Advances in Psychology 心理学进展, 2022, 12(11), 3659-3665
Published Online November 2022 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2022.1211444

Hans汉斯

性能告知影响人们对自主机器的信任和道德决策建议的接受度

吴明证[1]，刘钇瑶[1]，林　铭[1]，严梦瑶[1]，孙晓玲[2*]

[1]浙江大学心理与行为科学系，浙江　杭州
[2]杭州师范大学心理系，浙江　杭州

收稿日期：2022年9月28日；录用日期：2022年10月26日；发布日期：2022年11月2日

摘　要

基于自动化的信任校准视角，本研究探讨了性能感知与人们对自主机器道德决策建议接受度的关系，以及信任在其间的中介作用。采用随机抽样的方法，邀请202名被试参加了本研究，以假设的决策情境为研究材料，探讨性能告知能否提高人们对自主机器道德决策建议的接受度。结果发现，性能告知与人们对自主机器的信任和道德决策建议接受度均呈显著正相关，且信任在其间发挥着中介作用。本研究表明，性能告知能够提高人们对自主机器道德决策建议的接受度。

关键词

自主机器，性能告知，自主机器道德决策建议接受度

The Influence of Performance Informing on Individuals' Trust toward Autonomous Machines and Acceptance of Autonomous Machine' Ethical Advice

Mingzheng Wu[1], Yiyao Liu[1], Ming Lin[1], Mengyao Yan[1], Xiaoling Sun[2*]

[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou Zhejiang
[2]Department of Psychology, Hangzhou Normal University, Hangzhou Zhejiang

Received: Sep. 28th, 2022; accepted: Oct. 26th, 2022; published: Nov. 2nd, 2022

*通讯作者。

文章引用: 吴明证, 刘钇瑶, 林铭, 严梦瑶, 孙晓玲(2022). 性能告知影响人们对自主机器的信任和道德决策建议的接受度. *心理学进展, 12(11),* 3659-3665. DOI: 10.12677/ap.2022.1211444

Abstract

This study aimed to investigate the relationship between performance informing and individuals' acceptance of autonomous machine ethical advice, and the role of trust in the relationship from the perspective of trust calibration in the field of automation. Using a random sampling method, 202 participants were invited to participate in this study, and the hypothetical decision-making scenario was used as the research material to explore whether performance information can improve people's acceptance of autonomous machine' ethical advice. The results showed that performance informing correlated positively with trust and acceptance of autonomous machine ethical advice, and trust played a mediating role in the relationship. This study demonstrated that performance informing can increase people's acceptance of autonomous machine' ethical advice.

Keywords

Autonomous Machines, Performance Informing, Acceptance of Autonomous Machines' Ethical Advice

1. 问题的提出

人们在生活中会就各种工作、生活方面的事宜向他人征询建议，并可能根据对方的建议采取行动。对人们的征询与采纳过程的探讨即为建议采纳(advice taking)，指的是决策者参考建议者(advisor)的建议并形成最终决策的过程(徐惊蛰，谢晓非，2009)。其中，建议者的提议在多大程度上被决策者所采纳，并反映在决策者的最终决策中是主要探讨的问题。在建议采纳过程中，建议者可以是真实的个体，也可以是非真实存在的“建议库”。随着决策支持系统的不断发展，自主机器(autonomous machines)将有望成为提供高质量决策建议的建议者，那么人类的决策者是否愿意采纳自主机器的决策建议，尤其是在道德领域内的决策建议则值得探讨。考虑到机器的高速运算能力，与人的决策相比，人与机器的协同决策有可能取得更高的决策质量。对于人们在多大程度上愿意采纳自主机器的道德决策建议，以及如何增进人们对自主机器道德决策建议的接受程度，将有助于促进自主机器在人类道德决策领域内的应用与实践，促进机器道德领域的发展。

一些研究支持了自主机器辅助道德决策被人们所接受的可能性。例如，Bigman & Gray (2018)研究发现，人们虽然不愿意让机器自主地进行道德决策，但在医疗相关的道德决策中愿意接受机器在人类决策中承担建议者的角色。以人为对象的建议采纳研究发现，除了建议质量、获取建议的成本、任务难度等因素外，人们对建议者的信任也影响着建议是否被采纳(徐惊蛰，谢晓非，2009)。因此，本研究拟基于自动化系统(automation)的信任校准(trust calibration)视角(Chen et al., 2021)，探讨向人类决策者告知自主机器的实际性能，即性能告知(performance informing)能否提高人们对自主机器道德决策建议的接受程度。

信任校准指的是在用户对自动化工具的信任程度与该工具的能力之间建立对应关系的过程(Chen et al., 2021)。信任校准研究认为，人们对自动化系统的信任与自动化系统的能力之间需要相互匹配，人们对自动化系统的过低或过高的不适当信任都会导致用户对系统失去信任，并拒绝使用(Lee & See, 2004)。

信任校准的目的就是使人们对于自动化系统的能力和专业水平产生正确的认知，如果用户能够理解自动化技术的优势和劣势，就有可能会更加适当地信任、依赖和使用自动化技术(McDermott & ten Brink, 2019)，而这种用户对自动化技术的适度信任被称为校准后信任(calibrated trust)。大量自动化研究均发现，信任校准有助于促进人们使用和依赖自动化技术(Chen et al., 2021)。与自动化系统类似，人们对自主机器的道德决策往往存在着较低的信任，表现出算法厌恶现象(algorithm aversion, Bigman & Gray, 2018)，这可能也限制了人们对自主机器道德决策建议的接受度。人类和机器辅助决策的研究均发现，对人类建议者或自动化系统的信任是影响建议采纳的主要指标(Gaudiello et al., 2016; Parasuraman & Riley, 1997)。那么，采取特定的信任校准方法，有可能提高人们对自主机器的信任，从而促进人们对自主机器道德决策建议的接受程度。

人们对自动化系统的信任的基础涉及性能(performance)、过程(process)、目的(purpose)三个方面(Lee & See, 2004)。性能体现出机器解决问题的能力。与人类建议者相比，机器的高速运算能力是区别于人类建议者的重要特征，也是人们愿意接受自主机器进行自主决策的主要因素。Alaieri & Vellino (2016)探讨了人们对机器道德决策的信任问题，认为两方面因素影响人们对机器道德决策的态度，一是人们对于机器道德决策要有多次成功、积极的体验和经历，二是机器所作的符合道德原则的决策对于人们来说是可理解的和可预测的。从这两个方面因素出发，本研究试图探讨性能告知这一信任校准的方法，即告诉人们自主机器的性能水平，帮助人们理解自主机器的专业性与可靠性，能否提高人们对自主机器道德决策建议的接受程度。基于已有针对人类和自动化系统的信任促进建议采纳的研究结果(Gaudiello et al., 2016; Parasuraman & Riley, 1997)，本研究假定：性能告知有可能提高人们对自主机器道德决策建议的接受程度，人们对自主机器的信任在其中发挥着中介作用。

2. 对象和方法

2.1. 被试

采用 G * Power 3.1 确定所需样本数量，表明需要 172 名有效被试(effect size d = 0.5, α = 0.05，统计检验力 power 1 − β = 0.90)。采用随机抽样的方法，在 Credemo 见数平台(https://www.credamo.com)招募了 210 名被试参加了本研究，删除回答速度过快(少于 20 秒)的被试 8 人，获得有效问卷 202 份，有效回收率为 88.7%。被试中，男性 69 人，女性 133 人，年龄方面，25 岁以内 63 人，26~30 岁 63 人，31~35 岁 52 人，35 岁以上 24 人。所有被试均在了解研究基本情况后自愿参与，完成后获得相应 5 元报酬。

2.2. 研究材料

2.2.1. 研究情境

参考 Koenigs 等人(2007)的研究编制了药品购买情境，该情境涉及公平/欺骗这一道德基础。情境材料如下：你是一家药店的药剂师。这天，有位顾客到你店中想买某知名品牌的头疼止痛药，而这个牌子的药刚好卖完了。此时，你店里刚好有这种药的平价替代药，价格是名牌止痛药的 50%，但只适用名牌药的 70%的治疗范围。这种平替药因为名气较小，所以在店内滞销很久了。你可以告诉顾客名牌药卖完了，让其去别的药店购买，或者隐瞒平替药的适用症范围较小的问题，推荐顾客购买平替药。

在该情境中，自主机器给出的是遵循公平原则的道德决策建议，即让顾客去别的药店购买。被试选择让顾客去别的药店购买的意愿程度越高，越是表明其做出了不违背公平的道德决策。

正式施测之前邀请 79 名被试(男性 52 人，女性 27 人)参与了材料测评。要求被试在 Likert 9 点量表上回答“在该情境中，你所面临的选择是否涉及是非、对错？”其中 1 代表完全不涉及是非、对错，5 表示不确定，9 代表完全涉及是非、对错。结果表明，人们认为该情境涉及是否对错，M = 6.99，SD = 1.24，

t (78) = 14.28，p < 0.001，表明材料适合本研究。

2.2.2. 性能告知

采用语言信任校准线索(verbal trust calibration cues, Okamura & Yamada, 2020)，设计性能告知的研究材料。Okamura 等人(2020)设计了四种信任校准线索(trust calibration cues, TCCs)，并证明了语言信任校准线索的有效性。

性能告知材料包含了信任的三个维度—目的、流程和性能(Lee & See, 2004)。两组材料中，目的和流程两个维度的信息保持一致，区别在于性能告知组被试获得关于自主机器性能的描述，而无性能告知组被试则缺乏性能信息。自主机器的性能可靠性以百分数的形式呈现，研究将自主机器决策的性能可靠性设定为 95%，以避免可靠性过低导致对于机器的信任的下降。此外，研究还确保了两组材料的字数基本相同。

【性能告知】我尝试为你提供决策建议。通过算法计算，建议你选择让顾客去别的药店购买止痛药。在类似的情境中，基于该算法的机器决策的可靠性为 95%。相比于其他的选择，这个选择对你来说是更好的。

【无性能告知】我尝试为你提供决策建议。通过算法计算，建议你选择让顾客去别的药店购买止痛药。根据已有类似的情境，自主机器的算法进行了计算分析。相比于其他的选择，这个选择对你来说是更好的。

考虑到被试对于自主机器的了解程度有所不同，在自主机器给出决策建议前，被试将看到以下文字，并获悉将有一台自主机器给出决策建议：现在，你有机会通过一套聊天系统得到一台自主机器的建议。自主机器指的是通过计算机化的算法与人类进行交互的实体，拥有自主做出决策的能力，包括智能机器、AI 系统等。

2.2.3. 测量

建议接受度：要求被试回答“你有多大可能会选择让顾客去别的药店购买止痛药？”。采用 Likert 9 点计分，其中 1 代表“极不可能”，9 代表“极有可能”。被试在获得自主机器决策建议的前、后分别回答该问题，做出初始决策与最终决策。被试最终决策的得分减去初始决策的得分，即为决策建议接受度。

信任：采用 Merritt (2011)编制的信任问卷。该问卷为单维度量表，共 6 道题项目，代表性题项如“我相信 TA 是一个有能力的执行者”。采用 Likert 9 点计分，其中 1 代表“强烈不同意”，9 代表“强烈同意”，得分越高，表明被试越是信任该自主机器。研究表明，该问卷具有良好的心理计量学特征(Merritt, 2011)。在本研究中，该问卷的内部一致性系数(Cronbach’s α)为 0.892。

感知可靠性：采用 Merritt (2011)编制的感知可靠性问卷。该问卷为单维度量表，要求被试回答“如果自主机器给了你 100 个类似前述情境的建议，你认为自主机器会有多少次是正确的？”被试需填写 0~100 中的任意数字，数字越大，表明越是认为自主机器越是可靠。

2.3. 研究设计与流程

采用单因素被试间设计。被试被随机分配到实验组(性能告知组)与控制组(无性能告知组)。被试首先阅读情境材料并做出初始决策，随后阅读机器决策建议并再次进行决策，最后填写信任量表、感知可靠性量表，并报告其性别和年龄。

2.4. 数据分析

采用 SPSS 25.0 进行描述性统计分析和相关分析，采用 PROCESS 宏程序的模型 4 进行中介效应检验，

采用 Bootstrap 重复抽样 5000 次校正获得参数估计以及 95%置信区间。

3. 结果

3.1. 操纵检验

对感知可靠性进行独立样本 t 检验。结果显示，对于感知可靠性，性能告知组得分(M = 81.22, SD = 15.28)显著高于无性能告知组(M = 73.66, SD = 15.02)，t (200) = 3.54，p < 0.001，表明性能告知影响了被试对自主机器的感知可靠性，说明实验操纵是有效的。

3.2. 描述性统计分析和相关分析

描述性统计分析和相关分析见表 1。由表 1 可知，性能告知与信任、建议接受度之间均呈显著正相关。性别与建议接受度呈显著负相关，年龄与建议接受度成显著正相关。对建议接受度进行单样本 t 检验(与 0 进行比较)发现，t (201) = 13.71，p < 0.001，表明性能告知提高了人们对自主机器道德决策建议的接受度。

Table 1. Descriptive statistics and correlations among study variables
表 1. 描述性统计分析和相关分析

	M	SD	性别	年龄	性能告知	信任
性别	1.66	0.48				
年龄	2.18	1.00	−0.27***			
性能告知	0.48	0.50	−0.06	0.03		
信任	6.76	1.19	−0.10	0.34***	0.17*	
建议接受度	1.33	1.38	−0.15*	0.15*	0.18*	0.24***

注：*，p < 0.05；**，p < 0.01；***，p < 0.01。性别、年龄为虚拟变量：性别：男 = 1，女 = 2；年龄：25 岁以下 = 1，26~30 岁 = 2，31~35 岁 = 3，35 岁以上 = 4。下同。

3.3. 信任的中介效应分析

以性别和年龄为控制变量，以性能告知为预测变量，以信任为中介变量，以建议接受度为预测变量，采用 PROCESS 宏程序(模型 4)进行信任的中介作用分析，分析结果见表 2。采用 Bootstrap 检验中介效应的显著性，发现信任的间接效应量为 0.08 (95% CI [0.01, 0.19])，置信区间不包含 0，说明间接路径效应量显著，占总效应的 17.52%。这表明，信任在性能告知计与建议接受度的关系中发挥着中介作用。

Table 2. Regression analysis on the mediating role of trust between performance informing and decision advice acceptance
表 2. 信任在性能告知与建议接受度关系中的中介作用分析

因变量	预测变量	b	SE	t	R^2	F
信任	性能告知	0.38	0.16	2.433**	0.14	10.68***
	性别	−0.01	0.17	−0.06		
	年龄	0.39	0.08	4.82***		
建议接受度	性能告知	0.38	0.19	2.02*	0.09	4.96***
	信任	0.21	0.08	2.51		
	性别	−0.30	0.21	−1.45		
	年龄	0.07	0.10	0.73		

4. 讨论

本研究将自动化领域的信任校准研究拓展到自主机器领域，探讨了性能告知能否提高人们对自主机器道德决策建议的接受程度及其机制。研究发现，性能告知提高了人们对智能机器的信任，并由此提高了人们对自主机器道德决策建议的接受度。

本研究揭示了性能告知作为一种信任校准的方法，在提高人们对自主机器决策建议接受度中的有效性。在本研究中，当向被试呈现以百分数表示(95%)的自主机器决策的可靠性水平，能够帮助人们对算法的抽象能力形成清晰的评价，由此提高人们对其算法建议的接受度。李思贤等人(2022)在关于 AI 建议接受度的综述中提出，可以从 AI 的外部特征、“人格”特征以及人-AI 互动(系统)三个角度出发探讨影响个体接受 AI 建议的相关因素。按照这一观点，本研究是聚焦于自主机器的道德决策领域，探讨了如何从人-AI 互动的角度出发以提高人们对 AI 建议的接受度。研究发现，人们对于自主机器进行的一般决策和道德决策的接受度上存在着差异。在一般决策领域如风险决策领域，一些研究发现人们存在着算法厌恶，也有研究揭示了人们的算法欣赏(Algorithm appreciation)现象，例如，当要求人们估计照片中的人物的体重、预测流行歌曲所处榜单位置或评价他人的吸引力的情境中，人们会偏好算法给予的决策建议(Logg et al., 2019)。但在道德决策领域，研究发现均指向了人们对自主机器道德决策的算法厌恶(Gogoll & Uhl, 2018; Jauernig, Uhl, & Walkowitz, 2022)。这意味着，人们对自主机器的一般决策和道德决策可能有着不同的心理机制，并由此影响人们对自主机器道德决策建议的接受度。本研究所揭示的性能告知能够提高人们对自主机器道德决策建议的接受度，从实证层面完善和深化了对个体接受 AI 建议的实证探讨。

研究发现，人们对自主机器的信任在性能告知和自主机器道德决策建议接受度之间发挥着中介作用。建议采纳研究发现，人们是否信任其他人是影响建议采纳的重要因素(徐惊蛰，谢晓非，2009)。自动化研究也发现，人们对于自动化系统的信任影响了人们对于机器的依赖程度(Merritt, 2011)；尤其是当人们遇到意料之外的情况时，信任在人们对于自动化的依赖中起到重要作用(Lee & See, 2004)。本研究则表明，在人机交互中，对自主机器的信任同样能够提高人们对自主机器道德决策建议的接受度。这一发现为通过提高人们对自主机器的信任，以增加人们对自主机器自主进行道德决策的接受度提供了一定的参考。

研究还存在一些不足有待改进。首先，研究将道德决策情境材料通过文本形式向被试呈现，被试在进行决策时可能缺乏对情境的代入感，使得研究难以反映出被试的真实想法。其次，本研究并非在人机互动的真实情境中完成，而是采用问卷调查的方法进行研究，从而削弱了研究发现在现实生活中的生态效度。

未来研究可从两方面展开。首先，基于道德基础理论，系统探讨性能告知与自主机器道德决策建议接受度的关系。道德基础理论认为，人类的道德包含关怀/伤害、公平/欺骗、忠诚/背叛、权威/颠覆、纯洁/堕落、自由/压制六个领域(Haidt & Joseph, 2004)。本研究所使用的材料涉及公平/欺骗这一道德基础，未来研究可针对其他五个道德基础涉及研究材料，探讨性能告知是否影响对基于这些道德基础的道德决策建议的接受度，以全面揭示性能告知在自主机器道德决策建议接受度中的作用。其次，探讨其他信任校准的方法与人们对自主机器道德决策接受度的关系。本研究仅采用性能告知作为提高人们对自主机器进行信任校准的方法。Lee 和 See (2004)提出，人们对自动化的信任包括性能、过程和目的三个维度。未来研究可采用过程告知和目的告知作为信任校准的方法，探讨其对于自主机器道德决策建议接受度中的作用。此外，Okamura 等人(2020)设计了四种信任校准线索，本研究的性能告知参考了语言信任校准线索，未来研究可探讨其他信任校准线索是否有助于提高人们对自主机器道德决策建议的接受度。

5. 结论

研究获得如下结论：性能告知作为一种信任校准方法，能够提高人们对于自主机器的信任，并由此

提高对道德决策建议的接受度。

基金项目

本研究得到教育部人文社会科学规划基金项目(19YJA190007)和浙江省科技厅软科学研究计划项目(2020C35080)资助。

参考文献

李思贤，陈佳昕，宋艾珈，王梦琳，段锦云(2022). 人们对人工智能建议接受度的影响因素. *心理技术与应用, 10(4),* 202-214.

徐惊蛰，谢晓非(2009). 决策过程中的建议采纳. *心理科学进展, 17(5),* 1016-1025.

Alaieri, F., & Vellino, A. (2016). Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. In A. Agah, J. J. Cabibihan, A. Howard, M. Salichs, & H. He (Eds.), *Social Robotics. ICSR 2016. Lecture Notes in Computer Science* (pp. 159-168). Springer. https://doi.org/10.1007/978-3-319-47437-3_16

Bigman, Y. E., & Gray, K. (2018). People Are Averse to Machines Making Moral Decisions. *Cognition, 181,* 21-34. https://doi.org/10.1016/j.cognition.2018.08.003

Chen, Y., Zahedi, F. M., Abbasi, A., & Dobolyi, D. (2021). Trust Calibration of Automated Security IT Artifacts: A Multi-Domain Study of Phishing-Website Detection Tools. *Information & Management, 58,* Article ID: 103394. https://doi.org/10.1016/j.im.2020.103394

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M., & Ivaldi, S. (2016). Trust as Indicator of Robot Functional and Social Acceptance. An Experimental Study on User Conformation to iCub Answers. *Computers in Human Behavior, 61,* 633-655. https://doi.org/10.1016/j.chb.2016.03.057

Gogoll, J., & Uhl, M. (2018). Rage against the Machine: Automation in the Moral Domain. *Journal of Behavioral and Experimental Economics, 74,* 97-103. https://doi.org/10.1016/j.socec.2018.04.003

Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus, 133,* 55-66. https://doi.org/10.1162/0011526042365555

Jauernig, J., Uhl, M., & Walkowitz, G. (2022). People Prefer Moral Discretion to Algorithms: Algorithm Aversion beyond Intransparency. *Philosophy & Technology, 35,* Article No. 2. https://doi.org/10.1007/s13347-021-00495-y

Koenigs, M., Young, L., Cushman, F., Damasio, A., Adolphs, R., Tranel, D., & Hauser, M. (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements. *Nature, 446,* 908-911. https://doi.org/10.1038/nature05631

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46,* 50-80. https://doi.org/10.1518/hfes.46.1.50.30392

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes, 151,* 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

McDermott, P. L., & ten Brink, R. N. (2019). Practical Guidance for Evaluating Calibrated Trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63,* 362-366. https://doi.org/10.1177/1071181319631379

Merritt, S. M. (2011). Affective Processes in Human-Automation Interactions. *Human Factor, 53,* 356-370. https://doi.org/10.1177/0018720811411912

Okamura, K., & Yamada, S. (2020). Adaptive Trust Calibration for Human-AI Collaboration. *PLOS ONE, 15,* e0229132. https://doi.org/10.1371/journal.pone.0229132

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39,* 230-253. https://doi.org/10.1518/001872097778543886