Advances in Psychology 心理学进展, 2022, 12(8), 2663-2668
Published Online August 2022 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2022.128317

Hans汉斯

人机信任中的信任滥用和信任缺乏论述

王云霄，陈　华

西南交通大学心理研究与咨询中心，四川 成都

收稿日期：2022年7月8日；录用日期：2022年8月2日；发布日期：2022年8月11日

摘　要

信任是人与人之间互动的核心组成部分，也是在我们的社会中建立对人工智能的接受度的一个重要因素，但信任的复杂性使得设计合适的信任水平具有挑战性。这可能会导致人和机器之间出现信任或不信任的情况。本文基于人机信任的相关文献，梳理了人机信任概念和测量方法，并且对信任滥用和信任缺乏的实证研究进行了总结。

关键词

人机信任，测量方法，信任依赖，信任缺乏

A Discussion of Trust Abuse and Lack of Trust in Human-Computer Trust

Yunxiao Wang, Hua Chen

Psychological Research and Counselling Centre, Southwest Jiaotong University, Chengdu Sichuan

Received: Jul. 8th, 2022; accepted: Aug. 2nd, 2022; published: Aug. 11th, 2022

Abstract

Trust is a core component of human interaction and an important factor in building acceptance of AI in our society, but the complexity of trust makes it challenging to design the right level of trust. This can lead to situations of trust or mistrust between humans and machines. This paper compares the concept and measurement of human-machine trust based on the literature on human-

文章引用: 王云霄, 陈华(2022). 人机信任中的信任滥用和信任缺乏论述. *心理学进展, 12(8),* 2663-2668.
DOI: 10.12677/ap.2022.128317

machine trust, and summarizes empirical research on trust abuse and trust deficit.

Keywords

Human-Machine Trust, Measurement Methods, Trust Dependency, Trust Deficit

1. 引言

随着科学技术的飞速发展，能够极大地辅助和代替人类劳动的自动化技术纷纷出现，这些技术的好处为未来的移动性描绘了一幅充满希望的图景，但仍有一些障碍需要解决。主要的一点是人与机器之间的信任问题。例如探索在用户和自动驾驶车辆之间建立适当信任水平。虽然信任是接受自动驾驶汽车的一个关键因素(Adnan et al., 2018)，但对自动驾驶系统缺乏信任(不信任)或过度信任(不信任)可能会产生不利影响(Fraedrich & Lenz, 2014)。另外信任并不是一个常数，而是根据上下文和交互的变化而不断变化的，Hoffman (2017)在声明中指出，信任行为是“在不断变化的工作和不断变化的系统的范围内，积极探索和评估可信度和可靠性的持续过程”。因此本文基于人机信任的相关文献，梳理了人机信任概念和测量方法，并且对信任滥用和信任缺乏的实证研究进行了总结。

2. 人机信任概念

为了更好地操作自动化系统，研究者们在人与人之间的信任的概念基础上，拓展延伸出人机信任的概念(Hengstler, Enkel, & Duelli, 2016)。Lee 和 See (2004)提出的人机信任定义被研究者广为接受。他们认为态度、意愿与行为间存在内在联系：操作者对系统的态度影响其使用系统的意愿和依赖行为，但是依赖系统和有使用系统意愿并不代表信任系统。故 Lee 和 See 从态度角度定义信任，即信任是个体在不确定或易受伤害的情境下认为代理(如自动驾驶)能帮助其实现某个目标的态度。从信任的发展过程来看，Merritt 和 Ilgen (2008)认为操作者对自动化系统的信任是一个处于倾向性信任(Dispositional trust)与历史性信任(History-based trust)间的连续体，前者属于人对自动系统信任的固有倾向，后者属于人通过与自动系统交互后而形成的信任状态。

随着系统的持续使用，操作者对系统的信任逐渐从以倾向性信任成分为主转变为以历史性信任成分为主，并历经初始信任(Initial trust)、实时信任(Ongoing trust)和事后信任(Post-task trust)三种历史信任状态(French et al., 2018)。因此，信任发展分为 4 个发展阶段：倾向性信任、初始信任、实时信任和事后信任(高在峰等，2021)。初始信任在倾向性信任基础上发展而来，指操作者对即将使用的自动系统已具备一定认知、但尚未使用前，对其所持有的信任状态；实时信任指操作者在人机交互过程中对系统的信任状态；而事后信任则指操作者在结束交互后对系统的信任状态，属于事后对系统信任的总体评估。

初始信任从较低水平开始，并随着时间的推移而发展。这意味着对机器人人工智能的信任以一种类似于对人类的信任的方式发展，并在直接互动后增加。例如，Waytz 等人(2014)发现，驾驶部分自动驾驶汽车的参与者对其能力的信任度高于没有这种经验的参与者。

人机信任可以理解为：信任的付出方和被信任的对象(系统)之间进行分析、类比和情感这一系列交互过程的结果。信任的付出方对于被信任的对象(系统)的期望可能匹配，也可能不匹配被信任的对象(系统)

的行为或能力。当发生不匹配时，则可能表现为过度信任或信任缺乏。过度信任是指是信任的付出方的期望高于被信任的对象(系统)的能力。而缺乏信任则恰恰相反，指的是被信任的对象(系统)的能力高于预期水平。

3. 人机信任的测量

3.1. 主观报告法

主观报告法即采用人机信任的相关量表对被试的人机信任水平进行测量。这一定性的度量方法主要用于评估人机信任中的倾向性信任。这是由于倾向性信任本身所具备的稳定性所决定的。

Jian 等人(2000)编制的人机信任量表是目前运用最广泛的人机信任主观报告量表。这个量表一共有12 个题目，采用的是李克特 7 点评分的方式，总分是 84 分。Madsen 人机信任量表从自动化信任的结构入手，包括八个量表以及 38 向题目。认为信任包含 2 个因素：(对自动化系统的)信心及(使用系统的决定或建议的)意愿。Körber (2018)详细阐述了一份信任问卷，在可靠性/能力、理解/可预测性、熟悉度、开发人员的意图、信任倾向和信任自动化维度上有 19 个量表项目。Holthausen 等人(2020)基于 Hoff 和 Bashir 提出的信任模型，提出了自动驾驶情境信任量表(STS-AD)，用于评估情境信任不同方面的简短问卷。另外还包括一些信任程度的单个问题或者程度表示的简单问卷。

主观报告法操作简单，其有效性主要依靠量表的严谨程度。然而在实际操作中主观报告法受到练习效应和无关经验等的影响难以多次施测，另外在测试情景下被试报告态度是否真实有待商榷。

3.2. 动态测量

动态测量是一种定量的测量方法，适合于衡量动态的人机信任。一般通过自动控制系统、模拟器等具体环境对被试的行为以及生理变化进行测量，例如主要任务测量法、行为测量法、生理测量法等。主要任务测量法是指检查和预测给定环境中的自动化系统的激活状态，对于系统的启动和使用时机来测量信任缺乏或者过度信任。如 Lyons 和 Guznov (2019)使用人机交互模拟器来探索操作情境中的人机信任。行为测量是一种客观的测量方式，并且是通过测量影响信任的因素或受信任影响的行为进行一种间接的评估方式。如 Lee 等人(2021)使用自动驾驶系统通过被试的操作行为对于信任的动态变化进行了测量。通过借鉴传统驾驶领域中测量工作负荷等因素的众多生理手段，一部分研究者们探究了在自动驾驶背景下人机信任的生理测量方法。如 Hergeth 等人(2016)采用视觉的驾驶无关任务(NDRT)分析某一自动驾驶任务中眼球运动，发现并总结了一些能够在一定程度上反映人机信任的眼动指标，并发现监控频率与人机信任的主观评估得分呈现一定的反向相关关系。Nuamah 等人(2015)分析了采用 EEG 进行测量人机信任的理论基础和可行性。

动态测量可以得到相对实时的数据，但本质上属于间接测量的方法，不可避免的收到人机信任和具体行为不完全确定因果的影响。在动态过程中，呈现的信息变量相应增多，区分变得更加困难，动态测量的局限性凸显出来。

4. 人工智能中的过度信任

Hoff 和 Bashir (2015)认为，由于对新技术的普遍积极偏见，技术信任的发展方式与人类的发展方式不同。与不熟悉的人之间最初存在的信任度较低相比，新技术可能会对他们的能力和功能产生不切实际的乐观信念。因此，虽然通过频繁的互动，人们对人类的信任通常会随着时间的推移而增加，但基于遇到错误和故障，对技术的信任会随着时间的推移而降低(Matsui & Yamada, 2019)。然而，在人工智能方面，情况可能恰恰相反，当建议来自算法而不是来自人时，人们出现算法欣赏(Logg, Minson, & Moore, 2019)。

由于对人工智能的盲目信任会缠上技术上的积极偏见。这就导致了人类过度信任人工智能。换句话说，当人工智能做出错误的预测时，人类依赖它，即使他们本可以自己做出更好的决定。人工智能的自动化程度越高，人们对系统的初始信任水平可能也越高。因此，许多研究强调了校准信任在改善人工智能辅助决策方面的重要性(Buçinca et al., 2020; Jiang et al., 2018)。当人工智能预测不正确时，在简单可解释人工智能方法的帮助下，人们的表现不如没有人工智能支持的人。结果表明，与简单的可解释人工智能方法相比，认知强迫功能减少了对人工智能的过度依赖(Buçinca, Malaya, & Gajos, 2021)。

人工智能作为一种自动化并不体现意图。从形式上讲，AI 开发人员的能力体现在模型维护特定合同的能力上，而不是采用某种拟人化的意图概念。基于对 AI 开发者的信任的 AI 模型中的信任是一个通过代理实现的个人间信任的实例，而不是人类的信任(Lewis & Weigert, 1985)。由于专业不通导致的对开发者的盲目信任也会造成过高的人机依赖。

5. 人工智能中的信任缺乏

人工智能(AI)正日益自动化不同行业的决策。在许多高风险领域，如医疗保健、教育、刑事司法系统、组织管理和公共援助(Danaher, 2016; Danaher et al., 2017; Lee et al., 2015)。人工智能系统正在自动化或增强人类专家过去所做的决策。为了更好地了解人们对这一变化的反应，许多学者研究了人们如何看待算法决策与人类决策的比较(Castelo et al., 2019; Langer et al., 2019; Lee, 2018)。他们的研究结果表明，人们倾向于认为算法决策不如人类决策，并抵制遵循它们。在几项研究中，人们对数学决策的信任程度低于对人类决策的信任程度，也不太可能采纳这些决策，尤其是当任务被认为需要人类的独特能力、主观性或需要关注个人的独特性时。然而，并非每个人都对人类决策者有同等程度的信任。那些被其他人边缘化的人可能对人类的决定缺乏信心。

有研究表明不信任度较低的参与者对人类决策的信任度高于算法决策，认为人类决策更公平。然而，对人类系统高度不信任的参与者认为，算法和人类决策同样值得信赖和公平(Lee & Rich, 2021)。Lee 比较了人们对算法和人类管理者做出的管理决策的信任、公平和情感反应。Castelo 等人还比较了不同任务主观性的算法决策和人类决策的感知。这两项研究都发现，当任务是主观的时，人们不太可能信任和采用算法决策而不是人类决策。Longoni 等人比较了算法和人类物理学家在医疗诊断和治疗方面的决定，发现人们对算法决定更具抵抗力，可能是因为担心它不能解释个人的独特性(Longoni, Bonezzi, & Morewedge, 2019)。在招聘环境中，Langer 等人(2022)研究了求职者对评估应聘者面试视频的算法的反应，发现人们对算法评估的信任度低于对人的评估。

之前的研究是基于在线实验，Lee 和 Baykal 调查了人们实际算法结果与人类分配结果的实际经验，发现人们认为算法 mic 分配不太公平。综上所述，有越来越多的证据表明，人们对算法决策有着深刻的抵制，并认为人类决策具有优越性。可能是因为人工智能的环境通常高度复杂且部分随机，其行为具有不确定性，这意味着人工智能的决策可能很难预测，而每一个决策背后的逻辑往往很难理解。

6. 总结

人机信任和信任的行为结果之间是相互关联的，特别是如果在不确定性风险的环境之中，人机信任可以说是能够在很大程度上反映了人是否愿意去相信和依赖一个自动化系统。通常而言，实际上人们是会更加倾向于去接受并且去依赖他们所信任的自动化系统，警惕和拒绝他们不信任的自动化系统(Sethumadhavan, 2019)。例如，人们通常不需要对 ATM 机器中去除的钱进行计算核对，因为它没有出现过差错。但是，如果一个人只要自己有过一次在操作 ATM 机器的时候，机器出现了错误的亲身经历以后，他就会在下一次变得谨小慎微，反复进行各种检查核对。这也就是 Bansal 等人(2019)的研究中提出

的人工智能犯错误边界与人机信任的互补过程。有的时候，我们可能被迫依赖自动化系统，但并不会完全信任这个系统，比如当手动执行任务时所带来的快感远远高于自动化系统，这时就可能会选择手动执行任务。

信任是用户和 AI 之间相互作用的一个核心组成部分。人机信任问题的研究取决于脆弱性的概念，目前使用的许多评估方法都不能满足对于概念的绝对定义。在未来进行研究中需要考虑以下几个问题：1) 成功的预期虽然是信任的目标，但不保证其有效性。如果信任不是来源于可信的测量，那么结果可能是不可信的，在这种情况下，预期可能取决于其他情况下不存在的不同变量(例如系统界面的质量)。因此，尽管可模拟性方法作为评估该属性的方法是有用和有价值的，但仅仅依赖它们是不可靠的。2) 只有在合理的情况下，信任才是合乎道德的。无担保的信任并不能保证实现其目标，因为它并非来源于信任。这会导致 AI 被滥用、废弃或误用的问题。虽然未保证的信任可能符合某些当事人的利益，但是人工智能研究应努力通过识别相关评估可信度来诊断和消除未保证的信任。3) 如果有理由的话，不信任并不是绝对不可取的。正如可信和保证信任必须同时满足 AI 在实践中的实用性要求一样。完全可信的 AI 可能难以实现，在这种情况下，保证不信任是允许不完美 AI 发挥作用的机制。

基金项目

四川省心理学会年度科研规划重点项目成果(项目编号：SCSXLXH2021001)。

参考文献

高在峰, 李文敏, 梁佳文, 潘晗希, 许为, 沈模卫(2021). 自动驾驶车中的人机信任. *心理科学进展, 29(12),* 2172-2183.

Adnan, N., Md Nordin, S., bin Bahruddin, M.A., & Ali, M. (2018). How Trust Can Drive forward the User Acceptance to the Technology? In-Vehicle Technology for Autonomous Vehicle. *Transportation Research Part A: Policy and Practice, 118,* 819-836. https://doi.org/10.1016/j.tra.2018.10.019

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7,* 2-11.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI'20)* (pp. 454-464). ACM. https://doi.org/10.1145/3377325.3377498

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction, 5,* 1-21. https://doi.org/10.1145/3449287

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

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

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J. et al. (2017). Algorithmic Governance: Developing a Research Agenda through the Power of Collective Intelligence. *Big Data & Society, 4*. https://doi.org/10.1177/2053951717726554

Fraedrich, E., & Lenz, B. (2014). Automated Driving: Individual and Societal Aspects. *Transportation Research Record, 2416,* 64-72. https://doi.org/10.3141/2416-08

French, B., Duenser, A., & Heathcote, A. (2018). *Trust in Automation—A Literature Review*. CSIRO.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied Artificial Intelligence and Trust—The Case of Autonomous Vehicles and Medical Assistance Devices. *Technological Forecasting and Social Change, 105,* 105-120. https://doi.org/10.1016/j.techfore.2015.12.014

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust during Highly Automated Driving. *Human Factors, 58,* 509-519. https://doi.org/10.1177/0018720815625744

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.

Human Factors, 57, 407-434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R. (2017). A Taxonomy of Emergent Trusting in the Human-Machine Relationship. In P. J. Smith, & R. R. Hoffman (Eds.), *Cognitive Systems Engineering: The Future for a Changing World* (pp. 137-164). CRC Press. https://doi.org/10.1201/9781315572529-8

Holthausen, B. E., Wintersberger, P., Walker, B. N., & Riener, A. (2020). Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 40-47). Association for Computing Machinery. https://doi.org/10.1145/3409120.3410637

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics, 4,* 53-71. https://doi.org/10.1207/S15327566IJCE0401_04

Jiang, H., Kim, B., Guan, M. Y., & Gupta, M. (2018). To Trust or Not to Trust a Classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 5546-5557). Curran Associates Inc.

Körber, M. (2018). Theoretical Considerations and Development of a Questionnaire to Measure Trust In automation. In *Congress of the International Ergonomics Association* (pp. 13-30). Springer, Cham. https://doi.org/10.31234/osf.io/nfc45

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly Automated Job Interviews: Acceptance under the Influence of Stakes. *International Journal of Selection and Assessment, 27,* 217-234. https://doi.org/10.1111/ijsa.12246

Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in Artificial Intelligence: Comparing Trust Processes between Human and Automated Trustees in Light of Unfair Bias. *Journal of Business and Psychology,* 1-16. https://doi.org/10.1007/s10869-022-09829-9

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

Lee, J., Abe, G., Sato, K., & Itoh, M. (2021). Developing Human-Machine Trust: Impacts of Prior Instruction and Automation Failure on Driver Trust in Partially Automated Vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour, 81,* 384-395. https://doi.org/10.1016/j.trf.2021.06.013

Lee, M. K. (2018). Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management. *Big Data & Society, 5,* 1-16. https://doi.org/10.1177/2053951718756684

Lee, M. K., & Rich, K. (2021). Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445570

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1603-1612). https://doi.org/10.1145/2702123.2702548

Lewis, J. D., & Weigert, A. (1985). Trust as a Social Reality. *Social Forces, 63,* 967-985. https://doi.org/10.2307/2578601

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

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

Lyons, J. B., & Guznov, S. Y. (2019). Individual Differences in Human-Machine Trust: A Multi-Study Look at the Perfect Automation Schema. *Theoretical Issues in Ergonomics Science, 20,* 440-458. https://doi.org/10.1080/1463922X.2018.1491071

Matsui, T., & Yamada, S. (2019). Designing Trustworthy Product Recommendation Virtual Agents Operating Positive Emotion and Having Copious Amount of Knowledge. *Frontiers in Psychology, 10,* Article 675. https://doi.org/10.3389/fpsyg.2019.00675

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50,* 194-210. https://doi.org/10.1518/001872008X288574

Nuamah, J., Oh, S., & Seong, Y. (2015). Measuring Trust in Automation: A New Approach. In *Proceedings of the Modern Artificial Intelligence and Cognitive Science Conference*.

Sethumadhavan, A. (2019). Trust in Artificial Intelligence. *Ergonomics in Design: The Quarterly of Human Factors Applications, 27,* 34-34. https://doi.org/10.1177/1064804618818592

Waytz, A., Heafner, J., & Epley, N. (2014). The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *Journal of Experimental Social Psychology, 52,* 113-117. https://doi.org/10.1016/j.jesp.2014.01.005