Piecing Together the Puzzle: Understanding Trust in Human-AI Teams

Anna-Sophie Ulfert-Blank[a,1], Eleni Georganta[b], Myrthe Tielman[3], and Tal Oron Gilad
[a] *Eindhoven University of Technology, the Netherlands*
[b] *University of Amsterdam, the Netherlands*
[c] *Delft University of Technology, the Netherlands*
[d] *Ben Gurion University of the Negev, Israel*

ORCiD ID: Anna-Sophie Ulfert-Blank https://orcid.org/0000-0001-6293-4173
Eleni Georganta https://orcid.org/0000-0002-9070-5930
Myrthe Tielman https://orcid.org/0000-0002-7826-5821
Tal Oron Gilad https://orcid.org/0000-0002-9523-0161

Abstract. With the increasing adoption of AI as a crucial component of business strategy, the challenge of establishing trust between human and AI teammates remains a key issue. The project "We are in this together" highlights current theories on team trust in human-AI teams and proposes a research model that integrates insights from Industrial and Organizational Psychology, Human Factors Engineering, Human-Computer Interaction, and Computer Science. The proposed model suggests that trust in human-AI teams involves multiple actors and is critical for team success. We present three main propositions for understanding trust in human-AI team collaboration, focused on the importance of trustworthiness and trustworthiness reactions in interpersonal relationships between human and AI teammates. We further suggest that individual, technological, and environmental factors can impact trust relationships in human-AI teams. The goal of the project is to contribute to the development of effective human-AI teams by presenting and experimentally evaluating a research model of team trust in human-AI teams.

Keywords. Trust, Human-AI teams

1. Introduction

Artificial intelligence (AI) is rapidly transforming the workplace, with an increasing number of organizations adopting AI as a crucial component of their business strategy [1]. AI is no longer a simple tool but has become a teammate that works alongside humans to improve productivity, efficiency, and decision-making [2]. For example, SAP has utilized the AI-powered assistant "Olivia" to help with recruiting tasks, such as scheduling interviews and answering employee questions. With the introduction of new generative AI tools, in recent months, many companies have implemented new AI assistants using ChatGPT to support customers while shopping (e.g., Shopify) or to assist customer support staff (e.g. salesforce) [3]. In the light of these recent developments, researchers have argued that AI are transforming from tools to team members [2]. At the same time, this development has raised concerns about how such teams can collaborate

[1] Corresponding Author: Anna-Sophie Ulfert-Blank, a.s.ulfert.blank@tue.nl.

effectively [5] and how such collaborations should be implemented in the workplace while guaranteeing human safety and well-being [6].

According to both practice and research [7], human-AI teams are expected to become prevalent in the workforce. However, appropriate trust between team members remains a key challenge in the development of such teams [4]. This is because humans tend to have difficulties in developing appropriate trust in intelligent technologies [8], [9] and understanding their behaviors [10]. Moreover, it is unclear how AI teammates can express their status and intentions to be perceived as trustworthy by human teammates [7]. Recent research suggests that trust within human-AI teams will be essential for teams to communicate, integrate information, coordinate, and perform effectively [12], yet clear guidelines for how human-AI teams should be designed to foster appropriate trust are still missing. Research across disciplines already offers insights into how human teams develop trust, how humans interact with technology, and how AI systems should be designed.

2. Trust perspectives across disciplines

Across disciplines, the study of trust as a central influencing factor on human-technology interaction has a longstanding tradition which has led to a variety of perspectives and definitions [13], [14]. In psychology and human-technology interaction literature, trust in humans and trust between humans and AI is typically defined as the willingness to rely on and be vulnerable to another party [13]–[16]. Trust relationships in human-AI teams involve multiple actors, including human and AI teammates, and are critical to team success [17]. Further, trust is influenced by characteristics and states of the human and the AI team member as well as their shared environment [13]. Team trust describes the shared perception among team members that enables free sharing of information and views and reflects one of the most crucial properties for team success [18]. To achieve effective collaboration, both human and AI teammates must perceive each other as trustworthy and perceive that they are being trusted [17]. Computer science literature so far mainly focuses on dyadic trust relationships between humans and agents or trust between agents [19]. It is still unclear how team trust in human-AI teams can be understood and what mechanisms underlie team trust emergence [17]. Nevertheless, literature across disciplines agrees that to collaborate effectively, human-AI teams require appropriate levels of trust that are bidirectional (i.e., expressed by the human and the AI agent). Consequently, engineering literature has proposed design approaches with the goal of building appropriate bidirectional trust between humans and AI (i.e., trust engineering; [20]), for instance by addressing aspects such as explainability, security, or training.

Although these different literature streams have addressed trust as a central construct for effective collaboration, their definitions often differ depending on their unique disciplinary perspective [9]. For instance, a large body of literature focuses on the role of trust for team processes and how individual perceptions by team members impact collaboration, thus highlighting the perspective of the

human as a trustor in their collaboration with AI team members [21]. In contrast, trust engineering literature predominantly addresses technical challenges of human-AI teaming, such as data protection, transparency, or interface design [20]. Currently, there is still a lack of integration of technical system design perspectives and team processes in human-AI teams.

In addition a need for more integrated definition, new insights are needed with regards to the emergence of team trust. Human-AI teams can largely differ in their composition of human and agent team members (i.e., the number of human and agent team members). A team's composition can strongly impact how trust develops and how trustworthy team members are perceived [22]. Yet, current research predominantly highlights reactions to team member characteristics, rather than the dynamic development of trust in human-AI teams, trust reciprocity, or differing trust levels between team members [16]. Consequently, team trust needs to be considered from a dynamic and multi-level perspective, where team members differ in their characteristics (e.g.. AI or human; trustworthiness), they behaviors (e.g., how they display trust), and their relationships (e.g., their trust relationship with individual team members).

As human-AI teams are finding their way into work environments, we urgently require further integration of perspectives to answer pressing questions such as: How should trust in human-AI teams be defined considering different disciplinary perspectives? How much trust is needed in human-AI teams? How can we and should we reach theoretical unification of the vast and constantly growing trust literature? And, how do we guarantee work environments that foster not only performance but also human safety and well-being?

3. Addressing trust from multidisciplinary perspectives

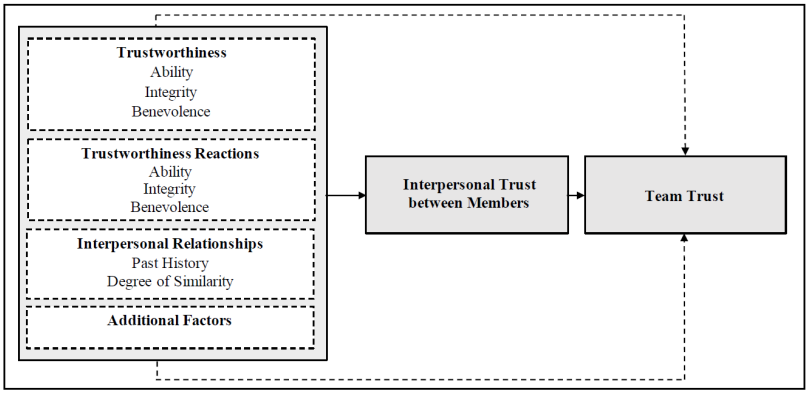

Figure 1. Research model of the project "We are in this together"

To address this research gap, our project "We are in this together"[2] aims to provide theoretical and empirical arguments for gaining a better understanding of trust in human-AI team collaboration (see Figure 1). Specifically, we combine knowledge from Industrial and Organizational Psychology, Human Factors Engineering, Human-Computer Interaction, and Computer Science and present three main propositions: (1) trust in human-AI teams considers human and AI teammates as trustors and trustees; (2) trust in human-AI teams depends on human and AI trustworthiness, their trustworthiness reactions, as well as interpersonal relationship between teammates; and (3) trust in human-AI teams is multilevel, including individual-, dyadic-, and team-level trust. We further propose that (4) additional factors, such as individual, technological, and environmental considerations, form and impact trust relationships in human-AI teams [16].

Our overall goal is to contribute to the development of effective human-AI teams by presenting a research model of team trust in human-AI teams and investigate its propositions in experimental studies. During the workshop, first results from the experimental studies will be presented.

References

[1] M. H. Jarrahi, 'Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making', *Business Horizons*, vol. 61, no. 4, pp. 577–586, Jul. 2018, doi: 10.1016/j.bushor.2018.03.007.

[2] L. Larson and L. DeChurch, 'Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams', *The Leadership Quarterly*, vol. 31, no. 1, pp. 1–18, 2020.

[3] A. Tellez, 'These Major Companies—From Snap To Salesforce— Are All Using ChatGPT', *Forbes*. https://www.forbes.com/sites/anthonytellez/2023/03/03/these-major-companies-from-snap-to-instacart--are-all-using-chatgpt/ (accessed Apr. 03, 2023).

[4] I. Seeber, L. Waizenegger, S. Seidel, S. Morana, I. Benbasat, and P. B. Lowry, 'Collaborating with technology-based autonomous agents', *Internet Research*, 2020.

[5] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, 'Human–Robot Collaboration in Manufacturing Applications: A Review', *Robotics*, vol. 8, no. 4, p. 100, Dec. 2019, doi: 10.3390/robotics8040100.

[6] J. Narayan, K. Hu, M. Coulter, and S. Mukherjee, 'Elon Musk and others urge AI pause, citing "risks to society"', *Reuters*, Mar. 29, 2023. Accessed: Apr. 03, 2023. [Online]. Available: https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/

[7] R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, '"An Ideal Human": Expectations of AI Teammates in Human-AI Teaming', *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW3, p. 246:1-246:25, Jan. 2021, doi: 10.1145/3432945.

[8] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, 'A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems', *Hum Factors*, vol. 58, no. 3, pp. 377–400, May 2016, doi: 10.1177/0018720816634228.

[9] A. Kaplan, T. T. Kessler, J. C. Brill, and P. A. Hancock, 'Trust in Artificial Intelligence: Meta-Analytic Findings', *Hum Factors*, p. 00187208211013988, May 2021, doi: 10.1177/00187208211013988.

[10] R. Singh, T. Miller, J. Newn, E. Velloso, F. Vetere, and L. Sonenberg, 'Combining gaze and AI planning for online human intention recognition', *Artificial Intelligence*, vol. 284, p. 103275, Jul. 2020, doi: 10.1016/j.artint.2020.103275.

[2] Funded by the Society of Industrial and Organizational Psychology Visionary Grant

[11] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, 'Ten challenges for making automation a "team player" in joint human-agent activity', *IEEE Intelligent Systems*, vol. 19, no. 6, pp. 91–95, Nov. 2004, doi: 10.1109/MIS.2004.74.

[12] A.-S. Ulfert and E. Georganta, 'A Model of Team Trust in Human-Agent Teams', in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, in ICMI '20 Companion. New York, NY, USA: Association for Computing Machinery, 2020, pp. 171–176. doi: 10.1145/3395035.3425959.

[13] K. A. Hoff and M. Bashir, 'Trust in automation: Integrating empirical evidence on factors that influence trust', *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

[14] J. D. Lee and K. A. See, 'Trust in automation: Designing for appropriate reliance', *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.

[15] R. C. Mayer, J. H. Davis, and F. D. Schoorman, 'An integrative model of organizational trust', *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[16] E. J. de Visser *et al.*, 'Towards a theory of longitudinal trust calibration in human-robot teams.', *International Journal of Social Robotics*, vol. 12, pp. 459–478, 2020, doi: 10.1007/s12369-019-00596-x.

[17] A.-S. Ulfert, E. Georganta, C. Centeio Jorge, S. Mehrotra, and M. L. Tielman, 'Shaping a multidisciplinary understanding of Team Trust in Human-AI Teams: A Theoretical Framework', *European Journal of Work and Organizational Psychology*, in press.

[18] B. A. De Jong, K. T. Dirks, and N. Gillespie, 'Trust and team performance: A meta-analysis of main effects, moderators, and covariates.', *Journal of Applied Psychology*, vol. 101, no. 8, pp. 1134–1150, 2016.

[19] C. Centeio Jorge, S. Mehrotra, M. Tielman, and C. M. Jonker, 'Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams', in *22nd International Trust Workshop co-located with AAMAS 2021*, London, UK, 2021.

[20] N. Ezer, S. Bruni, Y. Cai, S. J. Hepenstal, C. A. Miller, and D. D. Schmorrow, 'Trust Engineering for Human-AI Teams', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 322–326, Nov. 2019, doi: 10.1177/1071181319631264.

[21] Committee on Human-System Integration Research Topics for the 711th Human Performance Wing of the Air Force Research Laboratory, Board on Human-Systems Integration, Division of Behavioral and Social Sciences and Education, and National Academies of Sciences, Engineering, and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, D.C.: National Academies Press, 2022. doi: 10.17226/26355.

[22] B. G. Schelble *et al.*, 'Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming', *Hum Factors*, p. 001872082211169, Aug. 2022, doi: 10.1177/00187208221116952.