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

Anna-Sophie Ulfert-Blank[1], Eleni Georganta[2], Myrthe Tielman[3], and Tal Oron-Gilad[4]

[1] *Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands*
[2] *University of Amsterdam, Nieuwe Achtergracht 129, 1001 NK Amsterdam, the Netherlands*
[3] *Delft University of Technology, Mekelweg 5, 2628 CD Delft, the Netherlands*
[4] *Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 8410500, Israel*

Abstract

With the increasing adoption of Artificial intelligence (AI) as a crucial component of business strategy, establishing trust between humans and AI teammates remains a key issue. The project "We are in this together" highlights current theories on trust in Human-AI teams (HAIT) 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 in HAIT, trust involves multiple actors and is critical for team success. We present three main propositions for understanding trust in HAIT collaboration, focused on trustworthiness and trustworthiness reactions in interpersonal relationships between humans and AI teammates. We further suggest that individual, technological, and environmental factors impact trust relationships in HAIT. The project aims to contribute in developing effective HAIT by proposing a research model of trust in HAIT.

Keywords
Trust, Human-AI teams, Multidisciplinary research, Theoretical model

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 AI assistants using ChatGPT to support customers while shopping (e.g., Shopify) or to assist customer support staff (e.g. Salesforce) [3]. In light of these recent developments, researchers argue that AI is transforming from a tool to a teammate [2]. At the same time, this development has raised concerns about how such teams can collaborate effectively, [5] and how such collaborations should be implemented in the workplace while guaranteeing the safety and well-being of employees [6].

According to practice and research [7], HAIT are expected to become prevalent in the workforce. However, appropriate trust between team members remains a key challenge in

HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26—27, 2023, Munich, Germany
a.s.ulfert.blank@tue.nl (A.S. Ulfert-Blank)
0000-0001-6293-4173 (A.S. Ulfert-Blank); 0000-0002-9070-5930 (E.Georganta); 0000-0002-7826-5821 (M.Tielman); 0000-0002-9523-0161 (T. Oron-Gilad)

developing 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 HAIT will be essential for teams to communicate, integrate information, coordinate, and perform effectively [12]. Although research in different disciplines offers insights into how human teams develop trust, how humans interact with technology, and how AI systems should be designed, clear guidelines for how HAIT should be designed to foster appropriate trust are missing.

2. Trust perspectives across disciplines

Across disciplines, the study of trust as a central influencing factor on human-technology interaction has a longstanding tradition that has led to various perspectives and definitions [13], [14]. In the 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]. Definitions in these fields often build on organizational research, suggesting that interpersonal trust is impacted by the way a teammate's (trustee) trustworthiness is perceived by another teammate (trustor) [15]. Trustworthiness is defined by three dimensions: ability, integrity, and benevolence. Ability describes how a team member's knowledge or competences are perceived by the other team members [17]. Integrity, describes how a team member's credibility and consistency are perceived [18]. Benevolence reflects perceptions about a team member's concern for the good of the team [19]. Positive expectations of one's trustworthiness can positively impact accepting one's vulnerability [17]. Recent studies on trust in AI agents further show that reliable, autonomous and consistent AI agents are perceived as more trustworthy [20]. Based on the human or AI team members' trustworthiness evaluations, team members will adapt their behavior and collaboration (e.g. deciding not to share information in case of low perceived trustworthiness).

Further, antecedents of trust can include characteristics and states of the human and the AI teammates as well as their shared environment [13]. For instance, the effect of a system's reliability on trust may be influenced by differences in the type of AI system [21]. In contrast to interpersonal trust, team trust describes the shared perception among all team members that enables the free sharing of information and views and reflects one of the most crucial properties for team success [22]. Thus, trust in HAIT involves multiple actors, including human and AI teammates, whose trust is critical to team success [23]. To achieve effective collaboration, human and AI teammates must perceive each other as trustworthy and perceive that they are being trusted [23]. Literature across disciplines agrees that to collaborate effectively, HAIT require appropriate levels of trust that are bidirectional (i.e., expressed by the human and the AI agent). To achieve this goal, the engineering literature proposes design approaches with the goal of building appropriate bidirectional trust between humans and AI (i.e., trust engineering; [24]), for instance, by addressing aspects such as explainability, security, or training.

3. Further theoretical integration is needed

Although the different literature streams have addressed trust as a central construct for effective collaboration, the definitions differ slightly, depending on their unique disciplinary perspective [9]. For instance, a large body of literature in team research focuses on the role of trust in team processes and how individual perceptions by team members impact collaboration, thus, highlighting the perspective of humans as trustors in their collaboration with AI teammates [25]. In contrast, trust engineering literature predominantly addresses technical challenges of HAIT, such as data protection, transparency, or interface design [24]. Currently, there is still a lack of integration of technical system design perspectives and team processes in HAIT.

Further, across disciplines, the focus of trust research is often on dyadic trust relationships between humans and AI agents or between AI agents (e.g., in computer science) [24], rather than

trust between multiple team members and at different levels (dyadic and team). Consequently, it remains unclear how trust in HAIT can be understood and what mechanisms underlie the development of trust in HAIT [20].

To understand the emergence of trust in HAIT, more integrated definitions and interdisciplinary insights are needed . HAIT can largely differ in their composition of human and AI teammates (i.e., number of human and AI teammates or their roles within the team). A team's composition can strongly impact how trust develops and how trustworthy team members are perceived [26]. Yet, current research predominantly highlights reactions to team member characteristics rather than the dynamic development of trust in HAIT, trust reciprocity, or differing trust levels among team members [16]. Trust in HAIT should be considered from a dynamic and multilevel perspective, where team members differ in their characteristics (e.g., AI or human; trustworthiness), their behaviors (e.g., how they display trust), their roles, and their relationships (e.g., their trust relationship with individual team members) [23].

As HAIT find their way into work environments, we urgently require further integration of perspectives to answer pressing questions such as: How can trust in HAIT be defined considering different disciplinary perspectives? What is an appropriate level of trust in HAIT to enable effective collaboration? How can the vast and constantly growing trust literature be unified across disciplines? How can trust in HAIT foster performance as well as human safety and well-being?

4. Addressing trust from multidisciplinary perspectives

To provide first answers, our project "We are in this together" develops theoretical and empirical arguments for gaining a better understanding of trust in HAIT (see Figure 1) to move towards a workspace model that can be understood across disciplines. The research model focuses on individual characteristics, perceptions and reactions, as well as relationships between all team members, both human and AI aiming to define and map the components and emergence of trust in HAIT.

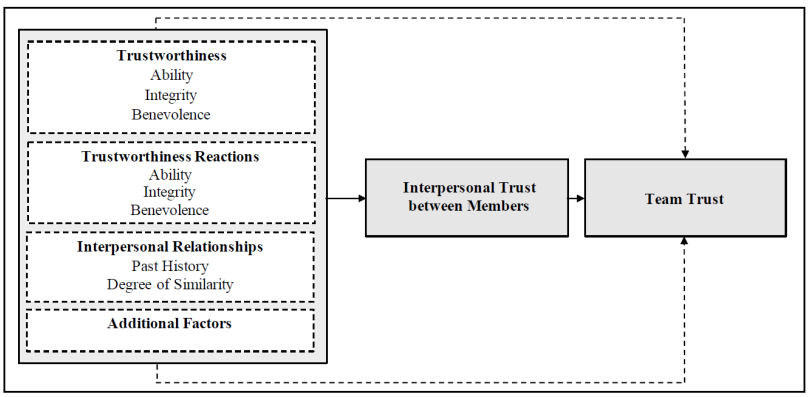

Figure 1. Research model of the project "We are in this together".

In this project, we integrate knowledge from Industrial and Organizational Psychology, Human Factors Engineering, Human-Computer Interaction, and Computer Science about teams, trust, and interaction between humans and (intelligent) technologies. In line with prior works [12], [23], we suggest that for trust to develop, both human and AI teammates must be able to perceive and express trust. Thus, we propose that:

Proposition 1: Trust in HAIT considers human and AI teammates as trustors and trustees.

For trust to develop, human and AI teammates will evaluate the trustworthiness of their teammates and appropriately react to the evaluations of their own trustworthiness. Yet, research

on how humans perceive, evaluate, and react to such trustworthiness reactions by AI agents is still scarce. Prior works on trustworthiness reactions in human-only interactions may be guiding in further exploring how humans may react to trust expressions by AI teammates [27], [28]. In addition, building on team research, we suggest that interpersonal trust further depends on the interpersonal relationships between team members [29]. Specifically, having a high degree of similarity and past experiences can have a positive impact on interpersonal trust [12]. This is because trust develops more naturally when team members have similarities. Recognizing similarities leads to assuming similarities in values and beliefs [30] and thus, results in a sense of comfort and a willingness to trust [31]. Given that similar social rules and evaluation strategies apply when humans build trust relationships with intelligent technologies [32], we expect that when human team members perceive an AI teammate as similar and, in consequence, as part of the team, their interpersonal trust will become stronger [33], [34]. At the same time, we expect that prior history of working together can support interpersonal trust between human and AI teammates. Related work suggests that having experience with intelligent technologies influences how much a human team member develops trust in this technology [14]. Overall, we propose that:

> *Proposition 2*: Trust in HAIT depends on human and AI trustworthiness, their trustworthiness reactions, and interpersonal relationship between teammates.

As previously described, in HAIT, individual team members will develop trust towards other individual team members, dyads, and the team as a whole [23]. Thus, we propose that:

> *Proposition 3*: Trust in HAIT is multilevel, including individual-, dyadic-, and team-level trust.

We further propose that (4) additional factors, such as individual, technological, and environmental considerations (e.g., system characteristics; complexity of the team and its roles), form and impact trust relationships in HAIT [16]. Our overall goal is to contribute to developing effective HAIT by presenting a research model of trust in human-AI teams that can serve as the foundation for unified conceptualizations and measurements of trust, as well as interdisciplinary empirical insights.

Acknowledgments

This research was funded by the SIOP Visionary Grant.

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.

[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] J. A. Colquitt, B. A. Scott, and J. A. LePine, 'Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance.', *Journal of applied psychology*, vol. 92, no. 4, p. 909, 2007.

[18] C. A. Fulmer and M. J. Gelfand, 'At what level (and in whom) we trust: Trust across multiple organizational levels', *Journal of management*, vol. 38, no. 4, pp. 1167–1230, 2012.

[19] F. D. Schoorman, R. C. Mayer, and J. H. Davis, *An integrative model of organizational trust: Past, present, and future*. Academy of Management Briarcliff Manor, NY 10510, 2007.

[20] M. Langer, C. J. König, C. Back, and V. Hemsing, 'Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias', *J Bus Psychol*, pp. 1–16, Jun. 2022, doi: 10.1007/s10869-022-09829-9.

[21] A. Weinstock, T. Oron-Gilad, and Y. Parmet, 'The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system', *Work*, vol. 41 Suppl 1, pp. 258–265, 2012, doi: 10.3233/WOR-2012-0166-258.

[22] 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.

[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] S. Utz, U. Matzat, and C. Snijders, 'On-Line Reputation Systems: The Effects of Feedback Comments and Reactions on Building and Rebuilding Trust in On-Line Auctions', *International Journal of Electronic Commerce*, vol. 13, no. 3, pp. 95–118, 2009.
[28] O. Eilam and R. Suleiman, 'Cooperative, pure, and selfish trusting: Their distinctive effects on the reaction of trust recipients', *Eur. J. Soc. Psychol.*, vol. 34, no. 6, pp. 729–738, Nov. 2004, doi: 10.1002/ejsp.227.
[29] A. C. Costa, C. A. Fulmer, and N. R. Anderson, 'Trust in work teams: An integrative review, multilevel model, and future directions', *Journal of Organizational Behavior*, vol. 39, no. 2, pp. 169–184, 2018.
[30] A. S. Tsui, L. W. Porter, and T. D. Egan, 'When both similarities and dissimilarities matter: Extending the concept of relational demography', *Human relations*, vol. 55, no. 8, pp. 899–929, 2002.
[31] J. L. Wildman *et al.*, 'Trust development in swift starting action teams: A multilevel framework', *Group & Organization Management*, vol. 37, no. 2, pp. 137–170, 2012.
[32] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, 'Automation failures on tasks easily performed by operators undermine trust in automated aids', *Human Factors*, vol. 48, no. 2, pp. 241–256, 2006, doi: 10.1518/001872006777724408.
[33] J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, and M. S. Wetherell, *Rediscovering the social group: A self-categorization theory.* Basil Blackwell, 1987.
[34] J. C. Turner, 'Social categorization and the self-concept: A social cognitive theory of group behavior.', 2010.