Trust Dispersion and Effective Human-AI Team Collaboration:

The Role of Psychological Safety

Tilman Nols[a], Anna-Sophie Ulfert-Blank [a] and Avi Parush [b]
[a] *Eindhoven University of Technology, the Netherlands*
[b] *Technion – Israel Institute of Technology, Israel*

ORCiD ID: Tilman Nols https://orcid.org/0009-0007-5551-4075

ORCiD ID: Anna-Sophie Ulfert-Blank https://orcid.org/ 0000-0001-6293-4173

ORCiD ID: Avi Parush https://orcid.org/ 0000-0003-4435-8576

Abstract. Trust is a crucial factor in team performance for human-human and human-AI teams. While research made significant advancements on factors affecting the human decision to trust their AI teammate, it disregards the potential dynamics of trust in teams with multiple team members. To address this gap, we propose that trust in AI is an emergent state that can be differentiated on the individual and team level. We highlight the importance of considering the dispersion of trust levels in human-AI teams to understand better how trust influences team performance. Furthermore, we transfer the concept of psychological safety from human psychology literature and propose its role in buffering the potential adverse effects of dispersed trust attitudes.

Keywords. Human-AI team, Trust, Psychological Safety, Emergent States

1. Introduction

As artificial intelligence (AI) advances, the potential for collaboration between humans and machines has become an increasingly important research topic. Human-AI teaming involves the integration of human and AI capabilities to achieve joint goals and has the potential to revolutionize a wide range of industries and fields [1]. Although promising, human-AI teamwork often faces challenges as human team members are unwilling to accept suggestions from their AI team member or overly rely on recommendations due to inappropriate trust levels [2], [3].

Trust describes "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control the other party" [4], p. 710). It can improve performance in human teams and human collaboration with

artificial intelligence [5], [6]. Specifically, when team members have high levels of trust, they are more willing to work together, share information, and accept suggestions from others, which can translate into improved teamwork processes, such as coordination, goal negotiation, or conflict management [5], [7]. Thus, in recent years, many researchers have highlighted the importance of trust for collaboration in human-AI teams [7], [8].

Although a large body of literature addresses human trust in technologies, prior works have predominantly focused on the trust of an individual human user in a specific system rather than teams of multiple humans or multiple AI systems. This disregards the diverse relationships and dynamics that may exist between team members within human-AI teams [7]. Psychological literature considers trust in teams to be an emergent state resulting from the interactions and relationships among team members [9]. Accordingly, emergent states like trust do not characterize the nature of team processes such as communication or collective decision-making. Rather, these properties emerge from the dynamics between multiple team members and serve as an input variable to subsequent team processes [10]. These states are temporary and subject to changes that result from various factors, such as the environment the team operates in and individual differences between team members.

Based on psychological and human-technology interaction literature, we formulate three propositions on how trust in human-AI teams impacts collaboration. Specifically, we suggest that in human-AI teams, (1) individual team members differ in their trust relationship with the AI team members, (2) these differences in trust impact overall team trust, and (3) interact with different emergent group-level phenomena, especially psychological safety, that may buffer negative effects of different trust beliefs of team members.

2. Not all trust is created equal

Past research made great efforts to understand what technological factors improve trusting behaviors towards AI (e.g., transparency; [11], [12]). Nevertheless, whether humans trust their artificial teammates does not only depend on AI characteristics but also human characteristics. Accordingly, research has demonstrated that individuals vary in their trustworthiness perception of the same AI [3].

Team members may further vary in their understanding of the AI team member (e.g., AI literacy; [13]), their perceptions of AI characteristics (e.g., usefulness; [14]), or their experience in interacting with such technologies [7] which can consequently impact their trust. In addition, individual differences (e.g., propensity to trust; [3]) or situational changes [15] may further yield differences between team members' trust in the AI and, subsequently, their trusting behaviors (e.g., relying on the AI or not; [16]). To summarize, humans evaluate their trust in the AI team member based on various factors specific to the individual. Thus, we propose that:

Proposition 1: Within the same human-AI team, human team members differ in their trust in an AI team member, depending on their understanding and perception of the AI, prior experience, and individual differences.

Until now, literature on trust in human-AI teams has predominantly focused on individual team members' trust towards AI [17]. However, given that human-AI teams

may be composed of multiple team members, it must be acknowledged that next to individual effects, these teams are also influenced by trust on the team level [10]. Regarding human-AI teams, low levels of team trust in the AI may lead team members to decide to reject or ignore recommendations by the AI collectively. In contrast, high team trust in the AI team member would increase the collective reliance on the AI.

As such, trust in AI is considered an emergent phenomenon that differs at the individual and team level [17], [10], [18]. That is, teams harbor individual level perceptions that compile or compose aggregations on the collective level (bottom-up; [19]). This differentiation is important since the two levels are often interdependent but conceptually distinct and may, therefore, jointly help to explain variance in observed behavior [20].

Moreover, recent trust literature suggests that team-level trust should consider the mean of individual perceptions and focus on the degree of agreement or consensus among team members [18], [21]. In fact, some authors argue that team-level constructs are only meaningful if sufficient agreement between team members is achieved [22]. Considering both magnitude and consensus of trust perceptions enables researchers to understand better how trust manifests itself, acknowledge the underlying trust dynamics and further delineate how trust on the team level impacts collaboration [23].

Given that individual trust perceptions can vary (see *Proposition 1*), team members' trust levels may either converge and create a shared sense of team trust or diverge and show high variance in trust magnitude perceptions. When trust in AI on the individual level varies greatly, we may speak about a large dispersion of trust. In contrast, when individual perceptions are shared among team members, the dispersion of trust is low.

A large dispersion of trust beliefs within human teams has been shown to negatively impact performance [21]. Asymmetric individual trust levels can impede the team's ability to make high-quality decisions as it cannot capitalize on the hypothesized positive effects of overall team trust [24] [25]. For instance, when individuals show high levels of trust, they are less skeptical and more willing to accept recommendations from the AI [3]. If, within a human-AI team, all human team members possess similar levels of trust towards the AI, the team will be more confident in their decision-making due to the high magnitude *and* similarity of trust levels. In contrast, when a team has highly dispersed trust perceptions, the usefulness of AI team member's recommendations may be evaluated differently per individual. This dispersion of trust towards the AI, in turn, may reduce the likelihood of finding consensus in collective decision-making.Consequently, a team might be more susceptible to conflict, process loss, and, subsequently, inferior decision-making quality. It is therefore proposed that:

Proposition 2: The influence of team-level trust in the AI team member on team processes depends on the magnitude and dispersion of individual trust in the AI.

To further elaborate on how trust affects teamwork, it may be worthwhile to consider if and why some team processes are more affected than others. Generally, literature categorizes team processes into reoccurring phases of action- (e.g., back-up) and transition-processes (e.g., mission analysis and formulation)[9]. Additionally, interpersonal processes such as conflict management influence the effectiveness of concurrent teamwork activities throughout both phases. Importantly, these team interactions and experiences give rise to emergent states like team trust that in turn,

influence subsequent team processes [9]. In line with this assumption, team trust is ubiquitous and may influence all teamwork processes.

For instance, due to an inherent relational uncertainty of highly dispersed team trust, transition-related processes such as situation assessment or plan formulation may suffer from more skepticism and less effective information integration. On the other hand, highly dispersed team trust may decrease the confidence in others [25] and increasingly prompt the reliance on risk-reducing control strategies (e.g., monitoring others).

However, to our knowledge there is currently no research differentiating the effect of team trust on team processes empirically. As a result, we refrain from postulating clear propositions on more fine-grained relationships between team trust and team processes.

3. Psychological Safety - Capitalizing on unequal trust perceptions

Acknowledging the complexity of trust in human-AI teams helps explain the effect of trust on performance in more detail. Differences in team members' trust towards AI can lead to unequal perceptions, such as understanding the AI's role or decision-making, perceived usefulness, or perceived risk. This can be critical for the team's decision-making, for instance, when deciding whether to rely on or reject a recommendation by the AI team member. Diverging perspectives can exacerbate the teams' difficulties in reaching a consensus in their collective decision-making, affecting reliance on the AI team member. However, the mere presence of conflicting attitudes does not automatically lead to negative consequences.

In general, conflicting viewpoints can be considered both an asset and a barrier to team processes [26]. While dispersed trust levels may cause inefficiencies in group decision-making, they can also enrich the group's perspective on the problem [27]. In human teams, differences between individual team members (e.g., personality, expertise, attitudes) have been linked to team collaboration and performance if team members are enabled to share these differences [28]. Similarly, if team members can present and discuss their conflicting points of view in a human-AI team, and divergent perspectives are taken seriously, the team's understanding of and collaboration with the AI may even be improved. By disclosing their attitudes and reasoning, raising doubts and concerns, or asking questions, the team not only expands the informational basis for a critical decision, but also increases the likelihood to align trust perceptions [29], [30].

The impact of dispersed levels of trust on group decision-making hinges on the team's ability to manage and reconcile conflicting attitudes effectively. In human-team research, psychological safety is one critical determinant of dealing successfully with disagreement (PS; [27]). PS describes the perception that it is safe to take interpersonal risks [31]. Like team-level trust, PS is considered an emergent group-level phenomenon. The idea is that in all teams relationship dynamics are at play that signal to team members whether they feel appreciated and whether the pushing and pulling of information are associated with negative consequences. Therefore, PS is associated with a "sense of confidence that the team will not embarrass, reject, or punish someone for speaking up" ([32] p. 354). Accordingly, studies show that PS moderates the effect of interpersonal processes, such as team conflict [27]. While team conflict is generally associated with a negative effect on team performance due to a loss in harmony and productivity, PS can invert that relationship. If teams show high PS, individuals are

invited to elaborate on their conflicting viewpoints, which can benefit the creativity of decision-making. Furthermore, teams may be more reluctant to reach an agreement too quickly, refrain from group thinking and improve their rigor in decision-making. Thus, PS may help perceive team members' conflict not as a barrier but as a potential resource that facilitates decision-making [27].

Although PS has a solid theoretical and empirical basis in human team research, its existence and effect in human-AI teams remain largely unexplored. Nonetheless, in line with other researchers (e.g., [33]), we advocate increasingly focusing on social dynamics such as PS in human-AI team performance. In particular, we argue that, similar to the conflict study cited above, PS can help teams to deal with diverging attitudes toward AI. Trust dispersion may become an issue if team members perceive low PS. Consequently, individuals suffer from relational uncertainty and reduce their investments in social exchange [24]. As such, the informational basis of a team is neither questioned nor enriched; in addition, trust perceptions are likely to remain dispersed. However, when PS is high, trust dispersion may benefit (or at least not harm) decision-making by prompting team members to contribute their perception and facilitating confidence in the collective action plan. It is thus proposed that:

Proposition 3: Psychological safety moderates the effect of highly dispersed trust in the AI so that higher psychological safety buffers potential negative consequences of variations in individual trust towards the AI team member.

4. Discussion

The present paper argues that trust and human-AI teaming research can greatly benefit from a more dynamic perspective. The future of human-AI teams is not limited to dyadic team compositions but may entail teams composed of multiple humans and/or AI agents. Individual and team-level factors in these teams influence how members accomplish their work together. In that regard, we proposed two group-level phenomena, trust dispersion and psychological safety, that may influence teamwork in human-AI teams. However, in the current propositions, we only argue for the emergence of trust toward AI between human teammates. It may be a worthwhile agenda for future research to also consider whether and how AI can contribute to a state of emergence (e.g., team cohesion, team trust) across all team members. For instance, in one study, the display of vulnerability by a robot positively contributed to trust across the rest of the team (i.e., ripple effect) [34]. In addition, team member may also engage in trust dampening and repairing behaviours to calibrate effective trust over time [2]. This raises further questions on how trust dynamics unfold and impact team interactions across situations. Finally, human team research established different forms of trust development (i.e., affective and cognitive trust) that have different trajectories and relationships with team performance over time [35]. Research should also consider to what extent these different trust conceptualizations play a role in human-AI teams, and whether within-and between-individual differences give rise to meaningful collaboration difficulties.

References

[1] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, 'Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature', *Hum Factors*, vol. 64, no. 5, pp. 904–938, Aug. 2022, doi: 10.1177/0018720820960865.

[2] E. J. De Visser *et al.*, 'Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams', *Int J of Soc Robotics*, vol. 12, no. 2, pp. 459–478, May 2020, doi: 10.1007/s12369-019-00596-x.

[3] 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.

[4] 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.

[5] A. C. Costa, C. A. Fulmer, and N. R. Anderson, 'Trust in work teams: An integrative review, multilevel model, and future directions', *J Organ Behav*, vol. 39, no. 2, pp. 169–184, Feb. 2018, doi: 10.1002/job.2213.

[6] 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', 2021.

[7] A. S. Ulfert and E. Georganta, 'A model of team trust in human-agent teams', presented at the ICMI 2020 Companion - Companion Publication of the 2020 International Conference on Multimodal Interaction, Association for Computing Machinery, Inc, Oct. 2020, pp. 171–176. doi: 10.1145/3395035.3425959.

[8] C. Centeio Jorge, M. L. Tielman, and C. M. Jonker, 'Artificial Trust as a Tool in Human-AI Teams', in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, in HRI '22. Sapporo, Hokkaido, Japan: IEEE Press, Mar. 2022, pp. 1155–1157.

[9] M. A. Marks, J. E. Mathieu, and S. J. Zaccaro, 'A Temporally Based Framework and Taxonomy of Team Processes', *The Academy of Management Review*, vol. 26, no. 3, p. 356, Jul. 2001, doi: 10.2307/259182.

[10] B. Fyhn, V. Schei, and T. E. Sverdrup, 'Taking the emergent in team emergent states seriously: A review and preview', *Human Resource Management Review*, vol. 33, no. 1, p. 100928, Mar. 2023, doi: 10.1016/j.hrmr.2022.100928.

[11] J. Y. C. Chen, S. G. Lakhmani, K. Stowers, A. R. Selkowitz, J. L. Wright, and M. Barnes, 'Situation awareness-based agent transparency and human-autonomy teaming effectiveness', *Theoretical Issues in Ergonomics Science*, vol. 19, no. 3, pp. 259–282, May 2018, doi: 10.1080/1463922X.2017.1315750.

[12] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. C. Chen, 'Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems', in *Unmanned Systems Technology XVI*, SPIE, Jun. 2014, pp. 112–123. doi: 10.1117/12.2050622.

[13] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, 'In AI we trust? Perceptions about automated decision-making by artificial intelligence', *AI & SOCIETY*, vol. 35, no. 3, pp. 611–623, 2020.

[14] H. Choung, P. David, and A. Ross, 'Trust in AI and Its Role in the Acceptance of AI Technologies', *International Journal of Human–Computer Interaction*, pp. 1–13, Apr. 2022, doi: 10.1080/10447318.2022.2050543.

[15] J. D. Lee and K. A. See, 'Trust in Automation: Designing for Appropriate Reliance', *Human Factors*, 2004.

[16] 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.

[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] S. W. J. Kozlowski and G. T. Chao, 'The Dynamics of Emergence: Cognition and Cohesion in Work Teams: THE DYNAMICS OF EMERGENCE', *Manage. Decis. Econ.*, vol. 33, no. 5–6, pp. 335–354, Jul. 2012, doi: 10.1002/mde.2552.

[19] S. W. J. Kozlowski and G. T. Chao, 'Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods.', *American Psychologist*, vol. 73, no. 4, pp. 576–592, May 2018, doi: 10.1037/amp0000245.

[20] S. M. Fiore and K. A. Kapalo, 'Innovation in Team Interaction: New Methods for Assessing Collaboration Between Brains and Bodies Using a Multi-level Framework', in *Innovative Assessment of Collaboration*, A. A. von Davier, M. Zhu, and P. C. Kyllonen, Eds., in Methodology of Educational Measurement and Assessment. Cham: Springer International Publishing, 2017, pp. 51–64. doi: 10.1007/978-3-319-33261-1_4.

[21] B. A. De Jong and K. T. Dirks, ‘Beyond shared perceptions of trust and monitoring in teams: Implications of asymmetry and dissensus.’, *Journal of Applied Psychology*, vol. 97, no. 2, pp. 391–406, 2012, doi: 10.1037/a0026483.
[22] A. C. Costa, C. A. Fulmer, and N. R. Anderson, ‘Trust in work teams: An integrative review, multilevel model, and future directions’, *J Organ Behav*, vol. 39, no. 2, pp. 169–184, Feb. 2018, doi: 10.1002/job.2213.
[23] 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.
[24] B. de Jong, N. Gillespie, I. Williamson, and C. Gill, ‘Trust Consensus Within Culturally Diverse Teams: A Multistudy Investigation’.
[25] G. R. Jones and J. M. George, ‘The Experience and Evolution of Trust: Implications for Cooperation and Teamwork’, *The Academy of Management Review*, vol. 23, no. 3, p. 531, Jul. 1998, doi: 10.2307/259293.
[26] T. A. O’Neill and M. J. W. McLarnon, ‘Optimizing team conflict dynamics for high performance teamwork’, *Human Resource Management Review*, vol. 28, no. 4, pp. 378–394, Dec. 2018, doi: 10.1016/j.hrmr.2017.06.002.
[27] B. H. Bradley, B. E. Postlethwaite, A. C. Klotz, M. R. Hamdani, and K. G. Brown, ‘Reaping the benefits of task conflict in teams: The critical role of team psychological safety climate.’, *Journal of Applied Psychology*, vol. 97, no. 1, pp. 151–158, 2012, doi: 10.1037/a0024200.
[28] U. R. Hülsheger, N. Anderson, and J. F. Salgado, ‘Team-level predictors of innovation at work: A comprehensive meta-analysis spanning three decades of research.’, *Journal of Applied Psychology*, vol. 94, no. 5, pp. 1128–1145, 2009, doi: 10.1037/a0015978.
[29] S. Tyagi, R. Sibal, and B. Suri, ‘Empirically developed framework for building trust in distributed agile teams’, *Information and Software Technology*, vol. 145, p. 106828, May 2022, doi: 10.1016/j.infsof.2022.106828.
[30] T. Savolainen, ‘Process dynamics of trust development: exploring and illustrating emergence in the team context’, in *Trust, Organizations and Social Interaction*, Edward Elgar Publishing, 2016, pp. 231–256. doi: 10.4337/9781783476206.00022.
[31] A. C. Edmondson and D. P. Bransby, ‘Psychological Safety Comes of Age: Observed Themes in an Established Literature’, *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 10, no. 1, pp. 55–78, Jan. 2023, doi: 10.1146/annurev-orgpsych-120920-055217.
[32] A. Edmondson, ‘Psychological Safety and Learning Behavior in Work Teams’, *Administrative Science Quarterly*, vol. 44, no. 2, pp. 350–383, Jun. 1999, doi: 10.2307/2666999.
[33] J. B. Lyons, K. Sycara, M. Lewis, and A. Capiola, ‘Human–Autonomy Teaming: Definitions, Debates, and Directions’, *Front. Psychol.*, vol. 12, p. 589585, May 2021, doi: 10.3389/fpsyg.2021.589585.
[34] S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati, ‘The Ripple Effects of Vulnerability: The Effects of a Robot’s Vulnerable Behavior on Trust in Human-Robot Teams’, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, Chicago IL USA: ACM, Feb. 2018, pp. 178–186. doi: 10.1145/3171221.3171275.
[35] S. S. Webber, ‘Development of Cognitive and Affective Trust in Teams: A Longitudinal Study’, *Small Group Research*, vol. 39, no. 6, pp. 746–769, Dec. 2008, doi: 10.1177/1046496408323569.