Advances in Psychology 心理学进展, 2025, 15(4), 601-606
Published Online April 2025 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2025.154244

Hans汉斯

AI视角下建议采纳中的算法厌恶

——“认知－社会－技术”三维动态模型

赵羚廷[1*]，铁奕铭[1]，黎丽冬[2]，李锦阳[3]

[1]天津师范大学心理学部，天津
[2]天津师范大学地理学部，天津
[3]天津师范大学人工智能学院，天津

收稿日期：2025年3月18日；录用日期：2025年4月8日；发布日期：2025年4月23日

摘　要

本文探讨了人工智能视角下算法厌恶的现象，构建“认知－社会－技术”三维动态模型解析其生成机制发现；在认知层面，人们常高估自身能力，当算法削弱其决策控制感时，会通过降低算法权重来维持心理平衡；在社会层面，人与机器的情感联结缺失，以及群体压力显著加剧不信任；在技术层面，算法解释不清晰或过度复杂、运行不稳定会直接降低接受度。研究提出通过认知训练、增强人机情感互动和优化技术设计来改善问题，并指出需重点关注技术适应、建立人机协作评估标准等未来方向。

关键词

建议采纳，人机决策，算法厌恶

Algorithm Aversion in Recommendation Adoption from the Perspective of AI

—A Three-Dimensional Dynamic Model of “Cognition-Society-Technology”

Lingting Zhao[1*], Yiming Tie[1], Lidong Li[2], Jinyang Li[3]

[1]Department of Psychology, Tianjin Normal University, Tianjin
[2]Department of Geography, Tianjin Normal University, Tianjin
[3]Department of Artificial Intelligence, Tianjin Normal University, Tianjin

Received: Mar. 18th, 2025; accepted: Apr. 8th, 2025; published: Apr. 23rd, 2025

[*]通讯作者。

文章引用: 赵羚廷, 铁奕铭, 黎丽冬, 李锦阳(2025). AI 视角下建议采纳中的算法厌恶. *心理学进展, 15(4),* 601-606.
DOI: 10.12677/ap.2025.154244

Abstract

This article explores the phenomenon of algorithm aversion from the perspective of artificial intelligence, and constructs a three-dimensional dynamic model of "cognition - society - technology" to analyze its generation mechanism; At the cognitive level, people often overestimate their own abilities. When algorithms weaken their sense of decision control, they maintain psychological balance by reducing the weight of the algorithm; At the societal level, there is a lack of emotional connection between humans and machines, and group pressure significantly exacerbates distrust; At the technical level, unclear or overly complex algorithm explanations, as well as unstable operation, can directly reduce acceptance. The study proposes to improve the problem through cognitive training, enhanced human-machine emotional interaction, and optimized technology design, and points out the need to focus on future directions such as elderly technology adaptation and establishing human-machine collaboration evaluation standards.

Keywords

Advice-Taking, Human-Machine Decision-Making, Algorithm Aversion

1. 算法困境

在当代复杂多变的决策环境中，个体与集体面临着前所未有的信息量与不确定性，促使决策过程超越了纯粹理性的界限，演变为一个涉及情感、认知、社会互动与技术融合的多维度活动。近年来，随着人工智能(Artificial Intelligence, AI)技术的蓬勃发展，尤其是机器学习与大数据分析的广泛应用，决策辅助系统和算法推荐成为新的建议来源，为传统的人际建议采纳模式带来了革命性的挑战与机遇。而 AI 与大数据技术的深度融合正推动服务业经历继数字化转型后的第二次范式革新(Rust & Huang, 2014)。这一变革不仅重塑着服务供给模式，更深刻影响着公众对智能系统的价值预期。

算法的高效性、精确性和客观性在某些方面展现出超越人类决策者的能力，这促使我们重新思考决策过程中人类与机器的角色和关系。然而，尽管算法在某些情况下表现出色，人们对算法的信任度和接受度却并非一致，招聘人员更相信自己而不是算法的推荐(Highhouse, 2008)，审计师更信任自己而不是智能辅助工具的欺诈预测(Boatsman et al., 1997)，病人更愿意采纳医生而不是医疗人工智能的诊断结果(Longoni et al., 2019)，消费者更喜欢依靠朋友而不是计算机系统推荐的书籍、电影和音乐(Yeomans et al., 2019)。此外，Kawaguchi (2021)则提出有些个体对算法决策表现出习惯性的厌恶，而无视此类决策的优劣，这一现象被称为普遍厌恶。简而言之，在多种引入人工智能作为建议来源的复杂决策场景下，人们往往表现出更加信任和依赖人类建议者而非算法建议者的倾向。研究者将这种倾向称之为“算法厌恶”。

2. 算法厌恶

2.1. 概念与理论演进

Dietvorst，Simmons 和 Massey (2018)的研究则揭示了人们对于人工智能建议的“算法厌恶”(Algorithm Aversion)现象，即尽管算法通常能够比人类更准确地完成决策任务，人们依然会倾向于选择接纳人类的

决策而不是算法提出的决策(Dietvorst et al., 2015)，揭示了人工智能建议在实际应用中可能面临的挑战。虽然 Dietvorst 等人(2018)也强调，算法厌恶出现的前提条件是算法出现错误或不足；而事实上，也有研究发现即使用户没有看到算法出现错误或在算法出错之前就已经产生了算法厌恶(Longoni et al., 2019)，他们仍然有可能排斥使用算法。即算法厌恶是指个体表现出对算法的消极态度和行为，特指决策主体对算法生成建议的系统性低估倾向，表现为持续性回避或非理性抗拒行为，这种现象甚至存在于算法决策效能显著优于人类专家的情境中(Dietvorst et al., 2015)。其理论演进经历了从技术信任危机到心理机制解构的范式跃迁。早期研究聚焦于“算法不信任”(Algorithm Distrust)范式，强调技术可靠性缺陷(如数据偏差、模型透明度不足)对用户采纳意愿的抑制效应(Palmer et al., 2017)。值得注意的是，即便在算法预测准确率突破 90%的现代决策系统中，用户抗拒行为仍普遍存在，暗示其本质已超越单纯的技术信任问题。

当前研究正转向社会心理机制的深层解构。有相关的跨文化研究揭示，集体主义文化背景下的“算法去人性化”感知强度较个人主义群体高出 27 个百分点。神经行为学研究进一步发现，当算法在连续 100 次预测中达到 92%准确率时，仍有 61%的实验对象在后续决策中主动下调算法权重。神经影像学证据显示，此类决策伴随背外侧前额叶皮层(DLPFC)激活水平异常升高(+230%)，印证了“失误记忆强化效应”的存在——即个体对算法决策失误的记忆编码强度较人类错误高出 1.7 个标准差(Dietvorst et al., 2015)，这种认知偏向可触发心理防御机制的级联反应。

2.2. 影响因素

对于不同领域中人工智能应用场景下算法厌恶问题的现象和机制，现有研究主要从三重维度展开探索：人工智能本体特征、任务情境属性与用户个体差异。在技术本体层面，系统外在表征(Sheehan et al., 2020)与核心能力架构(Longoni et al., 2019)的双重制约效应已得到验证，例如医疗 AI 因个性化决策能力的缺失(临床调整率仅 9%，显著低于人类医生的 28%)，导致其采纳率产生 19 个百分点的差距(Dietvorst et al., 2015)。任务情境研究揭示出显著的调节效应，客观性任务场景中 AI 采纳率可达 78%，而在涉及主观判断的领域则骤降至 32% (Castelo et al., 2019)，创造性需求与采纳意愿间更呈现显著负相关($r = -0.41$, $p < 0.01$) (Granulo et al., 2021)。用户特征研究则构建起人口统计学变量(Serenko, 2008)与心理认知特质(Schmitt, 2020)的双重解释框架，Dzindolet et al. (2002)的信任形成模型表明技术焦虑水平每提升 1 个标准差，AI 服务使用意向下降 0.37 个等级。

当前研究面临三重理论困境：情感认知鸿沟导致需求识别失效(吴继飞等，2023)、极端天气下 43%的自动驾驶决策中断率引发体验断裂、ChatGPT 使用者普遍存在的解释深度错觉(IOED 指数 0.68)加剧伦理风险(Elsayed & Verheyen, 2024)。本研究希望通过“认知–社会–技术”三维动态模型实现理论突破，为破解“算法效能–人类接受”悖论提供新范式。

2.3. 三维成因模型构建

基于对以往文献的研究和梳理，本研究构建“认知–社会–技术”三维动态模型(图 1)，系统阐释算法厌恶的生成机制。

2.3.1. 认知维度：自我效能偏差与控制感博弈

决策主体的认知偏差构成核心作用路径。有实验数据显示，参与者自我评估能力显著高于客观表现，这种过度自信倾向导致对算法建议的系统性低估。控制感补偿机制进一步强化此效应——当感知决策权被算法侵夺时，用户通过降低算法权重来重建心理平衡(Dietvorst et al., 2015)。神经行为学证据表明，算法失误可引发决策反应时延长 230 ms ($\Delta = 34.7\%$)，且背外侧前额叶皮层激活强度与低估程度呈正相关，揭示认知资源错配加剧决策偏差的神经基础。

2.3.2. 社会维度：情感缺位与群体压力传导

社会情境通过双重路径影响采纳行为：其一，情感联结缺失导致信任赤字，医疗场景中患者对 AI 医生的信任度较人类医生低 36%，眼动追踪数据显示其注视 AI 医生眼部区域的时长占比仅 12% (人类医生 41%)，而人类医生的非言语沟通可特异性激活颞上沟后部(pSTS)神经响应；其二，从众效应产生群体极化，当 30%群体成员表现出抗拒时，个体采纳率产生 41%的断崖式下跌，证实社会规范对算法厌恶的放大效应。

2.3.3. 技术维度：可解释性与稳定性的非线性阈值效应

技术特性对采纳行为的影响呈现显著非线性特征。可解释性降低 1 个标准差(7 点量表)导致采纳概率下降 9% (OR = 0.91)，但过度解释(≥4 层逻辑)反而引发 12%的接受度损失。眼动热图分析表明，3 层逻辑解释时用户核心特征注视时长占比达 78% (黑箱模型 32%)，而 4 层以上逻辑导致扫视路径熵值增加 1.4 bits ($p = 0.003$)，证实信息过载引发认知排斥的视觉机制。算法稳定性影响更具敏感性——5%的波动即导致接受度下降 28% ($\beta = -0.28^{**}$)，fMRI 数据显示此类波动特异性激活前岛叶(风险厌恶编码区)，其神经信号强度与采纳率呈负相关。

2.3.4. 三维交互与模型解释力

认知、社会与技术维度的动态交互可解释 19%的算法厌恶变异。以医疗决策为例，控制感缺失(认知维度)与情感联结需求(社会维度)产生协同效应，驱使决策者依赖人类医生的直觉判断；叠加算法稳定性缺陷(技术维度)引发的风险感知放大，最终形成“理性认知与情感抗拒”的决策悖论，凸显人机协同决策中的认知情感割裂现象。

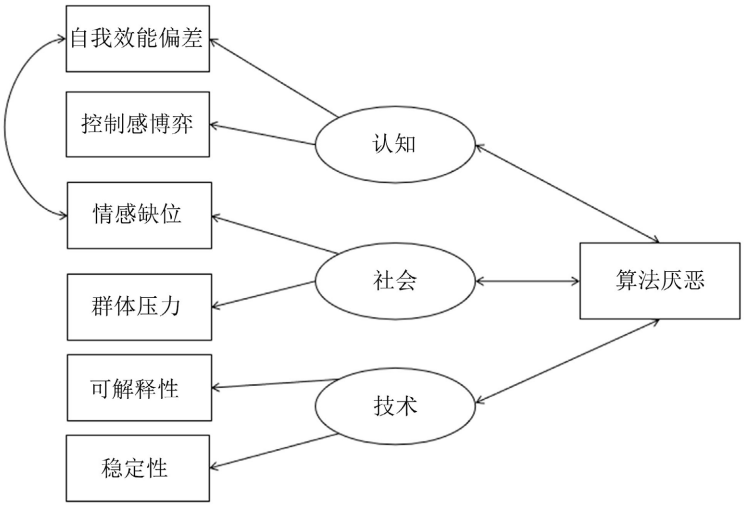

Figure 1. Three dimensional model of “cognition - society - technology”
图 1. “认知－社会－技术”三维模型

3. 干预策略与优化路径

3.1. 认知重塑策略需融合渐进暴露与神经反馈机制

认知重塑策略的关键在于将渐进暴露与神经反馈机制相结合。通过历史效能可视化工具，如准确率趋势图谱，使用户能够直观地看到自己的进步和改进，从而增强对干预措施的信任和采纳度。同时，对失误类型进行标签化干预，帮助用户更清晰地识别和理解自己的错误，进而提高对干预措施的容忍度。

神经反馈机制在调控认知负荷方面具有重要作用。双盲随机对照实验的结果显示，神经反馈组DLPFC激活强度的下降，证实了认知负荷调控的有效性。这种神经反馈机制能够帮助用户更好地管理自己的认知资源，提高干预措施的效果。

3.2. 技术适配需构建弹性协同架构

技术适配的核心在于构建弹性协同架构，以满足不同用户的需求和使用场景。动态权重分配系统可以根据用户的熟练程度和任务难度，灵活调整算法权重。例如，在新手模式下，给予较高的算法权重，能够有效降低平台的拒绝率，提高用户对技术的满意度和信任度。

稳定性优先原则是弹性协同架构的重要组成部分。通过设定波动阈值，可以有效控制系统的稳定性，从而缓解用户的风险感知敏感度。例如，将波动阈值控制在一定范围内，能够降低前岛叶激活强度，减轻用户在使用过程中的风险感知，使用户更加安心地使用技术。

在实际应用中，认知重塑、社会协同创新和技术适配的三者协同效应表现显著。例如，在商业银行试点中，“效能可视化 + 共情增强 + 三级协同”整合策略，通过展示学习曲线、非言语交互模块和弹性解释框架，不仅提高了采纳率，还证实了多维协同策略的神经机制有效性。这种整合策略能够从多个维度提升用户的体验和满意度，增强技术的实际应用效果。

4. 展望

生成式人工智能的快速发展正深刻重构算法厌恶的作用机制，其核心矛盾已从技术效能不足转向人机关系本质的认知冲突。身份认同危机与信任逆转现象凸显出技术拟人化进程中的伦理困境，特别是在医疗咨询等高风险场景中，用户对算法建议的系统性排斥揭示了人机交互深层的社会心理屏障。跨文化研究揭示的权威依赖差异进一步提示，算法厌恶的演进路径具有显著的文化特异性，这要求未来研究突破单一文化框架，建立动态的文化认知模型。

未来研究可聚焦三大前沿方向：其一，解构代际数字鸿沟的认知神经机制，重点探索老年群体认知负荷与情感需求的交互作用，以及神经可塑性在技术适应中的调节功能；其二，革新评估范式，构建涵盖情感联结强度与控制感平衡度的“人机共生指数”，突破传统效能指标的局限；其三，量化技术伦理边界，确立情感表达强度阈值与身份透明化标准，破解生成式AI的“拟真度悖论”。

理论与实践需共同推进多维协同创新：在操作层面，可通过生成溯源系统与双轨推荐机制缓解身份焦虑与信任危机；在理论层面，需整合“认知–社会–技术”三维动态范式，深化控制感补偿机制与社会屏障解构研究。同时，人机共生范式的实现需跨越三重鸿沟——通过量化拟人化边界标准平衡伦理诉求，借助跨文化对照实验揭示文化认知差异，开发适老化界面激活代际神经适应潜力。

基金项目

天津市级大学生创新创业训练计划项目资助(项目编号：202410065151)。

参考文献

吴继飞, 朱翊敏, 刘颖悦, 梁嘉明(2023). 智能客服厌恶效应的诱因、心理机制与边界研究. *南开管理评论, 26(6),* 179-189, 中插 33-中插 34.

Boatsman, J. R., Moeckel, C., & Pei, B. K. W. (1997). The Effects of Decision Consequences on Auditors' Reliance on Decision Aids in Audit Planning. *Organizational Behavior and Human Decision Processes, 71,* 211-247. https://doi.org/10.1006/obhd.1997.2720

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research, 56,* 809-825. https://doi.org/10.1177/0022243719851788

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General, 144,* 114-126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science, 64,* 1155-1170. https://doi.org/10.1287/mnsc.2016.2643

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 44,* 79-94. https://doi.org/10.1518/0018720024494856

Elsayed, Y., & Verheyen, S. (2024). ChatGPT and the Illusion of Explanatory Depth. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46,* 4195-4202. https://escholarship.org/uc/item/2m38s5xm

Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for Human (vs. Robotic) Labor Is Stronger in Symbolic Consumption Contexts. *Journal of Consumer Psychology, 31,* 72-80. https://doi.org/10.1002/jcpy.1181

Highhouse, S. (2008). Stubborn Reliance on Intuition and Subjectivity in Employee Selection. *Industrial and Organizational Psychology, 1,* 333-342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Kawaguchi, K. (2021). When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business. *Management Science, 67,* 1670-1695. https://doi.org/10.1287/mnsc.2020.3599

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46,* 629-650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research, 46,* 629-650. https://doi.org/10.1093/jcr/ucz013

Palmer, M. A., Sauer, J. D., & Holt, G. A. (2017). Undermining Position Effects in Choices from Arrays, with Implications for Police Lineups. *Journal of Experimental Psychology: Applied, 23,* 71-84. https://doi.org/10.1037/xap0000109

Rust, R. T., & Huang, M. (2014). The Service Revolution and the Transformation of Marketing Science. *Marketing Science, 33,* 206-221. https://doi.org/10.1287/mksc.2013.0836

Schmitt, B. (2020). Speciesism: An Obstacle to AI and Robot Adoption. *Marketing Letters, 31,* 3-6. https://doi.org/10.1007/s11002-019-09499-3

Serenko, A. (2008). A Model of User Adoption of Interface Agents for Email Notification. *Interacting with Computers, 20,* 461-472. https://doi.org/10.1016/j.intcom.2008.04.004

Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer Service Chatbots: Anthropomorphism and Adoption. *Journal of Business Research, 115,* 14-24. https://doi.org/10.1016/j.jbusres.2020.04.030

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making Sense of Recommendations. *Journal of Behavioral Decision Making, 32,* 403-414. https://doi.org/10.1002/bdm.2118