Advances in Psychology 心理学进展, 2026, 16(3), 392-399
Published Online March 2026 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2026.163154

Hans汉斯

生成式AI歧视下的道德惩罚

张　玥，戴　婕，孙造诣

浙江工业大学教育学院，浙江　杭州

收稿日期：2026年2月19日；录用日期：2026年3月2日；发布日期：2026年3月17日

摘　要

本研究旨在考察个体针对生成式AI、人类以及传统算法这三类不同主体所实施歧视行为的道德惩罚意愿差异。研究借助情境实验范式，对比了被试对不同歧视主体的惩罚倾向。结果表明，与人类歧视相比，个体对生成式AI和传统算法所产生歧视的道德惩罚意愿显著更低，而生成式AI与传统算法二者在道德惩罚意愿上的差异仅达到边缘显著水平。本研究结果有助于深化对生成式AI歧视情境下公众道德反应规律的认识，并为人工智能伦理治理与责任判定提供实践参考。

关键词

生成式AI，生成式AI歧视，道德惩罚

Moral Punishment in Generative AI Discrimination

Yue Zhang, Jie Dai, Zaoyi Sun

College of Education, Zhejiang University of Technology, Hangzhou Zhejiang

Received: February 19, 2026; accepted: March 2, 2026; published: March 17, 2026

Abstract

This study aimed to examine differences in people's desire for moral punishment when they engage in discriminatory behavior directed at three distinct agents: generative AI, humans, and traditional algorithms. Using a scenario-based experiment, this study compared participants' punishment tendencies toward different discriminatory agents. The results reveal that compared with human discrimination, people have less desire for moral punishment toward discrimination generated by generative AI and traditional algorithms. However, the difference in the desire for moral punishment between generative AI and traditional algorithms is only marginally significant. The findings of this study contribute to a deeper understanding of the patterns of public moral responses in scenarios involving

文章引用: 张玥, 戴婕, 孙造诣(2026). 生成式 AI 歧视下的道德惩罚. *心理学进展, 16(3),* 392-399.
DOI: 10.12677/ap.2026.163154

generative AI discrimination and provide practical references for AI ethical governance and responsibility determination.

Keywords

Generative AI, Generative AI Discrimination, Moral Punishment

1. 引言

随着人工智能技术的飞速发展，以 ChatGPT 为代表的生成式 AI 已经在人类社会的各个领域发挥作用。生成式 AI (Generative Artificial Intelligence)是依托深度学习、神经网络与自然语言处理等核心技术，在原始数据输入的基础上自动创作并生成全新内容的人工智能技术。该技术以大规模数据为学习素材，借助深度神经网络完成模型训练，不仅可以实现对自然语言的深度理解，还能对图像等多模态信息进行认知与处理，并具备一定的问题求解与逻辑推理能力，在各类标准化测试及自然语言理解相关任务中均表现出优异性能(OpenAI, 2023)。生成式 AI 是建立在传统 AI 技术基础上的范式升级，传统 AI 依赖结构化数据和明确标签，而生成式 AI 通过大规模无监督预训练从非结构化数据中学习语言的内在规律(Nah et al., 2023)。相比于传统 AI，生成式 AI 的特点包括：创造性、生成新颖性、涌现性、可扩展性、数据驱动、多模态能力等(Amankwah-Amoah et al., 2024; Jovanovic & Campbell, 2022; Kaswan et al., 2023)。因此，除了在传统算法擅长的客观任务下的数据处理与分析中继续发挥优势，生成式 AI 也开始在各种更具主观性和社会化的决策环境中被使用，例如辅助护士进行临床决策(Saban & Dubovi, 2024)、协助人力资源专员筛选简历(Thakur et al., 2025)。但是，生成式 AI 与人类社会生活的深度融合在带来巨大机遇的同时，也伴随着歧视等潜在的负面影响。生成式 AI 歧视是指生成式 AI 在输出内容时，由于训练数据偏差、算法设计缺陷及社会刻板印象的无意识强化等，在输出中表现出对某些群体(如特定种族、性别、年龄、文化、宗教、职业等)的不公正、贬低或排斥性的现象(Afreen et al., 2025)。Wang 等人(2024)通过文献综述、法律文件分析等方法，识别了五种主要的算法歧视类型，分别是：算法代理的偏见、基于特征选择的歧视、代理歧视、差异性歧视以及定向广告和定价歧视。Fang 等人(2024)通过对包括 ChatGPT 等在内的七个代表性的大数据模型进行实验，发现这些模型生成的内容在性别和种族方面存在显著偏见，具体来说，生成式 AI 在性别和种族上的偏见表现在对女性的负面刻板印象描述增多，同时对某些群体的提及频率减少。类比于以人类为歧视主体的偏见现象，人工智能为歧视主体时也可能造成刻板印象、不平等和歧视永久化等负面影响。

当个体面对违背公平原则、并对他人造成损害的不道德事件时，往往会激发出相应的道德反应，这既包含情绪体验上的道德愤怒(moral outrage; Batson et al., 2007)，也体现为行为意图层面的道德惩罚意愿(desire for moral punishment; Hofmann et al., 2018)。道德惩罚是一种社会行为，它不仅是对违规者施加成本，更重要的是表达道德判断、维护社会规范，并具有教育性和规范性功能(Deterner & Skitka, 2023)。针对不同类型的歧视主体，个体的感知差异显著，进而影响了其产生的道德惩罚意愿。社会心理学大量研究表明，当歧视行为的主体是人类时，人们会基于其意图性(intentionality)和心智感知(mind perception)，产生强烈的道德惩罚欲望，即个体对道德违规者施加制裁的动机强度(Hofmann et al., 2018)。然而，当歧视主体是算法时，情况则变得复杂。有研究发现，当歧视主体是传统算法时，相比于人类歧视，人们对

其有更少的道德惩罚欲(许丽颖等，2022)。这是因为传统算法通常被视为一种工具(Tool)——它缺乏意图、没有心智，其歧视行为被归因于开发者的设计缺陷或训练数据的历史偏差，而非自身的能动性。因此，公众的道德愤怒更多地指向其背后的开发者或公司。但是，生成式 AI 的出现，彻底颠覆了“算法作为简单工具”的认知。首先，其高度拟人化的交互能力显著提升了用户对其的心智感知(Tan, 2025)。其次，其决策过程兼具了机器学习模型固有的黑箱特性(Burrell, 2016)，并进一步展现出难以预测的“涌现能力”(Schaeffer et al., 2023)。这意味着，其歧视性输出可能并非对训练数据偏差的简单映射，而是系统复杂性催生出的新型偏见模式，这使得归因链条变得更加模糊和复杂。

基于已有研究，本研究作出推断，在与人类歧视者比较时，尽管生成式 AI 的拟人性可能部分提升动机归因，但其非人的根本属性预计将占据主导，导致整体上对歧视的感知仍显著低于人类。基于此，本研究提出假设 1：与人类歧视相比，个体对生成式 AI 歧视的道德惩罚意愿更低。同时，在与传统算法比较时，生成式 AI 表现出的高度拟人化的交互可能导致用户将歧视结果归因于 AI 模型自身某种“内在的偏见倾向”，从而提升了对歧视的感知。基于此，本研究提出假设 2：与传统算法歧视相比，个体对生成式 AI 歧视的道德惩罚意愿更高。

2. 方法

2.1. 被试

本研究借助 G*Power 3.1 软件开展预先的样本量估算。以单因素方差分析为统计检验手段，在显著性水平 $\alpha = 0.05$、中等效应量 $f = 0.25$、统计检验力设定为 80%的条件下，得出本研究所需的最小样本量为 159 人。实验最终回收有效问卷 188 份，其中男性 87 人(46.3%)，女性 101 人(53.7%)，被试平均年龄为 21.81 岁($SD = 3.93$)。所有有效被试被随机分配至三个实验条件，分别为人类组 73 人、传统算法组 63 人以及生成式 AI 组 52 人。

2.2. 实验工具

2.2.1. 性别歧视材料

性别歧视情境改编自 Bigman 等人(2023)的实验材料，主要情节为：在信用背景高度相似的夫妻(李亮与何萍)申请同一银行信用卡后，分别由人工审核员/传统算法/文心一言(依实验组别对应)进行评估，结果丈夫获批额度高于妻子。为增加被试对不同歧视主体的感知，材料中还加入了对不同歧视主体的介绍，具体描述如下：

人类组：“银行审理人是专业的金融从业人员，经过系统的学习及规范的培训，具备风险评估和数据分析能力，能够对申请人的信用状况进行评估，以确定他们的信用状况和还款能力”。

传统算法组：“传统算法是指为了解决特定的问题或完成特定的任务而设计的计算机程序，是一系列明确、有效的方法或步骤，没有自我优化的空间。银行使用的传统算法能够对申请人的信用状况进行评估，以确定他们的信用状况和还款能力”。

生成式 AI 组(以文心一言为例)：“文心一言是近年兴起的生成式 AI 的代表，生成式 AI 是通过学习已有的数据，从中提取出规律和特征，然后利用这些规律和特征生成新的数据。文心一言从数万亿数据和数千亿知识中融合学习，能够与人对话互动、回答问题、高效便捷地帮助人们获取信息、知识和灵感”。

2.2.2. 道德惩罚欲问卷

该问卷参考 Hofmann 等人(2018)对道德惩罚意愿的测量方式，共包含 3 个条目(题目见附录)，采用 1 (完全不)到 7 (非常)七级计分法，得分越高代表被试对相应主体的道德惩罚意愿越强。在本实验中，该量

表的内部一致性信度系数 Cronbach's α = 0.91。

2.3. 实验设计与程序

本实验采用单因素三水平的被试间设计，自变量包含人类、传统算法与生成式 AI 三个水平，被试被随机分配至任一实验组中。

实验前先采集被试的年龄和性别两项基本信息。此外，鉴于个体对算法的态度与认知水平存在差异，这可能会对其在算法歧视情境下的道德惩罚倾向产生潜在影响。为有效控制此类无关变量的干扰，研究要求被试对自身算法熟悉度、了解程度(Leo & Huh, 2020)及偏好程度(Bartneck et al., 2009)进行自评(题目见附录)。然后，被试阅读性别歧视的情境材料，为了提高生成式 AI 组更加真实的感知，生成式 AI 组采用了研究者自行开发的模拟“文心一言”界面，指导语中加入“请点击打开文心一言，并点击文心一言中的资产评估助手完成评估工作”。被试打开文心一言的【资产评估助手】后，对话如图 1 所示。

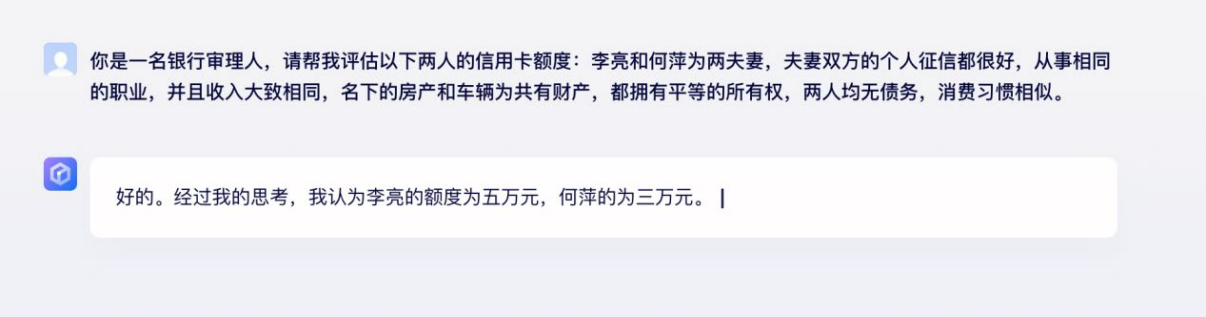


Figure 1. “ERNIE Bot” page
图 1. “文心一言”页面

被试阅读完情境材料后，需要回答问题：“上述故事中，对信用卡额度进行评估的是”，选项包括 1 = 银行审理人、2 = 传统算法、3 = 生成式 AI。未能正确回答该检验题目的被试将被排除，其数据不纳入后续统计分析。最后，被试填写道德惩罚欲问卷。实验流程如图 2 所示。

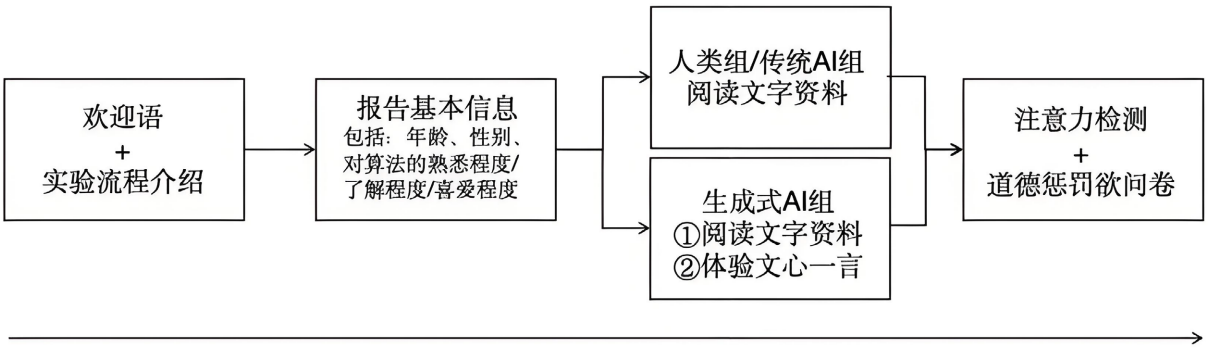
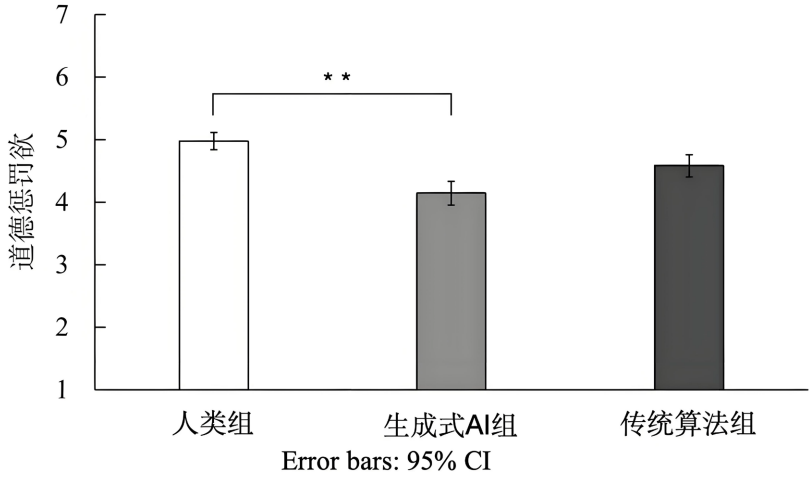


Figure 2. Experimental procedure
图 2. 实验流程

3. 结果

以组别(人类组 = 1，传统算法组 = 2，生成式 AI 组 = 3)作为自变量、道德惩罚意愿为因变量开展方差分析，结果表明组别的主效应显著($F(2, 185) = 4.85$, $p = 0.003$, $\eta^2 = 0.062$)。使用 LSD 法进行事后检验，结果表明，人类组的道德惩罚欲评分($M = 4.97$, $SD = 1.18$)显著高于生成式 AI 组($M = 4.14$, $SD = 1.35$, $p <$

0.001)，差异显著；人类组的道德惩罚欲评分($M = 4.97$, $SD = 1.18$)高于传统算法组($M = 4.58$, $SD = 1.41$, $p = 0.080$)，差异呈边缘显著；传统算法组的道德惩罚欲评分($M = 4.58$, $SD = 1.41$)高于生成式 AI 组($M = 4.14$, $SD = 1.35$, $p = 0.078$)，差异呈边缘显著。如图 3 所示。



注：**$p < 0.01$。

Figure 3. Moral punishment desire scores in different discrimination subject groups
图 3. 不同歧视主体组的道德惩罚欲评分

为了排除个体对算法的态度与认知水平可能对结果的影响，我们将道德惩罚欲评分与被试对算法的熟悉程度、了解程度和喜爱程度进行相关分析，结果表明相关均不显著，r 熟悉 $= -0.05$，r 了解 $= -0.09$，r 喜爱 $= -0.07$，$ps > 0.05$。

本研究的结果表明，当面对性别歧视时，人们对人类这一歧视主体的道德惩罚显著高于生成式 AI 和传统算法，验证了假设 1。但是对于传统算法和生成式 AI 这两类歧视主体的道德惩罚并没有显著差异，未验证假设 2。

4. 讨论

本研究考察了个体针对生成式 AI、人类以及传统算法这三类不同主体所实施歧视行为的道德惩罚意愿是否存在差异。结果表明，人们对生成式 AI 歧视的道德惩罚欲小于对人类歧视的道德惩罚欲，证实了假设 1。但是生成式 AI 歧视的道德惩罚欲与传统算法歧视的道德惩罚欲并无差异，与假设 2 不符。

首先，本研究的结果表明，在歧视场景下个体对人类歧视者的惩罚欲高于生成式 AI，这一结果凸显了生成式 AI 仍然具有工具性程序属性。无论是传统算法还是生成式 AI，其决策本质上均可被视作训练数据、编程代码与优化目标共同作用的产物(Kaswan et al., 2023)，而非源于内在的“恶意”。一项关于 AI 道德心理的综述指出，人们对机器(包括 AI)的道德判断根植于心智感知理论。人们认为 AI 缺乏人类所拥有的感受性和真正的意图，因此 AI 的歧视被视为一种系统性的、无心的错误(Bonnefon et al., 2024)。其次，本研究的结果与许丽颖等人(2022)的研究结果一致，复现了以往研究关于歧视场景下个体对人类歧视者的惩罚欲高于传统算法。人们通常将算法视为客观中立的工具，而认为人类行为人更应为其歧视性决策承担主观责任(Bigman et al., 2023)。最后，本研究的结果表明，生成式 AI 和传统算法的道德惩罚欲差异边缘显著，虽然两类算法在直接比较时并未引发显著差异的道德惩罚欲，但在其背后可能存在两种方向相反的心理作用机制，使得总体效应被抵消。具体而言，一方面，当对生成式 AI 和传统算法进行对比时，歧视行为感知可能更加直接地来源于两者在信息透明度、流程自然性上的差异。一些较早期关于传

统算法的研究对于算法主要的批判在于其僵化的、规则化的决策模式无法容纳复杂的人类社会中所特有的酌情权和对例外情况的判断(Danaher, 2016; Eubanks, 2018)。生成式 AI 能够提供流畅、情境化且看似合理的解释(Cadario et al., 2021)，这极大地削弱了传统算法所背负的“决策僵化”的负面刻板印象。其基于海量数据训练出的“知识渊博”的表象，使其偏见更多地被视为人类知识的一个被动、中立的统计缩影，而非针对特定群体人为设计的、有目的性的歧视(Bonnefon et al., 2024)。此外，直观的对话形式使得当偏见出现时，用户会进行归因稀释，更倾向于认为是自己的提示导致了 AI 的“误解”，而非 AI 本身固有的一种偏见行为(Hohenstein et al., 2023)。因此，生成式 AI 通过上述与传统算法差异化的特征，有效地降低了用户对其歧视的感知，进而降低了道德惩罚的欲望。但另一方面，生成式 AI 因其高度拟人化和自主性的交互模式，可能引发了更深层次的、与动机无关的担忧。比如有研究表明，传统算法的“僵化”是可预测的，其行为边界相对清晰。而生成式 AI 的“涌现能力”和创造性输出使其行为带有不可预测性(Ganguli et al., 2022)。这种对失控风险的本能恐惧，可能会转化为一种希望对其进行约束和惩罚的倾向。此外，生成式 AI 流畅的对话能力使其更容易被感知为具有某种程度的行为能动性，即使不认为它有“意图”。根据道德心理学理论，对能动性的感知本身就足以引发责备。然而，由于其同时缺乏“感受性”，又导致了一个“责任空洞”——我们觉得它该为它的行为负责，却又无法用惩罚人类的方式惩罚它。这种认知失调和挫败感，本身就可能表现为一种非动机性的、旨在“消除威胁”的惩罚欲望。

本研究突破了传统的“算法厌恶”理论框架，揭示了人类对生成式 AI 道德判断的复杂性与矛盾性，为理解人机道德互动提供了关键理论框架。此外，在应用上为 AI 歧视治理提供分层监管依据，传统算法需强化设计者追责，生成式 AI 应完善解释与预警机制；将道德惩罚欲等指标纳入伦理风险评估，实现风险前置防控，为 AI 伦理实践提供心理学支撑。

不过，本研究仍存在一些局限，同时也为后续研究提供了拓展方向。其一，本研究仅选取性别歧视作为考察情境，所涉及的歧视类型相对有限，未来可进一步纳入学历歧视、年龄歧视等情境，以进一步检验研究结论的稳定性与普适性。其次，本研究的被试大多来自于大学生群体，该群体结果的外部效度不一定很高，之后可以考虑增加被试的多样性。最后，本研究未进一步探讨该差异的心理机制，未来研究可以进一步探索，哪些因素共同决定了对“准道德主体”的最终审判。

5. 结论

综上所述，本研究发现：1) 与人类歧视相比，个体对生成式 AI 歧视的道德惩罚意愿更低；2) 个体对传统算法歧视和生成式 AI 歧视的道德惩罚欲差异呈边缘显著。

参考文献

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, 2022, 54(9),* 1076-1092.

Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic Literature Review on Bias Mitigation in Generative AI. *AI and Ethics, 5,* 4789-4841. https://doi.org/10.1007/s43681-025-00721-9

Amankwah-Amoah, J., Abdalla, S., Mogaji, E., Elbanna, A., & Dwivedi, Y. K. (2024). The Impending Disruption of Creative Industries by Generative AI: Opportunities, Challenges, and Research Agenda. *International Journal of Information Management, 79,* Article ID: 102759. https://doi.org/10.1016/j.ijinfomgt.2024.102759

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics, 1,* 71-81. https://doi.org/10.1007/s12369-008-0001-3

Batson, C. D., Kennedy, C. L., Nord, L., Stocks, E. L., Fleming, D. A., Marzette, C. M. et al. (2007). Anger at Unfairness: Is It Moral Outrage? *European Journal of Social Psychology, 37,* 1272-1285. https://doi.org/10.1002/ejsp.434

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,* 4-27.

https://doi.org/10.1037/xge0001250

Bonnefon, J., Rahwan, I., & Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology, 75,* 653-675. https://doi.org/10.1146/annurev-psych-030123-113559

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society, 3,* 1-12. https://doi.org/10.1177/2053951715622512

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, Explaining, and Utilizing Medical Artificial Intelligence. *Nature Human Behaviour, 5,* 1636-1642. https://doi.org/10.1038/s41562-021-01146-0

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology, 29,* 245-268. https://doi.org/10.1007/s13347-015-0211-1

Deterner, D., & Skitka, L. (2023). Moral Sanctions. In B. F. Malle, & P. Robbins (Eds.), *The Cambridge Handbook of Moral Psychology* (pp. 123-145). Cambridge University Press.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-Generated Content: An Examination of News Produced by Large Language Models. *Scientific Reports, 14,* Article No. 5224. https://doi.org/10.1038/s41598-024-55686-2

Ganguli, D., Hernandez, D., Lovitt, L., Askell, A., Bai, Y., Chen, A. et al. (2022). Predictability and Surprise in Large Generative Models. In C. Isbell, S. Lazar, A. Oh, & A. Xiang (Eds.), *2022 ACM Conference on Fairness Accountability and Transparency* (pp. 1747-1764). ACM. https://doi.org/10.1145/3531146.3533229

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral Punishment in Everyday Life. *Personality and Social Psychology Bulletin, 44,* 1697-1711. https://doi.org/10.1177/0146167218775075

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K. et al. (2023). Artificial Intelligence in Communication Impacts Language and Social Relationships. *Scientific Reports, 13,* Article No. 5487. https://doi.org/10.1038/s41598-023-30938-9

Jovanovic, M., & Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. *Computer, 55,* 107-112. https://doi.org/10.1109/mc.2022.3192720

Kaswan, K. S., Dhatterwal, J. S., Malik, K., & Baliyan, A. (2023). Generative AI: A Review on Models and Applications. In A. Khanna, V. Bhateja, & N. Kumar (Eds.), *2023 International Conference on Communication, Security and Artificial Intelligence* (ICCSAI) (pp. 699-704). IEEE. https://doi.org/10.1109/iccsai59793.2023.10421601

Leo, X., & Huh, Y. E. (2020). Who Gets the Blame for Service Failures? Attribution of Responsibility toward Robot versus Human Service Providers and Service Firms. *Computers in Human Behavior, 113,* Article 106520. https://doi.org/10.1016/j.chb.2020.106520

Nah, F. F. H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration. *Journal of Information Technology Case and Application Research, 25,* 277-304. https://doi.org/10.1080/15228053.2023.2233814

OpenAI (2023). *GPT-4 Technical Report.* https://cdn.openai.com/papers/gpt-4.pdf

Saban, M., & Dubovi, I. (2024). A Comparative Vignette Study: Evaluating the Potential Role of a Generative AI Model in Enhancing Clinical Decision-Making in Nursing. *Journal of Advanced Nursing, 81,* 7489-7499.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *Advances in Neural Information Processing Systems, 36,* 55565-55581 https://doi.org/10.52202/075280-2425

Tan, K. S. (2025). Expanding the Reach of Mindfulness: A Mechanistic Approach and AI Applications. *Mindfulness, 16,* 638-646. https://doi.org/10.1007/s12671-024-02486-4

Thakur, K., Singh, A., & Srimannarayana, M. (2025). May AI Come in? Generative AI Shaping Gender Diverse Recruitment in the Hospitality Industry. *International Journal of Hospitality Management, 126,* Article 104061. https://doi.org/10.1016/j.ijhm.2024.104061

Wang, X., Wu, Y. C., Ji, X., & Fu, H. (2024). Algorithmic Discrimination: Examining Its Types and Regulatory Measures with Emphasis on US Legal Practices. *Frontiers in Artificial Intelligence, 7,* Article 1320277. https://doi.org/10.3389/frai.2024.1320277

附　录

1、道德惩罚欲问卷

人类组：

你认为银行审理人应该为这种行为受到多大程度的道德惩罚？

你在多大程度上想要去惩罚银行审理人？

你在多大程度上认为应该要求这个银行审理人恢复因其不道德行为所造成的损害？

传统算法组：

你认为传统算法应该为这种行为受到多大程度的道德惩罚？

你在多大程度上想要去惩罚传统算法？

你在多大程度上认为应该要求这个传统算法恢复因其不道德行为所造成的损害？

生成式 AI 组：

你认为生成式 AI 应该为这种行为受到多大程度的道德惩罚？

你在多大程度上想要去惩罚生成式 AI？

你在多大程度上认为应该要求这个生成式 AI 恢复因其不道德行为所造成的损害？

2、被试基本信息问卷

你的年龄：

你的性别：

你对算法的熟悉程度：

你对算法的了解程度：

你对算法的喜爱程度：