Please quote as: Schmitt, A. (2022). Examining Trust in Conversational Systems: Conceptual and Empirical Findings on User Trust, Related Behavior, and System Trustworthiness. AAAI/ACM Conference on AI, Ethics, and Society (AIES). Oxford, United Kingdom.

Examining Trust in Conversational Systems

Conceptual and Empirical Findings on User Trust, Related Behavior, and System Trustworthiness

Anuschka Schmitt
University of St.Gallen
St.Gallen, Switzerland
anuschka.schmitt@unisg.ch

ABSTRACT

Machine learning (ML)-based conversational systems represent a value enabler for human-machine interaction. Simultaneously, the opacity, complexity, and humanness accompanied by such systems introduce their own issues, including trust misalignment. While trust is viewed as a prerequisite for effective system use, few studies have considered calibrating for appropriate trust, and empirically testing the relationship between trust and related behavior. Moreover, the desired implications of transparency-enhancing design cues are ambiguous. My research aims to explore the impact of system performance on trust, the dichotomy between trust and behavior, and how transparency might help attenuate the effects caused by low system performance in the specific context of decision-making tasks assisted by ML-based conversational systems.

ACM Reference format:

Anuschka Schmitt. 2022. Examining Trust in Conversational Systems: Conceptual and Empirical Findings on User Trust, Related Behavior, and System Trustworthiness. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). August 1-3, 2022, Oxford, UK.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3514094.3539525

1 Research Background and Aim

Conversational systems have become a ubiquitous part of our everyday life, being characterized by increasing amount of data and model complexity [1]. While such technological advancements enable a more natural interaction, issues of opacity and complexity are exacerbated. Research narratives around anthropomorphism have framed our understanding of conversational systems and demonstrate the (unintended) implications of systems' increasing agency [2]. Overall, the opacity, complexity, and agency associated with conversational systems give rise to the importance of trust. While "[t]rust is an important mechanism for coping with the cognitive complexity that accompanies increasingly sophisticated technology" [3], simply enhancing user trust is not reasonable per se as users might be induced to over-trust the system beyond its capabilities and performance [4]. To develop appropriate trust, researchers such as [5] have proposed efforts of system trustworthiness and thereby making system confidence or insights into underlying ML models transparent to the user.

My goal is to map the discourses around how ML-based conversational systems produce, maintain, and modify (the need for) trust, and to highlight the relations between trust and downstream behavior. I aim to contribute to the field of trust and behavioral research by exploring the following research question: *How does the accuracy and transparency cues of ML-based conversational systems affect user perceptions and behavior?* More specifically, I plan to investigate whether and how variations in advice accuracy affect the user, as well as to what extent measures suggested to mitigate overtrust hold up to their claimed goal of preventing undesirable relationships with ML-based systems. I thereby consider 1) varying confidence and accuracy levels, and 2) transparency statements as instantiated methods of system trustworthiness in different learning tasks and scenarios deploying ML-based conversational systems.

ACKNOWLEDGEMENTS

We thank the Swiss National Science Foundation for funding parts of this research (192718).

REFERENCES

[1] Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G. and P. De Hert, P. 2022. Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. In IEEE Computational Intelligence Magazine, 17 (1), 72-85.

[2] Puranam, P., & Vanneste, B. 2021. Artificial Intelligence, Trust, and Perceptions of Agency. 2021/42/STR, Available at https://ssrn.com/abstr.

[3] Lee, J.D. and See, K.A. 2002. Trust in Computer Technology and the Implications for Design and Evaluation. Available at: www.aaai.org.

[4] Zhang, Y., Vera Liao, Q. and Bellamy, R.K.E. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery (FAT* 2020), Inc, 295–305.

[5] Hoffman, R.R., Klein, G. and Mueller, S.T. 2018. Explaining explanation for "explainable AI. In Proceedings of the Human Factors and Ergonomics Society, 1, Human Factors and Ergonomics Society Inc., 197–201.

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3539525