 AI 建议则较难引发同样的效应（Kim et al., 2024）。

4.3 人机协作中的道德影响

人机协作日益普遍，人机混合工作因其在提升生产效率、保障安全与促进创新等方面的重要作用受到广泛关注。一方面，当不道德行为对个体更有利时，人机协作会增加其不道德行为。例如，当人们发现 AI 同伴的不道德行为时，往往选择“同流合污”而非纠正（Krügel et al., 2023a）。AI 同事的存在会引发职业倦怠，间接导致员工贬低同事等不文明行为（Yam et al., 2023）。当 AI 作为管理者时，个体更可能夸大工作表现以获取经济利益（Cohn et al., 2022）。使用 AI 算法管理还会增加员工物化同事的倾向，降低帮助同事的意愿，在人机共同管理时这种负面影响也依然存在（Granulo et al., 2024）。

另一方面，与 AI 协作也可能增强人们对其他人类的亲社会行为。例如，频繁与 AI 互动会增强人们的社交和归属感需求，使其更愿意帮助同事（Tang et al., 2023）。在混合人类 - 机器人团队中，人类工作者比纯人类团队更倾向于表现出亲社会行为（Gorny et al., 2023）。在道德决策中，人们拒绝 AI 主管不道德指令的程度显著高于人类主管，体现出对 AI 互动不利结果的抵制（Lanz et al., 2024）。因此，在人机协作中，AI 既可能带来负向影响促使人类产生不道德行为，也可能产生正向影响激发人类的亲社会行为，其影响方向取决于该行为本身是否能给人类带来实际利益或心理满足。当 AI 被感知为高能力的竞争威胁或压力来源时，可能会促进人类通过增加不道德行为、减少亲社会行为来应对这种威胁或压力，以维护自身地位和利益。相反，当 AI 被视作提供情感和社会支持的温暖同伴，特别是当人们认为 AI 与自身利益一致时（McKee et al., 2023），可能会促进亲社会行为。这种效应在长期参与人机协作的个体中可能表现得更为显著。长时间与 AI 互动可能会削弱人们从真实的人际互动中获取的情感支持和社会联系，促使人们通过增加对人类同事的亲社会行为来进行补偿。

4.4 AI 作为道德模范

人类很多道德认知和行为源于观察学习。AI 作为道德模范理论上可有效促进道德行为，但在成年人中的效果不佳。例如，AI 的亲社会行为示范在激发成年消费者捐赠行为方面效果较差（Chen & Huang, 2023），且企业通过虚拟影响者传达社会责任信息时，公众认为其信息可信度较低，公众参与意愿也更低（Yang et al., 2023）。有限的几项研究表明，AI 作为“道德榜样”对儿童的影响可能更大。例如，儿童对 AI 行为表现出更显著的遵从倾向（Vollmer et al., 2018）；智能语音玩具也能显著影响儿童对道德越界和刻板印象的认知（Song-Nichols & Young, 2020; Williams et al., 2018）；观察到机器人积极示范的儿童在后续任务中更倾向于表现出亲社会行为（Peter et al., 2021）。

5 总结与未来研究方向

本文从道德认知、道德情绪、社会评价和反应三个方面探讨了 AI 应用对人类道德的影响，并分析了其认知神经机制及在不同应用情境中引起的行为表现（图 1）。但该领域还有很多问题没有解决，有待进一步研究和深入探讨。

首先，现有研究表明，AI 对个体不道德或道德行为的促进作用与其行为对个体利益的正负面影响及个体对 AI 的温暖和能力感知密切相关。然而，这种关系是否具有因果性以及在个体需同时考虑群体利益时是否会改变，仍需进一步研究探索。未来研究可以借助认知神经科学的方法，如利用 fMRI 等技术观察在不同的人 -AI 互动情境中，个体与意图推测、情绪、社会感知等相关脑区的活动差异以及与自我利益和道德行为之间的冲突处理相关大脑区域活动的变化，帮助厘清在各种应用情境中人们的道德认知、情绪和社会评价与反应发挥作用的先后顺序和边界条件，从而有针对性地制定干预措施。

其次，已有研究多聚焦 AI 对人类即时道德判断及行为的影响，而对其长期效应的探讨较少。随着人 -AI 互动日益频繁，特别是 AI 在儿童认知与发展阶段的重要性日益凸显，亟需深入探究其对个体道德认知与情绪体验的长期影响。此外，AI 的情绪识别、感知及表达能力的不断提升模糊了人与 AI 之间的界限，可能导致个体对他人和自己的物化和去人性化（Dang & Liu, 2025; Kim & McGill, 2024）。AI

决策的不公平性还会引发道德冷漠，降低个体在后续决策中的亲社会行为（Zhang et al., 2025）。另一些研究则显示出 AI 在倾听人类情绪，通过对话调节和改善人类负面情绪方面的巨大潜力（Sabour et al., 2023; Yin et al., 2024）。长期人 -AI 互动究竟会强化物化与去人性化倾向导致冷漠自私及不道德行为，还是通过情绪调控促进人们之间的信任与互惠、增强道德认知与行为，仍需更多的纵向研究进行检验。

最后，未来研究可通过 AI 的道德训练和外显设计优化，提升其对人类道德的正面影响，降低负面影响。一方面，AI 处于不断学习和进化中，通过训练使其具备与人类相似的道德判断、情绪感知和决策能力，能够有效提升 AI 作为道德模范的能力，从而提升和干预 AI 对人类的道德影响。另一方面，现有研究将 AI 的外显设计（如拟人化）作为干预其道德影响的有效机制。例如，在募捐情境中，赋予 AI 适度的拟人化特征，如社会性应答行为（拥抱、微笑和眼神交流等）和恰当的表达方式（如奉承和感激），可以有效促进人们的捐赠意愿和捐赠金额（Baek et al., 2022; Namkoong et al., 2024; Oliveira et al., 2021; Quach et al., 2024）。然而，最小程度的拟人化无法产生作用，过度拟人化则可能引发恐怖谷效应（Rosenthal-Von der Pütten et al., 2019）、造成对人类的去人性化等反向效果（Kim & McGill, 2024）。因此，未来研究应进一步探索 AI 外显设计的优化策略，以减轻其负面影响、提升其正面影响，同时考虑个体专业知识、特质及文化背景的调节作用。

参考文献

聂春燕 , 汪涛 . (2025). AI 让人更环保？ AI 推荐对消费者的绿色产品购买意愿的影响 . *南开管理评论 , 28*(3), 51–62.

徐岚 , 陈全 , 崔楠 , 辜红 . (2024). “共赢” vs. “牺牲”：道德消费叙述框架对消费者算法推荐信任的影响 . *心理学报 , 56*(2), 179–193.

Anders, S., Heussen, Y., Sprenger, A., Haynes, J. D., & Ethofer, T. (2015). Social gating of sensory information during ongoing communication. *NeuroImage, 104,* 189–198.

Arango, L., Singaraju, S. P., & Niininen, O. (2023). Consumer responses to AI–generated charitable giving ads. *Journal of Advertising, 52*(4), 486–503.

Baek, T. H., Bakpayev, M., Yoon, S., & Kim, S. (2022). Smiling AI agents: How anthropomorphism and broad smiles increase charitable giving. *International Journal of Advertising, 4*1(5), 850–867.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences, 274*(1610), 749–753.

Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335–1348.

Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology, 83*(2), 330–339.

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

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88*(1), 1–45.

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

Chaminade, T., Da Fonseca, D., Rosset, D., Cheng, G., & Deruelle, C. (2015). Atypical modulation of hypothalamic activity by social context in ASD. *Research in Autism Spectrum Disorders, 10,* 41–50.

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5,* 126.

Chen, F., & Huang, S. C. (2023). Robots or humans for disaster response? Impact on consumer prosociality and possible explanations. *Journal of Consumer Psychology, 33*(2), 432–440.

Cohn, A., Gesche, T., & Maréchal, M. A. (2022). Honesty in the digital age. *Management Science, 68*(2), 827–845.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*(1), 59–70.

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current Opinion in Psychology, 17,* 7–14.

Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1771), 20180034.

Dang, J., & Liu, L. (2025). Dehumanization risks associated with artificial intelligence use. *American Psychologist.* Advance online publication.

de Melo, C. M., Marsella, S., & Gratch, J. (2018). Social decisions and fairness change when people' s interests are represented by autonomous agents. *Autonomous Agents and Multi–Agent Systems, 32,* 163–187.

Delgado, M. R., Schotter, A., Ozbay, E. Y., & Phelps, E. A. (2008). Understanding overbidding: Using the neural circuitry of reward to design economic auctions. *Science, 321*(5897), 1849–1852.

Di Cesare, G., Fasano, F., Errante, A., Marchi, M., & Rizzolatti, G. (2016). Understanding the internal states of others by listening to action verbs. *Neuropsychologia, 89,* 172–179.

Dietvorst, B. J., & Bartels, D. M. (2022). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology, 32*(3), 406–424.

Dixon, M. L., Thiruchselvam, R., Todd, R., & Christoff, K. (2017). Emotion and the prefrontal cortex: An integrative review. *Psychological Bulletin, 143*(10), 1033.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: Artificial agents

may help to evade punishment. *Science and Engineering Ethics, 28*(2), 19.

Fernández Domingos, E., Terrucha, I., Suchon, R., Grujić, J., Burguillo, J. C., Santos, F. C., & Lenaerts, T. (2022). Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Scientific Reports, 12*(1), 8492.

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research, 47*(2), 272-291.

Giroux, M., Kim, J., Lee, J. C., & Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and AI. *Journal of Business Ethics, 178*(4), 1027-1041.

Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., & Pietrini, P. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience, 23*(8), 1911-1920.

Gorny, P. M., Renner, B., & Schäfer, L. (2023). Prosocial behavior among human workers in robot-augmented production teams—A field-in-the-lab experiment. *Frontiers in Behavioral Economics, 2,* 1220563.

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 56-67.

Granulo, A., Caprioli, S., Fuchs, C., & Puntoni, S. (2024). Deployment of algorithms in management tasks reduces prosocial motivation. *Computers in Human Behavior, 152,* 108094.

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

Gray, K., Waytz, A., & Young, L. (2012a). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry, 23*(2), 206-215.

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

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

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*(8), 322-323.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108.

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

Haidt, J. (2008). Morality. *Perspectives on Psychological Science, 3*(1), 65-72.

Harris, L. T. (2024). The neuroscience of human and artificial intelligence presence. *Annual Review of Psychology, 75*(1), 433-466.

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems, 32*(3), 101772.

Hogenhuis, A., & Hortensius, R. (2022). Domain-specific and domain-general neural network engagement during human-robot interactions. *European Journal of Neuroscience, 56*(10), 5902-5916.

Hohensinn, L., Willems, J., Soliman, M., Vanderelst, D., & Stoll, J. (2024). Who guards the guards with AI-driven robots? The ethicalness and cognitive neutralization of police violence following AI-robot advice. *Public Management Review, 26*(8), 2355-2379.

Holthöwer, J., & Van Doorn, J. (2023). Robots do not judge: Service robots can alleviate embarrassment in service encounters. *Journal of the Academy of Marketing Science, 51*(4), 767-784.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences, 13*(1), 1-6.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience, 24*(6), 102679.

Kim, B., Wen, R., de Visser, E. J., Tossell, C. C., Zhu, Q., Williams, T., & Phillips, E. (2024). Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice. *International Journal of Human-Computer Studies, 184,* 103217.

Kim, H. Y., & McGill, A. L. (2024). AI-induced dehumanization. *Journal of Consumer Psychology, 34*(4), 123-145.

Kim, T., Lee, H., Kim, M. Y., Kim, S., & Duhachek, A. (2023). AI increases unethical consumer behavior due to reduced anticipatory guilt. *Journal of the Academy of Marketing Science, 51*(4), 785-801.

Kobis, N., Bonnefon, J. F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behavior, 5*(6), 679-685.

Krügel, S., Ostermaier, A., & Uhl, M. (2023a). Algorithms as partners in crime: A lesson in ethics by design. *Computers in Human Behavior, 138,* 107483.

Krügel, S., Ostermaier, A., & Uhl, M. (2023b). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports, 13*(1), 4569.

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage, 54*(3), 2492-2502.

Lanz, L., Briker, R., & Gerpott, F. H. (2024). Employees adhere more to unethical instructions from human than AI supervisors: Complementing experimental evidence with machine learning. *Journal of Business Ethics, 189*(3), 625-646.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis) honesty. *The Economic Journal, 134*(658), 766-784.

Li, H., & Zhang, R. (2024). Finding love in algorithms: Deciphering the emotional contexts of close encounters with AI chatbots. *Journal of Computer-Mediated Communication, 29*(5), zmae015.

Liu, P., & Du, Y. (2022). Blame attribution asymmetry in human-automation cooperation. *Risk Analysis, 42*(8), 1769-1783.

Liu, Y., Wang, X., Du, Y., & Wang, S. (2023). Service robots vs. human staff: The effect of service agents and service exclusion on unethical consumer behavior. *Journal of Hospitality and Tourism Management, 55,* 401-415.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., & Söllner, M. (2019). AI-based digital assistants: Opportunities, threats, and research perspectives. *Business and Information Systems Engineering, 61,* 535-544.

Mars, R. B., Neubert, F. X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. (2012). On the relationship between the "default mode network" and the

"social brain". *Frontiers in Human Neuroscience, 6,* 189.

Martin, D. U., Perry, C., MacIntyre, M. I., Varcoe, L., Pedell, S., & Kaufman, J. (2020). Investigating the nature of children's altruism using a social humanoid robot. *Computers in Human Behavior, 104,* 106149.

McKee, K. R., Bai, X., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *iScience, 26*(8), 107256.

Mell, J., Lucas, G., Mozgai, S., & Gratch, J. (2020). The effects of experience on deception in human–agent negotiation. *Journal of Artificial Intelligence Research, 68,* 633–660.

Melo, C. D., Marsella, S., & Gratch, J. (2016). People do not feel guilty about exploiting machines. *ACM Transactions on Computer–Human Interaction (TOCHI), 23*(2), 1–17.

Namkoong, M., Park, G., Park, Y., & Lee, S. (2024). Effect of gratitude expression of AI chatbot on willingness to donate. *International Journal of Human–Computer Interaction, 40*(20), 6647–6658.

Nash, K., Lea, J. M., Davies, T., & Yogeeswaran, K. (2018). The bionic blues: Robot rejection lowers self–esteem. *Computers in Human Behavior, 78,* 59–63.

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

Oliveira, R., Arriaga, P., Santos, F. P., Mascarenhas, S., & Paiva, A. (2021). Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Computers in Human Behavior, 114,* 106547.

Peter, J., Kühne, R., & Barco, A. (2021). Can social robots affect children's prosocial behavior? An experimental study on prosocial robot models. *Computers in Human Behavior, 120,* 106712.

Petisca, S., Paiva, A., & Esteves, F. (2020). *Perceptions of people's dishonesty towards robots. In Peoceddings of the Social Robotics. ICSR 2020*(pp. 132–142). Lecture Notes in Computer Science, Springer, Cham.

Quach, S., Cheah, I., & Thaichon, P. (2024). The power of flattery: Enhancing prosocial behavior through virtual influencers. *Psychology and Marketing, 41*(7), 1629–1648.

Rosenthal–Von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *Journal of Neuroscience, 39*(33), 6555–6570.

Sabour, S., Zhang, W., Xiao, X., Zhang, Y., Zheng, Y., Wen, J., & Huang, M. (2023). A chatbot for mental health support: Exploring the impact of Emohaa on reducing mental distress in China. *Frontiers in Digital Health, 5,* 1133987.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo–parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835–1842.

Schindler, S., Kruse, O., Stark, R., & Kissler, J. (2019). Attributed social context and emotional content recruit frontal and limbic brain regions during virtual feedback processing. *Cognitive, Affective, and Behavioral Neuroscience, 19,* 239–252.

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

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 and Society, 22*(5), 648–663.

Song–Nichols, K., & Young, A. G. (2020, July). *Gendered robots can change children' s gender stereotyping.* Poster at The CogSci 2020 Virtual Meeting, Seattle, WA, United States.

Stern, D. N. (2010). *Forms of vitality: Exploring dynamic experience in psychology, the arts, psychotherapy, and development.* Oxford University Press.

Tang, H., Ye, P., Wang, S., Zhu, R., Su, S., Tong, L., & Liu, C. (2017). Stimulating the right temporoparietal junction with tDCS decreases deception in moral hypocrisy and unfairness. *Frontiers in Psychology, 8,* 2033.

Tang, P. M., Koopman, J., Mai, K. M., De Cremer, D., Zhang, J. H., Reynders, P., & Chen, I. (2023). No person is an island: Unpacking the work and after–work consequences of interacting with artificial intelligence. *Journal of Applied Psychology, 108*(11), 1766–1789.

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict.* Cambridge University Press.

Vollmer, A. L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics, 3*(21), eaat7111.

Wang, K., Lu, L., Fang, J., Xing, Y., Tong, Z., & Wang, L. (2023). The downside of artificial intelligence (AI) in green choices: How AI recommender systems decrease green consumption. *Managerial and Decision Economics, 44*(6), 3346–3353.

Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs. human–robot interactions. *Social Cognitive and Affective Neuroscience, 10*(11), 1515–1524.

Williams, R., Machado, C. V., Druga, S., Breazeal, C., & Maes, P. (2018, June). *"My doll says it's ok": A study of children's conformity to a talking doll.* Poster at Interaction Design and Children, Trondheim, Norway.

Yam, K. C., Tang, P. M., Jackson, J. C., Su, R., & Gray, K. (2023). The rise of robots increases job insecurity and maladaptive workplace behaviors: Multimethod evidence. *Journal of Applied Psychology, 108*(5), 850.

Yang, J., Chuenterawong, P., Lee, H., Tian, Y., & Chock, T. M. (2023). Human versus virtual influencer: The effect of humanness and interactivity on persuasive CSR messaging. *Journal of Interactive Advertising, 23*(3), 275–292.

Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences, 121*(14), e2319112121.

Zhang, R. Z., Kyung, E. J., Longoni, C., Cian, L., & Mrkva, K. (2025). AI–induced indifference: Unfair AI reduces prosociality. *Cognition, 254,* 105937.

Zhou, Y., Fei, Z., He, Y., & Yang, Z. (2022). How human–chatbot interaction

impairs charitable giving: The role of moral judgment. *Journal of Business Ethics, 178*(3), 849–865.

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human–Computer Studies, 100*, 48–54.

How AI Reshapes Human Morality: Behavioral, Psychological, and Neural Perspectives

Guo Xueli[1], *Su Song*[1], *Liu Chao*[2,3,4], *Tang Honghong*[1]

([1]Business School, Beijing Normal University, Beijing, 100875)([2] State Key Laboratory of Cognitive Science and Learning & IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, 100875)([3] Center for Collaboration and Innovation in Brain and Learning Sciences, Beijing Normal University, Beijing, 100875)([4] Beijing Key Laboratory of Safe AI and Superalignment, Beijing Normal University, Beijing, 100875)

Abstract The rapid development and widespread use of Artificial Intelligence (AI) have given rise to significant ethical concerns, with a key question being how AI influences human morality. This paper reviews relevant literature to examine the multifaceted impacts of AI on human morality.

First, we propose potential underlying mechanisms to illustrate how AI applications influence human morality. Studies show that people typically attribute less "mind" to AI than to humans. This perception fundamentally influences humans' cognitive judgments and emotional responses during human-AI interactions compared to human-human interactions, resulting in distinct patterns of moral cognition, moral emotions, and social evaluations. The most notable influence on moral cognition may be how humans change their attribution patterns and judgments of responsibility in morally relevant behaviors within Human-AI interactions. According to previous studies, when AI systems cause harm, people tend to blame the AI more and attribute a higher degree of responsibility for the harm to the AI than to human agents. People consider AI as a "scapegoat" for immoral behaviors in many situations, facilitating their immoral behaviors. Meanwhile, Human-AI interactions could enhance humans' tendency for utilitarianism in moral judgment, which may reduce prosocial behavior. Furthermore, human-AI interactions typically mitigate humans' emotional responses toward stimuli or behaviors related to morality. This tendency may contribute to increasing immoral behavior and decreasing prosocial behavior. Specifically, when AI initiates prosocial actions instead of humans, people's moral emotional response is diminished, which reduces their prosocial tendencies. Additionally, human-AI interactions could undermine people's concerns for social image. When interacting with AI, individuals tend to adhere less to moral principles and become less sensitive to social evaluation. This reduced social awareness may subsequently increase unethical behavior or decrease prosocial behavior.

Then, we discuss the neural mechanisms underlying how human-AI interactions influence human morality. When people interact with AI rather than humans, their brain regions involved in social perception, emotion, and cognition show significant differences. Three primary brain networks demonstrate such differences: the interoception and human stereotype network, the social cognition network, and the valuation network. The interoception and human stereotype network primarily includes insula and MCC. These brain regions are less active during interactions with AI than with humans. The social cognition network involves the TPJ, STS, TP, SPL, PCC, and MTG. Activities in these regions are also decreased when people interact with AI. The valuation network primarily involves the MPFC, including the VMPFC and DMPFC. During human-AI interactions, compared to human-human interactions, both the ventromedial prefrontal cortex (VMPFC) and the dorsomedial prefrontal cortex (DMPFC) exhibit reduced activity during decision-making. This suggests diminished processing and evaluation of decision-related information.

Finally, we discuss the positive and negative effects of AI on human morality across four different forms of human-AI interaction: AI delegator, AI advisor, AI collaborator, and AI moral model. When AI plays as a delegator, people may use AI as a proxy for executing immoral actions, or promote cooperation and fairness to benefit themselves. When AI serves as an advisor, its immoral suggestions can promote deceptive behavior, yet its moral advice does not necessarily encourage honesty. When AI functions as a collaborator, it has a mixed impact on human moral behaviors, revealing both positive and negative effects. When AI acts as a moral role model, it shows the potential to positively shape moral cognition and behavior. Its impact on children's moral development is particularly pronounced, surpassing that observed in adults.

Future research should focus more on exploring how different AI applications impact human morality, particularly their long-term effects. These efforts will provide valuable theoretical support and guidance for the design, development, and application of AI.

Key words artificial intelligence, human-AI interaction, moral, social brain networks