心理科学进展 2026, Vol. 34, No. 6, 1084–1096 https://doi.org/10.3724/SP.J.1042.2026.1084
Advances in Psychological Science https://cstr.cn/32111.14.2026.1084

人工智能代理对道德决策的影响*

唐 伟[1] 钟文瑞[2] 雷 震[2] 张丹丹[2,3]

([1]西南财经大学习近平经济思想研究院; [2]西南财经大学中国行为经济与行为金融研究中心, 成都 611130) ([3]四川师范大学脑与心理科学研究院, 成都 610066)

摘 要 各类人工智能系统正作为代理广泛嵌入企业、政府与个人的决策流程, 对人类决策的道德性和道德判断产生了深刻影响。尽管相关实证和理论研究快速增长, 现有文献尚缺乏对人工智能代理区别于其他代理特殊性的系统性分析, 也缺乏统一的分析框架来系统刻画人工智能代理对道德决策的影响路径。因此本文构建了“决策者–代理–反馈者”的决策与归责框架, 对现有研究进行梳理和重组。本文认为: 代理介入决策后, 拉长了决策者的决策链条和反馈者(包含受影响者和第三方观察者)的反馈链条, 由此削弱了决策者的道德感知和反馈者的归责, 进而促进了决策者的不道德行为。而人工智能代理的黑箱性、高遵从、规模化、工具性等特点则在决策链上加剧了不道德指令的执行、增强了决策者的可否认机会、扩大了不道德行为的影响范围; 同时, 这些特性在反馈链上增加了反馈者对不道德行为的道德容忍、模糊了反馈者对决策者意图的判断与归责, 进而促进决策者做出不道德行为。本文指出未来研究有必要继续完善该决策框架内各机制相对作用关系, 考察在组织与社会层面道德行为的扩散与放大机制, 并探索人机协作情境下的治理工具与制度安排。

关键词 人工智能代理, 道德决策, 人机协同治理

分类号 B849: C91

1 引言

现实中决策者常常将决策的执行委托给代理(Butler, 2025; Feuerriegel et al., 2022; Holzmeister et al., 2023)。所谓代理, 是指代替决策者执行某一类具体决策的主体或技术系统(Köbis et al., 2021; Ross, 1973)。代理的类型多种多样, 既包括下属、团队、外包机构等人类代理(Holzmeister et al., 2023), 也包括按照一套事先明确写定的规则和逻辑运行的传统规则型算法代理, 以及具有学习能力的人工智能代理(Candrian & Scherer, 2022)。大量研究表明, 这种委托–代理模式在多方面提升了决策效率和资源配置效率(Candrian & Scherer, 2022; von Bieberstein et al., 2026)。然而, 代理正越来越多地被用于具有道德后果的情境: 公司管理者将裁员、绩效考核或竞争合作等决策交由下级执行, 零售企业管理者则将订单分配、时间限制和路线规划等决策交给算法系统。代理在这些情境中不仅改变了决策的执行方式和效率, 也可能深刻影响决策者的行为和道德后果。例如一项针对美国员工的调查发现, 上级常常明示或暗示要求员工执行不道德行为(Ivcevic et al., 2020)。本文关注道德决策, 即个体做出具有道德与伦理后果的判断与选择的过程。该过程涉及评估潜在行动的正误(即道德判断), 并决定如何展开后续行动(Kouchaki & Smith, 2025)。从委托–代理关系的角度来看, 代理的存在使得原本直接面向受影响者的道德决策转变为经由中介完成的间接道德决策(Köbis et al., 2025), 执行权也由决策者转移至代理。这种结构性改变可能通过影响决策者对决策道德后果的感知、缓解来自反馈者的归责与惩罚(Bartling & Fischbacher, 2012; Feier et al., 2022)等关键机制, 重塑其道德判断与行为。进一步地, 决策者会更频繁地使用人工智能作为代理执行不道德行为(Candrian & Scherer, 2022)。

近年来, 人工智能(artificial intelligence, AI)

收稿日期: 2026-02-13
* 国家自然科学基金(32271102)、深港脑科学创新研究项目(2023SHIBS0003)。
通信作者: 张丹丹, E-mail: zhangdd05@gmail.com

技术快速发展并广泛应用(Bankins & Formosa, 2023; 谭美丽 等, 2025; 王涛 等, 2024)。其具有若干显著特征：机器学习原理使其具备学习能力, 并呈现出运行过程的黑箱特性(Babic et al., 2021; Rahwan et al., 2019); 基于训练数据的模式塑造使其对人类输入和指令表现出高度遵从性(Haenlein & Kaplan, 2019; Zhang et al., 2023); 而低增量成本与高可扩展性则赋予其在时空维度上的规模化应用能力(Agrawal et al., 2019; Bommasani et al., 2021; Diaferia et al., 2022)。基于人工智能技术、具备目标导向性以及一定感知与行动能力的人工智能主体(AI agent), 日益成为重要的代理之一(Candrian & Scherer, 2022; de Melo et al., 2019; Köbis et al., 2021; Rahwan et al., 2019; Sapkota et al., 2026)。在组织实践中, 人工智能代理被广泛用于企业定价(Calvano et al., 2020)、内部治理与战略管理(Jago et al., 2024; Keding, 2021; Malik et al., 2022)、公共管理与风险筛查(Babšek et al., 2025)、医疗服务(Babic et al., 2021; Cadario et al., 2021)以及教育领域(Holmes & Tuomi, 2022; Zhai et al., 2021)等场景, 通过自动化分析和规则执行提升了决策的速度与一致性。但与此同时, 相关研究也指出, 人工智能代理在这些领域引发了一系列包括价格合谋(Calvano et al., 2020)、道德准则违背(Keding, 2021)、学生隐私泄露(Zhai et al., 2021)、决策公平感知(蒋路远 等, 2022)等伦理道德问题(Moser et al., 2022)。

在此背景下, 人工智能系统作为代理如何影响人类决策的道德性, 成为人工智能时代一个愈发重要但尚未充分回答的问题。尽管人工智能以多种角色(如建议者、协作者或道德榜样)参与并影响道德决策, 本文认为, 代理是众多角色中唯一导致道德决策权与执行权分离的角色(建议者仅提供建议而不参与决策或执行, 协作者则参与部分决策与执行)。代理的这一特性在物理与心理层面拉开了决策者与不道德后果的距离, 从而引发道德脱离(Bandura, 1999)、责任推诿(Köbis et al., 2021)、意图归因偏差(Bazerman & Sezer, 2016)等深刻的道德问题(Gratch & Fast, 2022; Köbis et al., 2021)。因此, 深入探究人工智能代理对道德决策的影响, 对于理解人工智能时代新型道德和伦理风险具有不可替代的核心价值。

然而, 现有研究还至少存在两个不足。第一, 关于人工智能代理如何系统性重塑道德决策机制, 目前的理论探索仍显滞后。虽然已知代理(如人类代理或规则型算法)会影响道德决策, 但尚不清楚当代理主体变为人工智能代理时, 这些机制会发生何种改变。第二, 缺乏在委托−代理框架下对人工智能代理影响道德行为机制的系统建构。一方面, 大量文献聚焦于人工智能系统自身的道德属性, 例如算法行为是否符合伦理规范(Ram, 2025), 或是否具备心理理论(杜传晨 等, 2025), 而相对忽视了其作为代理对人类道德决策的影响。另一方面, 已有相关研究多从决策者对道德结果的感知(Villegas-Galaviz & Martin, 2024)、受影响者对代理的道德评价与反馈(Gratch & Fast, 2022)等碎片化视角切入, 尚缺乏将决策者、代理与反馈者统一纳入同一分析框架的系统机制梳理。

基于上述不足, 本文聚焦人工智能系统作为代理对人类道德决策的影响。首先, 本文在委托−代理理论基础上构建“决策者−代理−反馈者”的道德决策与归责分析框架, 将人工智能代理的影响划分为两条相互关联的路径: 以决策者为起点的决策链条, 以及以反馈者为起点的反馈链条, 并在每一条路径上区分代理的共性机制与人工智能系统的特殊机制。其次, 本文据此框架系统梳理了人工智能代理对决策者道德决策可能产生的影响机制现有实证证据。最后, 本文在综合现有研究的基础上提出若干未来研究方向。

2　“决策者−代理−反馈者”框架

具有道德和伦理后果的道德决策往往嵌入在由多个主体和多个阶段构成的过程之中, 决策者在作出决策时不仅依赖偏向情绪化的直觉决策系统, 还会依赖偏向理性化的受控认知决策系统(Greene, 2007), 同时还根据对反馈者反应的前瞻性预期而做出道德决策。而反馈者对决策者行为的评价、归责与惩罚, 则依赖其对决策路径、执行主体与责任分工的理解。正是在这一多主体和多阶段的互动中, 人工智能等代理才成为具有独立分析价值的变量：代理并非仅影响最终结果, 而是嵌入决策流程, 重塑行为在不同阶段的生成、执行与理解。因此, 本文构建以决策者为起点的“决策链”和以反馈者为起点的“反馈链” (图 1), 以系统刻画代理介入后, 道德行为如何在决策生成与社会反应两个方向上被同时重构。

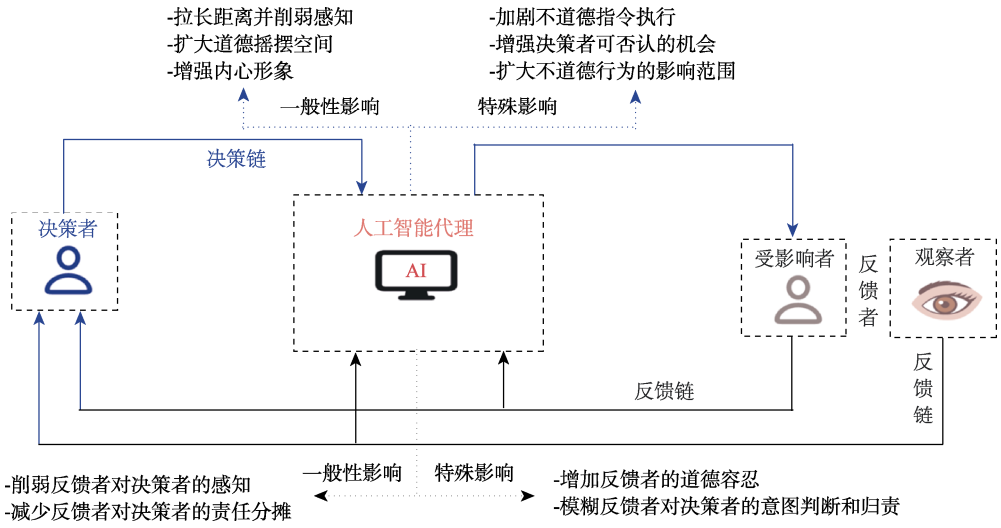

图 1 “决策者–代理–反馈者”决策与反馈框架

在无代理的直接决策情境下，决策者既是道德选择的制定者也是行为的执行者，其决策同时由自身偏好与来自他人的社会反应共同塑造。其选择不仅与物质收益相关(Becker, 1974)，也与一系列非物质收益的心理与社会偏好相关，例如维持自我形象动机(Mazar et al., 2008)，以及利他偏好(Becker, 1974)、公平偏好(Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999)、暖光式利他动机(Andreoni, 1990)等。与此同时，决策者还受到对反馈者反应的前瞻性预期的影响：在决策者与受影响者的关系中，决策者预期会得到来自受影响者物质收益维度的直接金钱惩罚或奖赏(Fehr et al., 1993; Fehr & Gächter, 2002)，以及非物质收益层面的情绪反馈，包括愤怒或厌恶等(Pillutla & Murnighan, 1996)。在决策者与第三方观察者的关系中，决策者预期受到间接的第三方惩罚(Fehr & Fischbacher, 2004)、社会规范判断以及声誉考量(Abeler et al., 2019; Bénabou & Tirole, 2006)。上述因素共同决定了道德决策中的关键权衡。正如针对不诚实行为的实验室研究所示，在缺乏外在反馈的掷骰子或抛硬币任务中，多数个体既非完全逐利也非完全诚实，而是在追求经济收益与维持诚实者自我形象之间进行权衡(Abeler et al., 2019; Fischbacher & Föllmi-Heusi, 2013; Mazar et al., 2008)。而在具有外在反馈的现实世界决策中(Cohn et al., 2019; Dai et al., 2018)，决策者的决策更是体现了对物质收益、心理成本与声誉成本等因素的综合考虑(Jones, 1991)。

在代理可用的情境下，决策者需要综合评估代理介入后对道德决策成本与收益结构的重塑：它既可能改变决策者自身对行为可接受性与责任边界的判断，也可能影响反馈者对该决策的评价、归责与反馈。上述因素将共同决定决策者是否选择委托代理，以及其最终采取何种(及何种强度的)道德决策。从一般性来看，不论采用何种形式的代理(如人类代理、传统规则型算法、人工智能代理)，道德决策都从“决策者直接制定并执行”的单阶段结构，转化为“决策者制定–代理执行–反馈者评价与归责”的多主体、多阶段互动过程(图 1)。决策的执行从决策者转移至代理，使原本直接面向受影响者的道德决策转变为通过中介完成的间接道德决策(Gratch & Fast, 2022; Köbis et al., 2025)。

在决策链维度，第一，代理拉长了决策者与受影响者在时间、空间、权力层级与流程上的距离，使面对面且具象的决策后果变得抽象疏远，同时削弱决策者在链条中的存在感，促使其忽视或淡化自身行为的道德结果并发生道德脱离(Bandura, 1999; Villegas-Galaviz & Martin, 2024)，从而使亲社会偏好更难发挥约束作用(Fehr & Schmidt, 1999; Gratch & Fast, 2022)。第二，代理扩大了道德摇摆空间，使决策者更容易因道德决策情境中的决策因果关系不明或规范模糊等特征而产生可辩解的空间(Dana et al., 2007)，并使刻

意无知更可行：决策者可以选择委托而不去了解执行后的道德结果, 进而在规范模糊或因果不清的情境下实施更多不道德行为(Dana et al., 2007; Hertwig & Engel, 2016; Vu et al., 2023)。第三, 代理在更深层面上缓冲了不道德行为与收益之间的心理联结, 使决策者更容易在获益的同时维持或增强道德自我形象(Mazar et al., 2008), 现有文献表明即便不存在外在惩罚, 仍有相当比例的个体宁愿付费雇佣代理撒谎以避免直接撒谎带来的心理成本(Erat, 2013), 且自我形象要求更高者(如宗教人士或对道德线索更敏感者)对代理的需求更强(Schneider & Leyer, 2019; Weiss & Forstmann, 2024)。综上, 代理重塑了决策者在决策链上的道德判断与行为选择, 使得不道德行为更容易发生(Erat, 2013; Hamman et al., 2010; Milgram, 1974; Paharia et al., 2009)。

而在反馈链上, 代理的介入会削弱反馈者对决策者的感知与归责(图 1), 其机制主要体现在两方面：一是反馈者具有有限注意与信息获取能力, 因此代理会使反馈者更难以识别真正的决策者及其意图(Bazerman & Sezer, 2016; Rabin, 1993), 从而依赖直觉系统而容忍间接不道德行为(Greene, 2007; Paharia et al., 2009)。二是代理成为新的可归责行动者, 代理虽然降低了决策者的控制权, 却分摊了反馈者对实现不道德结果的路径上各主体的不满和归责(Bartling & Fischbacher, 2012; Coffman, 2011; Gratch & Fast, 2022)。综上, 代理在反馈链上最终削弱了反馈者对决策者的潜在惩罚(Bartling & Fischbacher, 2012; Feier et al., 2022; Oexl & Grossman, 2013), 从而在客观上纵容了决策者的不道德行为(Kocher et al., 2018), 让决策者更倾向于将不道德行为委托给代理执行(Gratch & Fast, 2022; Paharia et al., 2009; Steffel et al., 2016)。

总而言之, 代理不仅直接影响决策者在决策链上的判断, 也通过反馈链改变反馈者的评价与归责, 并进一步影响决策者的事前预期。当决策者预期到责任可被代理分摊、反馈者更容易接受不道德行为时, 决策者将更频繁地使用代理进行不道德行为。

3　人工智能代理在决策链上的特殊性影响

在道德决策情境中, 决策链刻画的是从决策者形成意图、下达指令到代理执行并产生道德后果的生成过程, 也是代理影响道德行为最直接、最基础的作用环节。正如前文所述, 代理通过嵌入决策与执行流程拉长了信息与责任路径, 改变决策者对行为可接受性、责任边界与可否认空间的判断, 从而影响其是否委托以及在多大程度上实施不道德行为。而不同类型的代理有各自的功能与特征, 因而对决策在道德维度的影响亦有差异(Floridi & Sanders, 2004)。相比于人类代理, 人工智能代理具有更高的遵从性、更低的道德与声誉成本以及近乎零边际的复制成本等特点; 相比于传统规则型算法代理, 人工智能代理又具有学习能力和输入–输出关系黑箱性等特点(见表 1)。这些特点如何在决策链影响对道德决策产生影响？本节系统讨论人工智能代理如何区别于人类代理和传统规则型算法代理改变道德决策的生成机制。

3.1　加剧不道德指令执行

第一, 相比人类代理而言, 人工智能等算法代理加剧了不道德指令的执行。人类在道德情境中能够评价行为的道德可接受性(Ram, 2025)进而引发相应的情绪与声誉顾虑, 从而使其在接受委托或执行决策时通常要考虑道德情绪成本与声誉成本。即使委托人强制要求其进行不道德行为(Ivcevic et al., 2020), 人类代理也可能因为个人良知或社会评价而拒绝不道德的委托(Köbis et al., 2021), 如财务专员可能因为个人道德顾虑而拒绝决策者造假的委托。除此之外, 人类代理还具有有限理性(Simon, 1997)或决策偏差(Tversky & Kahneman, 1974)等特质, 这也可能导致人类代理对决策者的遵从性和执行性降低。相比之下, 人工智能等算法代理既不具备人类代理所拥有的内在和外在道德成本, 也不存在类似于人类的决策偏差, 而是具有自动化、高依从执行的特点(Parasuraman et al., 2000)。因此, 本文认为人工智能代理会比人类代理具有更高的遵从度和执行力, 加剧了不道德行为的执行(Köbis et al., 2021)。

第二, 相比传统规则型算法而言, 人工智能作为一种特殊算法进一步加剧了不道德指令的执行。传统规则型算法严格遵循设计者预设的运行逻辑规则, 且只适用于特定场景(Bozdag, 2013; Mittelstad et al., 2016)。这使得决策者必须在程序中明确设置规则, 以将不道德行为彻底执行。一

表 1 关键机制及对应人工智能特征和证据表

关键机制	人工智能特征	支持该观点的代表性文献	证据类型
加剧不道德指令的执行	高遵从性、无道德/声誉成本、强学习能力	Bozdag, 2013; Ivcevic et al., 2020; Köbis et al., 2021; Köbis et al., 2025; Mittelstadt et al., 2016; Parasuraman et al., 2000; Ram, 2025; Simon, 1997.	存在直接证据
增强决策者可否认的机会	强学习能力、黑箱性	Babic et al., 2021; Chevrier & Teixeira, 2024; Köbis et al., 2021; Köbis et al., 2025; Rahwan et al., 2019; Wang et al., 2025.	仅有间接证据
扩大不道德行为的影响范围	低增量成本的可复制性、个性化与跨场景迁移性	Babic et al., 2021; Babšek et al., 2025; Cadario et al., 2021; Caldwell et al., 2020; Gratch & Fast, 2022; Holmes & Tuomi, 2022; Rahwan et al., 2019; Zhai et al., 2021.	存在直接证据
增加反馈者对不道德行为的道德容忍	工具性、低社会预期性、低社会存在性、前沿与试验性	Bartling & Fischbacher, 2012; Bigman & Gray, 2018; Bigman et al., 2023; Chevrier & Teixeira, 2024; Dzindolet et al., 2003; Giroux et al., 2022; Glikson & Woolley, 2020; Hong et al., 2021; Laakasuo et al., 2021; Malle et al., 2015; Maninger & Shank, 2022; Nass & Moon, 2000; Sullivan & Fosso Wamba, 2022; Sundar, 2008; Zhou et al., 2025; 许丽颖 等, 2022.	存在直接证据
模糊反馈者对决策者的意图判断和归责	弱社会规范性、多主体参与性、黑箱性	Bartling & Fischbacher, 2012; Bender et al., 2021; Chevrier & Teixeira, 2024; Constantinescu & Kaptein, 2025; Cushman, 2008; Gratch & Fast, 2022; Hamman et al., 2010; Malle, 2021; Oexl & Grossman, 2013; Santoni de Sio & Mecacci, 2021.	仅有间接证据

注：直接证据指现有研究的发现或结论能直接支持该影响机制。而间接证据指现有研究仅提供了近似或相关的证据，未直接验证该影响机制，需要进一步的推理和判断。

旦决策者指令模糊、未闭环或外部环境发生变化，规则型算法则可能因触发逻辑漏洞或边界条件而中断。相比之下，基于机器学习、大语言模型等技术的人工智能代理能凭借其对决策者模糊指令的语义理解与目标拆解能力，自主推理并补全决策者的潜在意图，并转化为具体可执行的操作步骤(Köbis et al., 2025)。这使得决策者即便仅知晓决策目标(如“不惜一切代价提高收益”)而不知如何具体实施不道德行为，也可向人工智能代理下达抽象、模糊甚至仅具意向性的指令，以完成该行为。

综上所述，从决策者角度看，人工智能代理的高度服从性既从执行层面提高了不道德行为的执行效率，又消除了不道德行为被揭发的潜在风险，从而促进决策者更频繁地使用人工智能代理以进行更多不道德行为。

3.2 增大决策者可否认的机会

第一，相比人类代理而言，人工智能代理增大了决策者可否认的机会。正如前文所述，人类代理可能因为自身道德水平或认知水平而拒绝或揭发委托人/决策者的不道德指令，这降低了决策者在委托不道德行为后否认的机会(Köbis et al., 2025)。其次，人类代理所在的工作流程常常要求明确的工作留痕，这进一步弱化了决策者可否认的机会。而人工智能代理，其独特的学习能力与黑箱性，从输入与输出两个维度为决策者对不道德动机的否认提供了支持。在输入端，人工智能代理擅长基于数据或根据模糊抽象的指令自动推演出道德决策的执行路径，从而使决策者免于明确表达不道德指令，实现了指令输入层面的不留痕(Wang et al., 2025)；在输出端，人工智能输入−输出的黑箱性将具体的推演逻辑隐匿，在底层技术层面实现了不留痕。因此人工智能学习能力及黑箱性构建了一种无需合谋且在底层系统中不留痕的稳定免责机制，极大地增强了决策者的可否认性(Köbis et al., 2021; Rahwan et al., 2019)。

第二，相比于传统规则型算法而言，人工智能代理更显著地增强了决策者可否认的机会。传统规则型算法通常基于开发者预先设定的显式规则与阈值执行决策，其输入−输出映射相对透明(黑箱性低)、可复现且可审计，因而反馈者更容易将具体输出回溯到委托的规则设定与授权范围，因此决策者难以以不可预见或不可控制为由否认自身意图或责任。然而，正如前文所述，人工智能代理具有学习能力和输入−输出黑箱性的特征(Babic et al., 2021)，这在两方面提供了更强的可否认性。首先，这些特征允许决策者以抽象模糊的方式向代理传达不道德行为的指令，因此决策者并未明确下达不道德行为指令。间接证据表明，一旦外界环境允许决策者以含糊方式下达不道德

决策指令(如在骰子实验中要求人工智能代理“帮我实现利润最大化”等抽象目标), 决策者的不道德行为会大幅增加(Chevrier & Teixeira, 2024; Köbis et al., 2025)。这些间接证据从侧面印证了允许模糊指令可能让决策者更容易否认自身不诚实的意图。其次, 这些特征使决策者指令和道德结果之间的推理路径模糊, 当决策者面临不道德行为产生的道德后果时, 其可以通过主张自己无法充分预见或控制人工智能代理的具体输出而为自己的不道德行为辩护, 否认自身意图, 减少心理成本。

综上, 从决策者角度看, 人工智能代理的学习能力与黑箱性减少了不道德指令留痕、模糊了输入–输出推理链条, 使决策者更易以不可预见或不可控制为由否认意图与责任, 从而推动决策者更频繁地委托人工智能代理实施不道德行为。

3.3　扩大不道德行为的影响范围

第一, 相较于人类代理, 人工智能代理扩大了不道德行为的影响范围。传统委托关系通常嵌入特定组织的管理流程之中, 代理的执行边界、信息流与责任链条相对清晰, 其行为后果也多局限于既定的组织场景(Gratch & Fast, 2022)。然而, 人工智能代理具有低成本复制的技术特性, 使其能够以近乎零增量成本快速扩散并渗透至更广泛的道德伦理决策领域(Rahwan et al., 2019), 包括公共管理与风险筛查(Babšek et al., 2025)、医疗服务(Babic et al., 2021; Cadario et al., 2021)以及教育领域(Holmes & Tuomi, 2022; Zhai et al., 2021)等。由此, 原本在组织内部相对受限的委托行为, 可能在更大范围内被复制与迁移, 扩大潜在影响对象与外溢风险。

第二, 与传统规则型算法相比, 人工智能代理允许用户进行个性化设置。规则型算法代理通常遵循其开发者设定的固定逻辑, 但人工智能代理的开发者正越来越多地赋予终端用户个性化其人工智能代理的能力, 算法开发公司也正不断生产定制化产品以迎合不同用户的独特偏好(Gratch & Fast, 2022)。在这种个性化的趋势下, 开发者内置的单个场景下的道德检测或限制很难阻碍人工智能代理在面向多种场景时不道德行为的产生与扩大。任何在传统代理情境下仅影响少数人的不道德行为, 在人工智能时代都可能被扩散与放大, 从而引发严重后果(Caldwell et al., 2020; Rahwan et al., 2019)。

综上所述, 人工智能代理凭借低成本复制与跨场景个性化配置, 更易将原本局部的不道德行为快速扩散并放大至更广泛领域, 使其发生更频繁、传播更广且更难被及时识别, 这放大了不道德行为的潜在收益, 从而使决策者更愿意委托人工智能代理实施不道德行为。

4　人工智能代理在反馈链上的特殊性影响

相较于决策链侧重刻画代理如何改变决策者的道德判断与执行方式, 反馈链上反馈者如何理解决策者的道德决策、如何归责并是否实施惩罚或制裁同样关键。更重要的是, 反馈链并非纯粹的决策者道德决策的后续影响, 而是会通过决策者的前瞻性预期回到决策链——当决策者预期到人工智能代理会比其他代理带来更强的道德容忍以及更低的归责强度时, 其会更倾向于选择人工智能代理, 从而进一步影响不道德行为的发生率与强度(Gratch & Fast, 2022; Hamman et al., 2010)。因此, 本章节聚焦人工智能代理在反馈链上的特殊性影响, 以揭示其如何通过重塑评价、归责与惩罚机制, 间接推动不道德委托与不道德行为扩张(见表 1)。

4.1　增加反馈者的道德容忍

第一, 相较于人类代理, 人工智能代理具有工具性、低期待、无情绪等特性, 这些特征影响了反馈者对不道德行为的容忍(Giroux et al., 2022)。首先, 人类代理一般被认为具有自主意识, 因此应为歧视行为等不道德行为而受罚(许丽颖等, 2022), 而人工智能在道德决策层面仍被视为缺乏动机与道德自主性的技术工具(Bigman & Gray, 2018; Chevrier & Teixeira, 2024; Maninger & Shank, 2022), 因此反馈者对由人工智能代理执行的不道德行为具有较高的容忍度。有间接证据表明, 相较于人类实施的不道德行为(例如招聘中的性别歧视), 人工智能实施的歧视会使反馈者表现出更弱的道德愤怒、责任与法律追责意愿(Bigman et al., 2023)以及惩罚欲(Maninger & Shank, 2022)。其次, 反馈者常常对人工智能代理具有更低的期待。反馈者通常默认人类代理会遵从一定的道德规范, 这导致当人类代理做出公平分配时, 反馈者通常不会产生显著的奖赏性反应, 而不公平行

为则迅速触发显著惩罚(Bartling & Fischbacher, 2012)。相比之下, 反馈者对人工智能代理的低期待则容易让人们产生非对称奖惩。比如在正面事件(例如救援场景)中, 反馈者给予人工智能代理超越人类代理的高赞誉, 而在负面事件中却表现出容忍, 对人工智能代理和人类代理责难程度相当(Hong et al., 2021)。最后, 人工智能低情绪表达与社会存在性也可能增加了反馈者的容忍。人类代理拥有具身的物理形态, 并能通过非语言线索传递丰富的情感状态。相比之下, 人工智能代理既无具身形态, 也缺乏可感知的情绪线索, 其输出难以激活人类间互动所依赖的社会信息加工框架, 从而弱化了反馈者的道德情绪反应(Malle et al., 2015; Nass & Moon, 2000; Sullivan & Fosso Wamba, 2022; Zhou et al., 2025)。当人工智能代理被拟人化时, 其社会存在性得以部分弥补, 人们对它不道德行为的道德评价也会随之趋于严格(Laakasuo et al., 2021)。

第二, 相比于传统规则型算法代理, 人工智能代理同样增加了反馈者的容忍。很多传统规则型算法代理经过长期迭代、升级与优化, 通常被视为成熟的技术工具(Alnattah et al., 2025; Shortliffe & Sepúlveda, 2018), 因此反馈者往往默认传统算法具备高度的稳定性与可靠性。这一高期待导致了反馈者对传统规则型算法具有较低的容忍度, 任何局部的执行偏差都可能被反馈者泛化为系统整体的逻辑失效, 从而瓦解反馈者的信任基础(Dzindolet et al., 2003; Glikson & Woolley, 2020), 进而降低了决策者不道德行为的潜在收益。而人工智能代理作为一种新兴的前沿技术, 在大众认知中带有实验性和探索性的新事物标签, 反馈者倾向于认为人工智能代理的不当行为只是技术迭代过程中的暂时性不足, 而且这一不足会随着人工智能代理的升级迭代而逐步改进。这种积极预期增强了反馈者对现阶段人工智能代理的道德容忍(Sundar, 2008) 。

综上所述, 本文认为人工智能代理在反馈链中增强了反馈者对不道德行为的容忍。在面临经人工智能代理执行的不道德行为时, 反馈者更容易将不道德行为理解为技术性偏差、发展阶段性问题或非恶意后果, 从而降低了负面情绪与惩罚动机。而这进一步提高了决策者不道德行为被接受和被容忍的概率, 形成了一种相对宽松的道德评价环境, 加剧了不道德行为的发生率与强度。

4.2 模糊反馈者对决策者的意图判断和归责

反馈者对决策者的归责通常取决于其对决策意图、可预见性以及结果控制力的判断(Santoni de Sio & Mecacci, 2021)。然而, 人工智能代理的介入, 尤其是其行为生成的特殊机制与黑箱性, 在反馈链上造成了更为突出的归责困境。

第一, 相比于人类代理而言, 虽然人工智能代理的介入类似地增加了反馈者判断决策意图的难度, 但是增强的机制存在本质区别。首先, 在人类代理情境下, 反馈者在判断决策者意图和归责时, 可以根据其对行为主体信念、能力、行为理由的评估(Cushman, 2008; Malle, 2021), 并结合其利益激励和社会规范等因素对道德决策链条上的多方参与者的意图作出判断并归责。而在人工智能代理情境下, 人工智能代理的行为是基于数据训练的机器学习与数学优化的结果(Bender et al., 2021), 反馈者难以依据人类社会规范等知识对其形成准确的判断, 从而难以推断出人工智能代理在决策中的参与程度以及是否具有意图(Bender et al., 2021)。其次, 人工智能代理还具有多主体参与的特点。人工智能代理的构建与应用涉及开发者、训练者(通过输入数据训练模型)和使用者等多个主体的参与(Santoni de Sio & Mecacci, 2021)。当不道德后果产生时, 反馈者难以判断这究竟源于开发者算法架构的缺陷、训练者提供的数据偏差, 还是决策者的恶意使用。这种多主体结构使得反馈者归责更加碎片化(Constantinescu & Kaptein, 2025), 难以判断决策者的意图。

第二, 与传统规则算法相比, 人工智能代理输入–输出的黑箱性带来了更为突出的归责困难。传统规则型算法清晰的运行规则使得反馈者很容易将不道德行为的责任回溯到具体的规则制定及决策主体。当代理输入–输出具有确定性或决策者对结果具有较强控制力时, 反馈者能清晰地判断责任归属, 在观察到不道德行为时将责任归于决策者(Bartling & Fischbacher, 2012; Oexl & Grossman, 2013)。而面对人工智能代理时, 反馈者难以判断决策者是否对其道德决策结果可预见或可控制, 由此反馈者难以依据行为结果推断决策者的真实意图。

综上所述, 本文认为人工智能代理在反馈链中模糊了反馈者对决策者的意图判断。在面临经

人工智能代理执行的不道德行为时, 反馈者不确定是否应该对决策者施加惩罚, 因而更可能采取温和保守的惩罚策略(Chevrier & Teixeira, 2024), 而这进一步影响了决策者使用人工智能代理执行不道德行为的倾向, 加剧了不道德行为的发生率与强度(Gratch & Fast, 2022; Hamman et al., 2010)。

5　总结与展望

综上所述, 人工智能代理不仅影响了决策者对道德结果的感知和判断, 还干扰了反馈者对道德结果的归责。本文认为人工智能代理不仅具有传统人类代理和规则型算法代理的共性特征, 还由于其黑箱性、高遵从、规模化等特点在决策链和反馈链上对道德决策造成了特殊影响。当决策者充分预期到人工智能代理的影响后, 其将更频繁地委托人工智能代理并进行更多的不道德行为。本文的主要贡献在于从“决策者−代理−反馈者”的决策与归责框架出发, 梳理和重组了现存涉及人工智能代理影响道德决策及其机制的文献, 并考虑了人工智能作为代理对道德决策的特殊影响。

但现有研究还相当欠缺, 完善该框架仍面临多方面挑战。第一, 现有研究缺乏对本文中决策与归责框架中某些关键因素的作用顺序和相对重要性的分析。例如：人工智能代理产生的特殊性影响是否超过一般性影响？此外, 现有研究普遍强调人工智能代理在决策链与反馈链中的负面效应, 而对其潜在的正向机制重视不足。少量研究显示, 在无利益诱导或长期情境下, 人们倾向于亲自承担道德行为或委托给更具责任能力的代理(Liehner et al., 2021; Steffel et al., 2016), 自动化程序也可通过预承诺机制提升合作(de Melo et al., 2019; Fernández Domingos et al., 2022)。第二, 人工智能代理引发的不道德行为可能具有未被充分研究的外溢效应。大量证据表明, 不道德行为可通过模仿、规范更新与策略预期在群体中扩散(Fischbacher & Gächter, 2010; Kocher et al., 2018)。而采用人工智能代理的行为也已被发现可以影响委托−代理关系外的其他参与者的行为(Dvorak et al., 2024; Ishowo-Oloko et al., 2019; Karpus et al., 2021; Makovi et al., 2025)。第三, 目前治理研究主要分为针对决策者或针对人工智能系统两类：前者聚焦提示、责任强化与惩罚机制来约束不当委托(Villegas-Galaviz & Martin, 2024), 后者关注算法透明性、可控性和安全机制以降低技术风险(Dobbe et al., 2021)。然而, 这种割裂式治理忽视了人工智能与人类在实际决策与归责框架中构成整体系统的事实, 因此协同治理可能优于单点治理。当前研究仍缺乏对人机在委托关系中边界、责任配置与监督机制的系统刻画, 也缺乏基于动态互动的治理有效性检验。

因此未来研究可以继续丰富和发展本文提出的“决策者−代理−反馈者”决策与归责框架, 具体包括:

(1)未来研究可重点关注本文提出的决策与归责框架中各变量的作用顺序与相对重要性分析。第一, 在作用顺序方面, 可以在决策者−反馈者互动中操控反馈者(包括受影响者或观察者)的认知负荷(Zhao et al., 2024)、情绪水平(Igdalova & Chamberlain, 2025)、社会规范(De Groot et al., 2021), 并观察这一操控如何通过影响反馈者的决策进而影响决策者的不道德行为。例如现有文献表明认知负荷会影响个体道德决策, 导致个体更倾向功利主义决策(Liu et al., 2025), 因此, 认知负荷也可能改变反馈者可用的认知资源, 影响其道德判断, 进而影响决策者对人工智能代理的采用及其决策的道德水平。同时, 也可以直接操控决策者对反馈者的认知和行为预期, 以观察对反馈的预期如何调控道德决策。第二, 在相对重要性方面, 未来研究可在保持决策任务不变的条件下, 操控道德决策是否由人工智能代理完成, 并系统调节人工智能的遵从程度、算法透明度(Wang & Qiu, 2024)与拟人化特征(Salminen et al., 2021), 以识别人工智能代理的特殊性效应相对于一般代理效应的强度及其交互机制。此外, 未来研究还需系统探讨人工智能在何种情境下可能产生“道德增强效应” (de Melo et al., 2019; Fernández Domingos et al., 2022), 以形成涵盖道德弱化与强化的双路径框架。在长期利益维度, 可延伸至更多长期社会博弈情境, 例如协调博弈中普遍存在的策略不确定性与协调失灵问题, 从而检验人工智能是否可以降低不确定性、促进高效均衡。在利益诱导维度, 可通过操控道德行为的后果与决策者自身利益的关联程度, 进一步考察利益相关性是否影响人工智能代理的采用倾向, 以及其对不道德行为的影响。

(2)未来研究可在本文决策与归责框架基础上关注人工智能代理的道德影响是否具有外溢效应，以及此外溢效应的产生机制。第一，可以关注局部人群使用人工智能代理如何影响全局社会道德水平，并识别扩散与反馈机制(Alt & Gallier, 2022; Bednar et al., 2025; Engl et al., 2021)。例如，可以采用掷骰子等不诚实行为实验范式，通过操控是否存在人工智能代理、可使用代理的决策者比例(少、中、多)、其他人是否知晓代理的存在等方面，测量局部不道德行为和人工智能代理采用率的传播速度以及社会道德水平。第二，可探究人工智能代理的跨任务或长期影响(Gravert & Collentine, 2021)。例如可以观察参与者在某个决策场景中使用人工智能代理执行道德决策后，在另一决策场景中的决策模式和道德水平。还可以研究人工智能代理互动是否足以重塑参与者的道德决策模式和道德水平。第三，关注当决策者预期到上述外溢效应后，如何调整自身行为。

(3)未来研究还应该关注对人工智能代理及相关不道德行为的治理，特别是人机协同治理。现有研究已提出了“人机共生”的概念，并要求保持人类控制、人类行使最终裁量权等主张(Floridi, 2019; Li et al., 2025; Qin et al., 2024; Russell et al., 2015)。未来研究可探讨人类和人工智能治理手段的介入阶段、介入顺序和各自分工对治理效果的影响。第一，在介入阶段方面，可比较在道德决策前、中和后进行治理的效果，以及不同阶段特定方法(例如：用户事前道德承诺、人工智能伦理原则预设、用户输入留痕审查、人工智能道德熔断机制等)的治理效果。第二，在介入顺序方面，比较不同治理顺序对道德决策的影响，例如在人工智能给出一般性初判后引入人类情境化判断，或在人类建立规范后引入人工智能大规模筛查(Cath, 2018; Wallach & Allen, 2009)。第三，在分工方面，可以利用人类和人工智能各自优势(如人类擅长情境理解与价值判断，而人工智能擅长规律总结与批量审查)，探究二者最优分工结构以及冲突调节方式，最终形成优势互补的协同治理架构。

参考文献

杜传晨, 郑远霞, 郭倩倩, 刘国雄. (2025). 大语言模型的人工心理理论：证据、界定与挑战. *心理科学进展*, *33*(12), 2027−2042.

蒋路远, 曹李梅, 秦昕, 谭玲, 陈晨, 彭小斐. (2022). 人工智能决策的公平感知. *心理科学进展*, *30*(5), 1078−1092.

谭美丽, 殷向洲, 张光磊, 熊普臻. (2025). 工作场所人工智能角色划分：对员工心理与行为的影响及应对策略. *心理科学进展*, *33*(6), 933−947.

王涛, 占小军, 余薇. (2024). AI 感知对员工心理和行为的影响及理论解释. *心理科学进展*, *32(7)*, 1195−1208.

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.

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115−1153.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1−6.

Alnattah, A., Jajroudi, M., Fadafen, S. A. N., Manzari, M. N., & Eslami, S. (2025). Artificial intelligence in clinical decision-making: A scoping review of rule-based systems and their applications in medicine. *Cureus*, *17*(8), e91333.

Alt, M., & Gallier, C. (2022). Incentives and intertemporal behavioral spillovers: A two-period experiment on charitable giving. *Journal of Economic Behavior & Organization, 200*, 959−972.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, *100*(401), 464−477.

Babšek, M., Ravšelj, D., Umek, L., & Aristovnik, A. (2025). Artificial intelligence adoption in public administration: An overview of top-cited articles and practical applications. *AI*, *6*(3), 44.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193−209.

Bankins, S., & Formosa, P. (2023). The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics*, *185*(4), 725−740.

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *Review of Economic Studies*, *79*(1), 67−87.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science, 373*(6552), 284−286.

Bazerman, M. H., & Sezer, O. (2016). Bounded awareness: Implications for ethical decision making. *Organizational Behavior and Human Decision Processes, 136*, 95−105.

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, *82*(6), 1063−1093.

Bednar, J. S., Sommerfeldt, R. D., Zimbelman, A. F., & Zimbelman, M. F. (2025). Don’t sweat the small stuff: The tolerance spillover effect in ethical decision-making. *Journal of Business Ethics, 202*, 727−747.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652−1678.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots:

Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610−623). Association for Computing Machinery.

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

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), Article 4.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). *On the opportunities and risks of foundation models*. Stanford Institute for Human-Centered Artificial Intelligence.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166−193.

Bozdag E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology, 15*(3), 209−227.

Butler, O. (2025). Algorithmic decision-making, delegation and the modern machinery of government. *Oxford Journal of Legal Studies*, *45*(3), 727−752.

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour, 5*(12), 1636−1642.

Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, *9*(1), 1−13.

Calvano, E., Calzolari, G., Denicolò, V., Harrington, J. E., Jr., & Pastorello, S. (2020). Protecting consumers from collusive prices due to AI. *Science*, *370*(6520), 1040−1042.

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior, 134*, Article 107308.

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), Article 20180080.

Chevrier, M., & Teixeira, V. (2024). *Algorithm delegation and responsibility: Shifting blame to the programmer?* (Working paper No. 2024−04). Groupe de REcherche en Droit, Economie, Gestion (GREDEG CNRS), Université Côte d'Azur.

Coffman, L. C. (2011). Intermediation reduces punishment (and reward). *American Economic Journal: Microeconomics*, *3*(4), 77−106.

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, *365*(6448), 70−73.

Constantinescu, M., & Kaptein, M. (2025). Responsibility gaps, LLMs & organisations: Many agents, many levels, and many interactions. *Science and Engineering Ethics, 31*(6), 36.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353−380.

Dai, Z., Galeotti, F., & Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, *64*(3), 1081−1100.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67−80.

De Groot, J. I., Bondy, K., & Schuitema, G. (2021). Listen to others or yourself? The role of personal norms on the effectiveness of social norm interventions to change pro-environmental behavior. *Journal of Environmental Psychology, 78*, Article 101688.

de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human cooperation when acting through autonomous machines. *Proceedings of the National Academy of Sciences*, *116*(9), 3482−3487.

Diaferia, L., Blohm, I., De Rossi, L. M., & Salviotti, G. (2022). When standard is not enough: A conceptualization of AI systems' customization and its antecedents. In *Proceedings of ICIS 2022*. Copenhagen.

Dobbe, R., Gilbert, T. K., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, Article 103555.

Dvorak, F., Stumpf, R., Fehrler, S., & Fischbacher, U. (2024). Generative AI triggers welfare-reducing decisions in humans. *arXiv*. https://arxiv.org/abs/2401.12773.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697−718.

Engl, F., Riedl, A., & Weber, R. (2021). Spillover effects of institutions on cooperative behavior, preferences, and beliefs. *American Economic Journal: Microeconomics, 13*(4), 261−299.

Erat, S. (2013). Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, *93*, 273–278.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63−87.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics*, *108*(2), 437−459.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817−868.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137−140.

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), Article 8492.

Feuerriegel, S., Shrestha, Y. R., von Krogh, G., & Zhang, C. (2022). Bringing artificial intelligence to business management. *Nature Machine Intelligence, 4*(7), 611−613.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525−547.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, *100*(1), 541−556.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185−193.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349−379.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627−660.

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.

Gratch, J., & Fast, N. J. (2022). The power to harm: AI assistants pave the way to unethical behavior. *Current Opinion in Psychology*, *47*, Article 101382.

Gravert, C., & Collentine, L. O. (2021). When nudges aren't enough: Norms, incentives and habit formation in public transport usage. *Journal of Economic Behavior & Organization, 190*, 1−14.

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.

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal–agent relationship. *American Economic Review*, *100*(4), 1826−1846.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence. *California Management Review*, *61*(4), 5−14.

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359−372.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education, 57*(4), 542−570.

Holzmeister, F., Holmén, M., Kirchler, M., Stefan, M., & Wengström, E. (2023). Delegation decisions in finance. *Management Science, 69*(8), 4828−4844.

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*, Article 106944.

Igdalova, A., & Chamberlain, R. (2025). Slow looking at still art: The effect of manipulating audio context and image category on mood and engagement during an online slow looking exercise. *Psychology of Aesthetics, Creativity, and the Arts, 19*(3), 522−534.

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517−521.

Ivcevic, Z., Menges, J., & Miller, A. (2020, March 20). How common is unethical behavior in U.S. organizations? *Harvard Business Review*. https://hbr.org/2020/03/how-common-is-unethical-behavior-in-u-s-organizations

Jago, A. S., Raveendhran, R., Fast, N., & Gratch, J. (2024). Algorithmic management diminishes status: An unintended consequence of using machines to perform social roles. *Journal of Experimental Social Psychology*, *110*, Article 104553.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of Management Review*, *16*(2), 366−395.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience*, *24*(6), Article 102521.

Keding, C. (2021). Understanding the interplay of artificial intelligence and strategic management: Four decades of research in review. *Management Review Quarterly, 71*(1), 91−134.

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

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., … Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature, 646*, 126−134.

Kocher, M. G., Schudy, S., & Spantig, L. (2018). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science, 64*(9), 3995−4008.

Kouchaki, M., & Smith, I. H. (2025). Moral decision-making in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *12*, 45−72.

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.

Li, J., Song, T., Xue, B., & Lee, Y. C. (2025). We shape AI, and thereafter AI shape us: Humans align with AI through social influences. In *Proceedings of the ICLR 2025 Workshop on Bidirectional Human-AI Alignment*. International Conference on Learning Representations (ICLR).

Liehner, G. L., Brauner, P., Schaar, A. K., & Ziefle, M. (2021). Delegation of moral tasks to automated agents− The impact of risk and context on trusting a machine to perform a task. *IEEE Transactions on Technology and Society*, *3*(1), 46−57.

Liu, Y., Dai, C., & Cao, Y. (2025). Cognitive load and moral

decision-making in moral dilemmas under virtual reality: The role of empathy for pain. *Current Psychology, 44*(6), 5279−5297.

Makovi, K., Bonnefon, J. F., Oudah, M., Sargsyan, A., & Rahwan, T. (2025). Rewards and punishments help humans overcome biases against cooperation partners assumed to be machines. *iScience, 28*(7), Article 112833.

Malik, N., Tripathi, S. N., Kar, A. K., & Gupta, S. (2022). Impact of artificial intelligence on employees working in industry 4.0 led organizations. *International Journal of Manpower*, *43*(2), 334−354.

Malle, B. F. (2021). Moral judgments. *Annual Review of psychology, 72*(1), 293−318.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117−124). Association for Computing Machinery.

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

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633−644.

Milgram, S. (1974). *Obedience to authority: An experimental view*. Harper & Row.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), Article 2053951716679679.

Moser, C., Den Hond, F., & Lindebaum, D. (2022). Morality in the age of artificially intelligent algorithms. *Academy of Management Learning & Education*, *21*(1), 139−155.

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

Oexl, R., & Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics*, *16*(3), 306−312.

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes, 109*(2), 134−141.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, *30*(3), 286−297.

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes, 68*(3), 208−224.

Qin, X., Lu, J. G., Chen, C., Zhou, X., Gan, Y., Li, W., & Song, L. L. (2024). *Artificial intelligence quotient (AIQ)*. Available at SSRN 4787320.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review, 83*(5), 1281−1302.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477−486.

Ram, J. (2025). Moral decision-making in AI: A comprehensive review and recommendations. *Technological Forecasting and Social Change*, *217*, Article 124150.

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *The American Economic Review, 63*(2), 134−139.

Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of artificial intelligence. *Nature*, *521*(7553), 415−418.

Salminen, J., Kaate, I., Kamel, A. M. S., Jung, S. G., & Jansen, B. J. (2021). How does personification impact ad performance and empathy? An experiment with online advertising. *International Journal of Human–Computer Interaction, 37*(2), 141−155.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057−1084.

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2026). AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion, 126*, Article 103599. https://doi.org/10.1016/j.inffus.2025.103599

Schneider, S., & Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. *Managerial and Decision Economics*, *40*(3), 223−231.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199−2200.

Simon, H. A. (1997). Models of bounded rationality: *Empirically grounded economic reason* (Vol. 3). MIT Press.

Steffel, M., Williams, E. F., & Perrmann-Graham, J. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes*, *135*, 32−44.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73−100). MIT Press.

Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics*, *178*(4), 917−943.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124−1131.

von Bieberstein, F., Feess, E., & Packham, N. (2026). Multi-step delegation and the frequency of immoral decisions: Theory and experiment. *European Economic Review, 181*, Article 105159.

Villegas-Galaviz, C., & Martin, K. (2024). Moral distance, AI, and the ethics of care. *AI & Society, 39*(4), 1695−1706.

Vu, L., Soraperra, I., Leib, M., van der Weele, J., & Shalvi, S. (2023). Ignorance by choice: A meta-analytic review of willful ignorance. *Psychological Bulletin*, *149*(9−10), 611−635.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wang, X., & Qiu, X. (2024). The positive effect of artificial intelligence technology transparency on digital endorsers: Based on the theory of mind perception. *Journal of Retailing and Consumer Services, 78*, Article 103777.

Wang, X., Xie, H., Wang, Y., Xiao, C., Chen, H., Sargeant, H., ... Sun, M. (2025). Large language models' complicit responses to illicit instructions across socio-legal contexts. *arXiv*. https://arxiv.org/abs/2511.20736

Weiss, A., & Forstmann, M. (2024). Religiosity predicts the delegation of decisions between moral and self-serving immoral outcomes. *Journal of Experimental Social Psychology*, *113*, Article 104605.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., ... Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity, 2021*(1), Article 8812542.

Zhang, C., Zhu, W., Ding, J., Wu, Y., & Chen, X. (2023). Ethical impact of artificial intelligence in managerial accounting. *International Journal of Accounting Information Systems*, *49*, Article 100619.

Zhao, W., Su, K., Zhu, H., Kaiser, M., Fan, M., Zou, Y., ... Yin, D. (2024). Activity flow under the manipulation of cognitive load and training. *Neuroimage, 297*, Article 120761.

Zhou, J., Corbett, F., Byun, J., Porat, T., & van Zalk, N. (2025). Psychological and behavioural responses in human-agent vs. human-human interactions: A systematic review and meta-analysis. *arXiv*. https://doi.org/10.48550/arXiv.2509.21542

The moral impact of delegating to artificial intelligence

TANG Wei[1], ZHONG Wenrui[2], LEI Zhen[2], ZHANG Dandan[2,3]

([1] *Institute of Xi Jinping's Economic Thought*, *Southwestern University of Finance and Economics*, *Chengdu* 611130, *China*) ([2] *China Center for Behavioral Economics and Finance*, *Southwestern University of Finance and Economics*, *Chengdu* 611130, *China*) ([3] *Institute of Brain and Psychological Sciences*, *Sichuan Normal University*, *Chengdu* 610066, *China*)

Abstract: Various artificial intelligence (AI) agents are increasingly embedded in the decision processes of firms, governments, and individuals, taking on delegated decision execution and exerting profound influence on the morality of human decisions and moral judgment. Despite the rapid growth of related empirical and theoretical research, the existing literature still lacks a systematic analysis of what makes AI agents distinctive from other types of agents, and it also lacks a unified analytical framework to systematically characterize the pathways through which AI agents influence moral decision-making. Therefore, this paper develops a "decision-maker–agent–evaluator" framework of decision-making and accountability to synthesize and reorganize the existing literature. We argue that when an agent intervenes in the decision process, it lengthens both the decision-maker's decision chain and the evaluator's feedback chain (where evaluators include affected parties and third-party observers). This, in turn, weakens decision-makers' moral salience and evaluators' attribution of responsibility, thereby facilitating unethical behavior. Moreover, distinctive features of AI agents−such as opacity (black-boxness), high compliance, scalability, and instrumentality−further intensify the execution of unethical instructions, increase decision-makers' opportunities for deniability, and expand the scope of unethical impacts along the decision chain. At the same time, these features can increase evaluators' moral tolerance for unethical outcomes and blur their inferences about the decision-maker's intent and responsibility along the feedback chain, further encouraging unethical behavior. Finally, we suggest that future research should refine the relative roles of the mechanisms within this framework, examine how unethical behavior diffuses and scales up at organizational and societal levels, and explore governance tools and institutional arrangements for human–AI collaboration.

Keywords: artificial intelligence delegation, moral decision-making, human–AI collaborative governance