machine learning & knowledge extraction

MDPI

Comment

Comment on Novozhilova et al. More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts. *Mach. Learn. Knowl. Extr.* 2024, *6*, 342–366

Robertas Damaševičius

Centre of Real Time Computer Systems, Kaunas University of Technology, 51424 Kaunas, Lithuania; robertas.damasevicius@ktu.lt

The referenced article [1] aims to extend the understanding of public perceptions of artificial intelligence (AI) systems beyond individual user perspectives to encompass broader societal trust. It examines how demographic traits and technological familiarity influence public trust across different domains such as healthcare, education, and creative arts, through a large-scale survey (N = 1506). This study is particularly relevant to current state-of-the-art research as it addresses the complex dimensions of trust in AI, distinguishing between perceived capabilities and benevolence of AI systems in varied societal roles. The relevance of this research lies in its comprehensive approach to evaluating public trust in AI, which is crucial for developing and implementing AI technologies responsibly and ethically. By exploring both the capabilities and benevolence of AI systems in critical sectors, the study contributes valuable insights to ongoing discussions about AI governance and the need for human-centered AI design. These insights are essential for ensuring that AI development aligns with societal values and needs, thus supporting more informed policy-making and AI system design that foster public trust and acceptance. The article's exploration of demographic and technological familiarity as influencers of trust further contributes to understanding the socio-technical dynamics at play, providing a comprehensive view that supports the development of more targeted, human-centric AI governance and policy frameworks.

The methodology employed by Novozhilova et al. presents several strengths and potential weaknesses. A major strength is the robust sample size (N = 1506), which enhances the generalizability of the findings across the U.S. population. The use of a detailed survey instrument to assess diverse dimensions of trust across multiple domains also provides a comprehensive insight into the complex interplay of factors influencing public perceptions of AI. However, the methodology exhibits potential weaknesses. Primarily, it relies on self-reported data, which might skew the results due to participants' subjective understanding and experiences with AI, potentially leading to biased assessments of trust. The cross-sectional nature of the survey limits the ability to establish causality or observe changes over time, which are crucial for understanding the dynamic nature of trust as AI technologies evolve and become more integrated into everyday life. These methodological concerns suggest a need for a more robust approach, potentially incorporating longitudinal studies and qualitative methods, to provide a deeper and more accurate understanding of the complex interplay between demographic factors, technological familiarity, and trust in AI.

The original article's exploration of public trust in AI across various domains aligns with and diverges from the existing literature in nuanced ways. Like Choung's foundational concept of trust in AI technologies [2], the study emphasizes the importance of trust in AI adoption, focusing on both capabilities and benevolence, reflecting the multi-faceted nature of trust delineated in prior research. However, it uniquely extends these notions by demonstrating that public trust varies significantly across different contexts—healthcare, education, and creative arts—highlighting a domain-specific trust perspective not thoroughly examined in earlier works such as that by Herse et al. [3]. Previous studies, such as those by Nakao et al. [4], suggest a general aversion to algorithmic decision-making

Citation: Damaševičius, R. Comment on Novozhilova et al. More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts. *Mach. Learn. Knowl. Extr.* 2024, *6*, 342–366. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1667–1669. https://doi.org/10.3390/make6030081

Academic Editor: Andreas Holzinger

Received: 6 May 2024
Revised: 13 June 2024
Accepted: 14 June 2024
Published: 22 July 2024

despite its performance advantages; the original article complicates this view by showing that trust is not uniformly distributed across AI's capabilities and intentions, indicating a more sophisticated public discernment of AI's roles. This contrasts with simpler models of trust in technology that do not account for such granularity, thus enriching the theoretical landscape with insights into the complexities of public perceptions of AI within varied societal contexts [5].

To advance the theoretical and empirical grounding of the study [1], further research could be performed. Firstly, adopting a longitudinal research design would allow for the examination of changes in public trust over time, particularly in response to rapid advancements in AI technology and policy shifts. This approach would capture the dynamics of trust and its evolution as the public becomes more familiar with and potentially dependent on AI systems. Secondly, integrating qualitative methodologies, such as in-depth interviews or focus groups, would enrich the quantitative data, offering deeper insights into the reasons behind varying levels of trust across different demographics and contexts. Such mixed methods would provide a more nuanced understanding of the complex interplay between AI capabilities and perceived benevolence. Expanding the study to include cross-cultural comparisons could reveal how cultural differences influence trust in AI, thereby enhancing the generalizability and applicability of the findings across global contexts. This would be instrumental in designing AI systems and policies that are culturally sensitive and globally effective.

The findings from the study by Novozhilova's et al. [1] have implications for the development, deployment, and governance of AI technologies. By illustrating the discrepancy between the perceived capabilities and benevolence of AI systems, the study underscores a critical challenge for AI adoption: the public's nuanced apprehension towards AI's roles in society. This revelation is crucial for policymakers and developers as it highlights the need for AI systems that are not only technically proficient but also transparent, ethical, and aligned with human values to enhance their benevolence perception. Such insights are indispensable for informing strategies that aim to cultivate public trust, a key enabler of broader AI acceptance and integration into daily life. The differential trust across domains suggests that sector-specific approaches (such as that presented in [6]) might be necessary to address unique concerns and expectations, thereby guiding more targeted and effective regulatory frameworks. The emphasis on demographic and technological familiarity factors also suggests that educational and outreach programs tailored to various demographic groups could democratize AI literacy and empowerment, ultimately fostering a more informed and engaged public that can participate actively in the discourse surrounding AI technologies.

Building on the findings of Novozhilova et al.'s study, several areas for further research emerge. One critical avenue is to investigate the underlying psychological mechanisms that drive differential trust in AI's capabilities versus its benevolence. Experimental studies could manipulate variables such as AI transparency, ethical alignment, and user control to directly assess their impact on trust dynamics [7,8]. Further research could explore the intersection of AI trust with behavioral outcomes, such as willingness to use AI in critical decision-making scenarios, to link perceptual trust metrics to actual user behavior [9,10]. Another promising area involves the extension of trust research to include emerging AI applications in unexplored domains such as autonomous public transport or legal adjudication, where public trust could significantly influence the technology's adoption and regulatory oversight [11]. Conducting comparative studies across different cultures and regulatory environments would provide insights into how contextual factors influence public trust in AI, offering a more global perspective on the adoption challenges and opportunities for AI technologies [12]. These research directions could provide deeper insights into shaping policies and AI system designs that are more attuned to public expectations and concerns, thereby enhancing the societal integration of AI technologies.

Funding: This research received no external funding.

Acknowledgments: The author used Grammarly, an AI-assisted language editing tool, to enhance the grammatical correctness and fluency of language.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Novozhilova, E.; Mays, K.; Paik, S.; Katz, J.E. More Capable, Less Benevolent: Trust Perceptions of AI Systems across Societal Contexts. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 342–366. [CrossRef]
2. Choung, H.; David, P.; Ross, A. Trust in AI and Its Role in the Acceptance of AI Technologies. *Int. J. Hum.-Comput. Interact.* **2022**, *39*, 1727–1739. [CrossRef]
3. Herse, S.; Vitale, J.; Williams, M.-A. Using agent features to influence user trust, decision making and task outcome during human-agent collaboration. *Int. J. Hum.-Comput. Interact.* **2023**, *39*, 1740–1761. [CrossRef]
4. Nakao, Y.; Strappelli, L.; Stumpf, S.; Naseer, A.; Regoli, D.; Del Gamba, G. Towards responsible AI: A design space exploration of human-centered artificial intelligence user interfaces to investigate fairness. *Int. J. Hum.-Comput. Interact.* **2022**, *39*, 1762–1788. [CrossRef]
5. Danks, D. The Value of Trustworthy AI. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019. [CrossRef]
6. Plotnikova, V.; Dumas, M.; Nolte, A.; Milani, F. Designing a data mining process for the financial services domain. *J. Bus. Anal.* **2023**, *6*, 140–166. [CrossRef]
7. Schmidt, P.; Biessmann, F.; Teubner, T. Transparency and trust in artificial intelligence systems. *J. Decis. Syst.* **2020**, *29*, 260–278. [CrossRef]
8. Karran, A.; Demazure, T.; Hudon, A.; Sénécal, S.; Léger, P.M. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Front. Neurosci.* **2022**, *16*, 1–25. [CrossRef] [PubMed]
9. Ajenaghughrure, I.B.; Sousa, S.; Lamas, D. Psychophysiological Modeling of Trust In Technology. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–25. [CrossRef]
10. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI with a Case Study on IIoT Security. *IEEE Internet Things J.* **2022**, *10*, 2967–2978. [CrossRef]
11. Yu, L.; Li, Y. Artificial Intelligence Decision-Making Transparency and Employees' Trust: The Parallel Multiple Mediating Effect of Effectiveness and Discomfort. *Behav. Sci.* **2022**, *12*, 127. [CrossRef] [PubMed]
12. Liu, J.; Marriott, K.; Dwyer, T.; Tack, G. Increasing User Trust in Optimisation through Feedback and Interaction. *ACM Trans. Comput.-Hum. Interact.* **2022**, *29*, 1–34. [CrossRef]